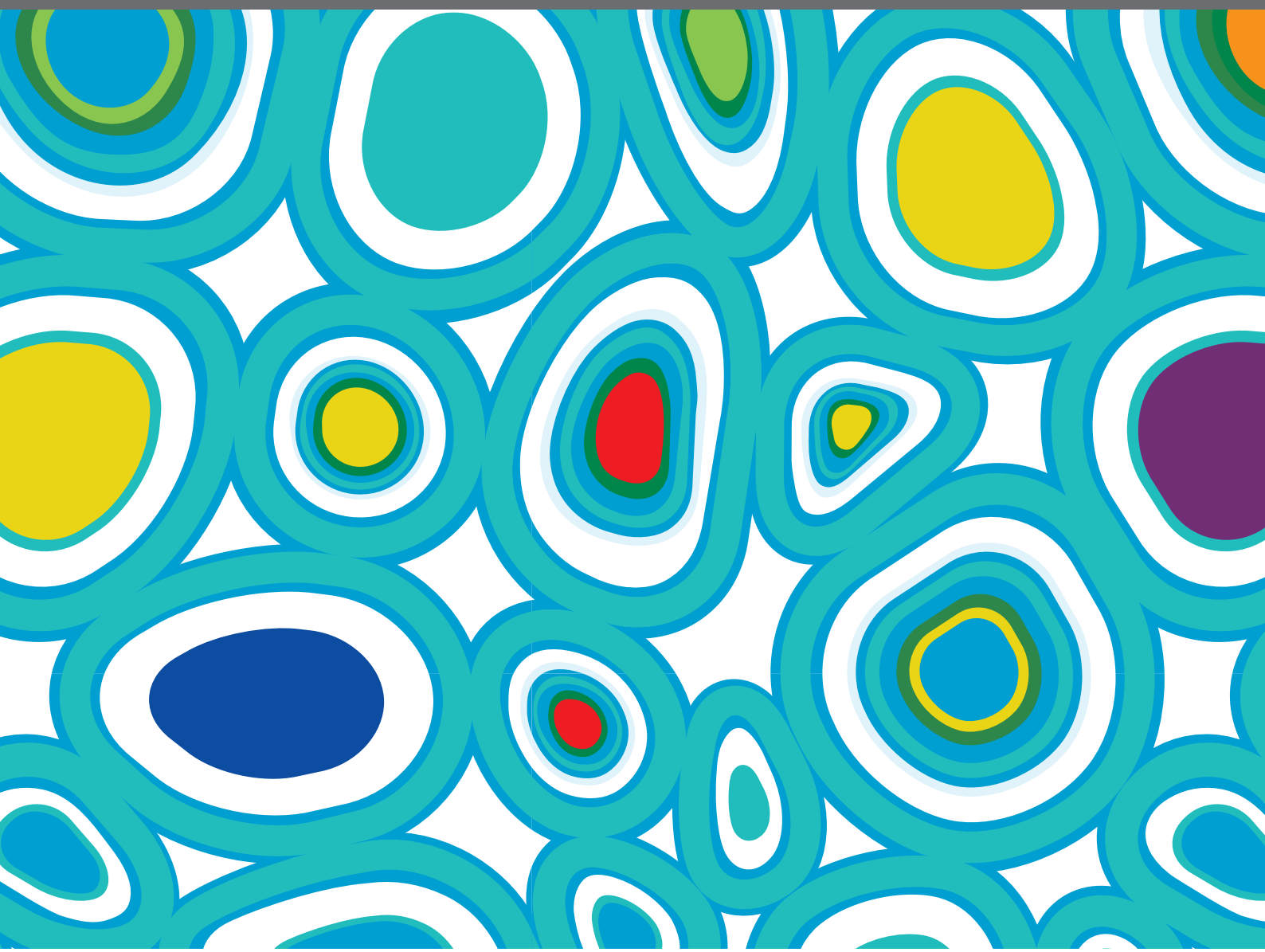
OMICS DATA INTEGRATION TOWARDS MINING OF PHENOTYPE SPECIFIC BIOMARKERS IN CANCER - VOLUME II

EDITED BY: Liang Cheng, Lei Deng, Chuan-Xing Li, Yan Zhang and Mingxiang Teng

PUBLISHED IN: Frontiers in Cell and Developmental Biology



frontiers Research Topics

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-83250-738-4
DOI 10.3389/978-2-83250-738-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

OMICS DATA INTEGRATION TOWARDS MINING OF PHENOTYPE SPECIFIC BIOMARKERS IN CANCER - VOLUME II

Topic Editors:
Liang Cheng, Harbin Medical University, China
Lei Deng, Central South University, China
Chuan-Xing Li, Karolinska Institutet (KI), Sweden
Yan Zhang, Harbin Institute of Technology, China
Mingxiang Teng, Moffitt Cancer Center, United States

Citation: Cheng, L., Deng, L., Li, C.-X., Zhang, Y., Teng, M., eds. (2022). Omics Data Integration Towards Mining of Phenotype Specific Biomarkers in Cancer - Volume II. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-738-4

Table of Contents

08 ***Identification of Key Genes With Differential Correlations in Lung Adenocarcinoma***
You Zhou, Bin Xu, Yi Zhou, Jian Liu, Xiao Zheng, Yingting Liu, Haifeng Deng, Ming Liu, Xiubao Ren, Jianchuan Xia, Xiangyin Kong, Tao Huang and Jingting Jiang

20 ***Pan-Cancer Metastasis Prediction Based on Graph Deep Learning Method***
Yining Xu, Xinran Cui and Yadong Wang

28 ***Identification and Validation of a Novel Prognosis Prediction Model in Adrenocortical Carcinoma by Integrative Bioinformatics Analysis, Statistics, and Machine Learning***
Xin Yan, Zi-Xin Guo, Dong-Hu Yu, Chen Chen, Xiao-Ping Liu, Zhi-Wei Yang, Tong-Zu Liu and Sheng Li

47 ***Identify Function of WASL in Prognosis of Cervical Cancer Based on Omics Data***
Jinxuan Hou, Chen Chen, Yingying Hu, Qing Gong, Lijuan Gan and Yu Xu

55 ***Uncovering the Subtype-Specific Molecular Characteristics of Breast Cancer by Multiomics Analysis of Prognosis-Associated Genes, Driver Genes, Signaling Pathways, and Immune Activity***
Xinhui Li, Jian Zhou, Mingming Xiao, Lingyu Zhao, Yan Zhao, Shuoshuo Wang, Shuangshu Gao, Yuan Zhuang, Yi Niu, Shijun Li, Xiaobo Li, Yuanyuan Zhu, Minghui Zhang and Jing Tang

68 ***Comprehensive Analysis of Ferroptosis Regulators With Regard to PD-L1 and Immune Infiltration in Clear Cell Renal Cell Carcinoma***
Song Wang, Shiming Chen, Yufan Ying, Xueyou Ma, Haixiang Shen, Jiangfeng Li, Xiao Wang, Yiwei Lin, Ben Liu, Xiangyi Zheng and Liping Xie

83 ***Characterizing the Metabolic and Immune Landscape of Non-small Cell Lung Cancer Reveals Prognostic Biomarkers Through Omics Data Integration***
Fengjiao Wang, Yuanfu Zhang, Yangyang Hao, Xuexin Li, Yue Qi, Mengyu Xin, Qifan Xiao and Peng Wang

96 ***DeepGP: An Integrated Deep Learning Method for Endocrine Disease Gene Prediction Using Omics Data***
Ningyi Zhang, Haoyan Wang, Chen Xu, Liyuan Zhang and Tianyi Zang

105 ***Multi-Omics Data Integration Analysis of an Immune-Related Gene Signature in LGG Patients With Epilepsy***
Quan Cheng, Weiwei Duan, Shiqing He, Chen Li, Hui Cao, Kun Liu, Weijie Ye, Bo Yuan and Zhiwei Xia

119 ***Identification of Disease-Related 2-Oxoglutarate/Fe (II)-Dependent Oxygenase Based on Reduced Amino Acid Cluster Strategy***
Jian Zhou, Suling Bo, Hao Wang, Lei Zheng, Pengfei Liang and Yongchun Zuo

129 ***Identification of Novel Susceptible Genes of Gastric Cancer Based on Integrated Omics Data***
Huang Yaoxing, Yu Danchun, Sun Xiaojuan, Jiang Shuman, Yan Qingqing and Jia Lin

135 ***Identification of Somatic Mutation-Driven Immune Cells by Integrating Genomic and Transcriptome Data***
Ying Jiang, Baotong Zheng, Yang Yang, Xiangmei Li and Junwei Han

145 ***Identify Inflammatory Bowel Disease-Related Genes Based on Machine Learning***
Lili Ye, Yongwei Lin, Xing-di Fan, Yaoming Chen, Zengli Deng, Qian Yang, Xiaotian Lei, Jizong Mao and Chunhui Cui

152 ***Probing Synergistic Targets by Natural Compounds for Hepatocellular Carcinoma***
Jian Gao, Zuojing Yin, Zhuanbin Wu, Zhen Sheng, Chao Ma, Rui Chen, Xiongwen Zhang, Kailin Tang, Jian Fei and Zhiwei Cao

163 ***Identification of Claudin-6 as a Molecular Biomarker in Pan-Cancer Through Multiple Omics Integrative Analysis***
Chiyuan Zhang, Cuishan Guo, Yan Li, Kuiran Liu, Qi Zhao and Ling Ouyang

182 ***Prediction of Genetic Factors of Hyperthyroidism Based on Gene Interaction Network***
Fei Shen, Wensong Cai, Xiaoxiong Gan, Jianhua Feng, Zhen Chen, Mengli Guo, Fang Wei, Jie Cao and Bo Xu

189 ***Identification of Biomarkers for Cervical Cancer Radiotherapy Resistance Based on RNA Sequencing Data***
Yue Feng, Zhao Wang, Nan Yang, Sijia Liu, Jiazhuo Yan, Jiayu Song, Shanshan Yang and Yunyan Zhang

199 ***Comprehensive Analysis of Glycolysis-Related Genes for Prognosis, Immune Features, and Candidate Drug Development in Colon Cancer***
Zhongqi Cui, Guifeng Sun, Ramesh Bhandari, Jiayi Lu, Mengmei Zhang, Rajeev Bhandari, Fenyong Sun, Zhongchen Liu and Shasha Zhao

215 ***Identification of Novel Metabolism-Associated Subtypes for Pancreatic Cancer to Establish an Eighteen-Gene Risk Prediction Model***
Yang Gao, Enchong Zhang, Xiang Fei, Lingming Kong, Peng Liu and Xiaodong Tan

232 ***An Ovarian Cancer Susceptible Gene Prediction Method Based on Deep Learning Methods***
Lu Ye, Yi Zhang, Xinying Yang, Fei Shen and Bo Xu

240 ***Psi-Caller: A Lightweight Short Read-Based Variant Caller With High Speed and Accuracy***
Yadong Liu, Tao Jiang, Yan Gao, Bo Liu, Tianyi Zang and Yadong Wang

251 ***Prioritizing Susceptible Genes for Thyroid Cancer Based on Gene Interaction Network***
Lin-kun Zhong, Chang-lian Xie, Shan Jiang, Xing-yan Deng, Xiao-xiong Gan, Jian-hua Feng, Wen-song Cai, Chi-zhuai Liu, Fei Shen, Jian-hang Miao and Bo Xu

257 ***Recognizing Pattern and Rule of Mutation Signatures Corresponding to Cancer Types***
Lei Chen, Xianchao Zhou, Tao Zeng, Xiaoyong Pan, Yu-Hang Zhang, Tao Huang, Zhaoyuan Fang and Yu-Dong Cai

268 ***Heterogeneous Information Network-Based Patient Similarity Search***
Hao-zhe Huang, Xu-dong Lu, Wei Guo, Xin-bo Jiang, Zhong-min Yan and Shi-peng Wang

278 ***Systematic Characterization of Novel Immune Gene Signatures Predicts Prognostic Factors in Hepatocellular Carcinoma***
Dafeng Xu, Yu Wang, Jincai Wu, Yuliang Zhang, Zhehao Liu, Yonghai Chen and Jinfang Zheng

297 ***Large Scale Identification of Osteosarcoma Pathogenic Genes by Multiple Extreme Learning Machine***
Zhipeng Zhao, Jijun Shi, Guang Zhao, Yanjun Gao, Zhigang Jiang and Fusheng Yuan

303 ***Prognostic Value and Immunological Role of MORF4-Related Gene-Binding Protein in Human Cancers***
Dongqi Chai, Lilong Zhang, Yongjun Guan, Jingping Yuan, Man Li and Weixing Wang

323 ***Prediction of Ovarian Cancer-Related Metabolites Based on Graph Neural Network***
Jingjing Chen, Yingying Chen, Kefeng Sun, Yu Wang, Hui He, Lin Sun, Sifu Ha, Xiaoxiao Li, Yifei Ou, Xue Zhang and Yanli Bi

329 ***Histopathological Images and Multi-Omics Integration Predict Molecular Characteristics and Survival in Lung Adenocarcinoma***
Linyan Chen, Hao Zeng, Yu Xiang, Yeqian Huang, Yuling Luo and Xuelei Ma

342 ***Identification of Key eRNAs for Spinal Cord Injury by Integrated Multinomial Bioinformatics Analysis***
Runzhi Huang, Siqiao Wang, Rui Zhu, Shuyuan Xian, Zongqiang Huang, Liming Cheng and Jie Zhang

361 ***Identification of an Immune-Related Prognostic Gene CLEC5A Based on Immune Microenvironment and Risk Modeling of Ovarian Cancer***
Jiacheng Shen, Tingwei Liu, Jia Lv and Shaohua Xu

375 ***Prediction of Gastric Cancer-Related Proteins Based on Graph Fusion Method***
Hao Zhang, Ruisi Xu, Meng Ding and Ying Zhang

381 ***Identification of Immune-Related Genes for Establishment of Prognostic Index in Hepatocellular Carcinoma***
Yinfang Li, Ling Zou, Xuejun Liu, Judong Luo and Hui Liu

395 ***Competitive Endogenous RNA Landscape in Epstein-Barr Virus Associated Nasopharyngeal Carcinoma***
Xiandong Lin, Steven Wang, Keyu Lin, Jingfeng Zong, Qianlan Zheng, Ying Su and Tao Huang

408 ***Therapeutic Response-Based Reclassification of Multiple Tumor Subtypes Reveals Intrinsic Molecular Concordance of Therapy Across Histologically Disparate Cancers***
Yue Xu, Jie Zheng, Zhaoqing Cai, Wang Li, Jens Köhler, Yao Dai, Xiaojie Cheng, Tao Wu, Fan Zhang and Haiyun Wang

422 ***Genomic Variation Prediction: A Summary From Different Views***
Xiuchun Lin

430 ***Colon Cancer-Related Genes Identification and Function Study Based on Single-Cell Multi-Omics Integration***
Xuepu Sun, Yu Guo, Yu Zhang, Peng Zhao, Zhaoqing Wang, Zheng Wei and Haiquan Qiao

437 ***Comprehensive Analysis of Subtype-Specific Molecular Characteristics of Colon Cancer: Specific Genes, Driver Genes, Signaling Pathways, and Immunotherapy Responses***
Fangjie Hu, Jianyi Wang, Minghui Zhang, Shuoshuo Wang, Lingyu Zhao, Hao Yang, Jinrong Wu and Binbin Cui

451 ***Research on the Computational Prediction of Essential Genes***
Yuxin Guo, Ying Ju, Dong Chen and Lihong Wang

460 ***Exploration of Lung Cancer-Related Genetic Factors* via *Mendelian Randomization Method Based on Genomic and Transcriptomic Summarized Data***
Nitao Cheng, Xinran Cui, Chen Chen, Changsheng Li and Jingyu Huang

468 ***An Apoptosis-Related Gene Prognostic Index for Colon Cancer***
Hanmin Tang, Jing Wang, Xuehui Luo, Qi Wang, Jie Chen, Xinyue Zhang, Qiuting Li, Chengyi Gao, Yuesen Li and Suxia Han

481 ***Association of* IL33, IL1RL1, IL1RAP *Polymorphisms and Asthma in Chinese Han Children***
Maolan Wu, Xiangrong Zheng, Juan Huang and Xiaolei Hu

490 ***Gene-Based Testing of Interactions Using XGBoost in Genome-Wide Association Studies***
Yingjie Guo, Chenxi Wu, Zhian Yuan, Yansu Wang, Zhen Liang, Yang Wang, Yi Zhang and Lei Xu

502 ***GBDTLRL2D Predicts LncRNA–Disease Associations Using MetaGraph2Vec and K-Means Based on Heterogeneous Network***
Tao Duan, Zhufang Kuang, Jiaqi Wang and Zhihao Ma

517 ***Comprehensive Analysis Revealed that* CDKN2A *is a Biomarker for Immune Infiltrates in Multiple Cancers***
Zheng Chen, Yingjie Guo, Da Zhao, Quan Zou, Fusheng Yu, Lijun Zhang and Lei Xu

530 ***Identification of Diagnostic Markers for Breast Cancer Based on Differential Gene Expression and Pathway Network***
Shumei Zhang, Haoran Jiang, Bo Gao, Wen Yang and Guohua Wang

541 ***WAFNRLTG: A Novel Model for Predicting LncRNA Target Genes Based on Weighted Average Fusion Network Representation Learning Method***
Jianwei Li, Zhenwu Yang, Duanyang Wang and Zhiguang Li

552 ***Identification of Novel Lung Cancer Driver Genes Connecting Different Omics Levels With a Heat Diffusion Algorithm***
Fei Yuan, Xiaoyu Cao, Yu-Hang Zhang, Lei Chen, Tao Huang, ZhanDong Li and Yu-Dong Cai

562 ***Interpreting Functional Impact of Genetic Variations by Network QTL for Genotype–Phenotype Association Study***
Kai Yuan, Tao Zeng and Luonan Chen

582 ***Identification of Ferroptosis-Related Biomarkers for Prognosis and Immunotherapy in Patients With Glioma***
Junfeng Shi, Donglin Lai, Xiaojia Zuo, Dingsheng Liu, Bing Chen, Yanjun Zheng, Changlian Lu and Xuefeng Gu

601 ***Identification of Prognostic Biomarkers for Bladder Cancer Based on DNA Methylation Profile***
Shumei Zhang, Jingyu Zhang, Qichao Zhang, Yingjian Liang, Youwen Du and Guohua Wang

613 ***Revealing Cavin-2 Gene Function in Lung Based on Multi-Omics Data Analysis Method***
Changsheng Li, Jingyu Huang, Hexiao Tang, Bing Liu and Xuefeng Zhou

622 ***The Characterization of Structure and Prediction for Aquaporin in Tumour Progression by Machine Learning***
Zheng Chen, Shihu Jiao, Da Zhao, Quan Zou, Lei Xu, Lijun Zhang and Xi Su

633 ***Discovered Key CpG Sites by Analyzing DNA Methylation and Gene Expression in Breast Cancer Samples***
Yan-Ni Cao, Qian-Zhong Li and Yu-Xian Liu

647 ***Multi-Omics Prognostic Signatures Based on Lipid Metabolism for Colorectal Cancer***
YuanLin Sun, Bin Liu, YuJia Chen, YanPeng Xing and Yang Zhang

667 ***CD93 Correlates With Immune Infiltration and Impacts Patient Immunotherapy Efficacy: A Pan-Cancer Analysis***
Zerui Zhang, Mengli Zheng, Qiang Ding and Mei Liu

683 ***Gene Expression Profile Reveals a Prognostic Signature of Non–MSI-H/pMMR Colorectal Cancer***
Zaoqu Liu, Hui Xu, Xiaoyong Ge, Siyuan Weng, Qin Dang and Xinwei Han

694 ***Based on Molecular Subtypes, Immune Characteristics and Genomic Variation to Constructing and Verifying Multi-Gene Prognostic Characteristics of Colorectal Cancer***
Lei Gu, Chunhui Jiang, Chunjie Xu, Ye Liu and Hong Zhou

711 ***Integrated Analysis of Multi-Omics Alteration, Immune Profile, and Pharmacological Landscape of Pyroptosis-Derived lncRNA Pairs in Gastric Cancer***
Chunguang Guo, Zaoqu Liu, Yin Yu, Shirui Liu, Ke Ma, Xiaoyong Ge, Zhe Xing, Taoyuan Lu, Siyuan Weng, Libo Wang, Long Liu, Zhaohui Hua, Xinwei Han and Zhen Li

723 ***A Novel Signature of Lipid Metabolism-Related Gene Predicts Prognosis and Response to Immunotherapy in Lung Adenocarcinoma***
Kai Zhang, Ying Qian, Xiaowei Quan, Tengteng Zhu and Biyun Qian

739 ***A Method of Optimizing Weight Allocation in Data Integration Based on Q-Learning for Drug-Target Interaction Prediction***
Jiacheng Sun, You Lu, Linqian Cui, Qiming Fu, Hongjie Wu and Jianping Chen

749 ***Identification of Potential Biomarkers and Small Molecule Drugs for Cutaneous Melanoma Using Integrated Bioinformatic Analysis***
Yong Liu, Jiayi Sun, Dongran Han, Shengnan Cui and Xiaoning Yan

762 ***Glycosylation Modification Patterns Reveal Distinct Tumor Metabolism and Immune Microenvironment Landscape in Lower-Grade Gliomas***
Guihua Tang, Liming Tan, Hao Yuan and Wen Yin

777 ***An Epithelial-Mesenchymal Plasticity Signature Identifies Two Novel LncRNAs With the Opposite Regulation***
Pranjal Kumar, Seema Khadirnaikar, Nikita Bhandari, Annesha Chatterjee and Sudhanshu Shukla

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 05 May 2021
doi: 10.3389/fcell.2021.675438

Identification of Key Genes With Differential Correlations in Lung Adenocarcinoma

You Zhou[1,2,3], Bin Xu[1,2,3], Yi Zhou[1,2,3], Jian Liu[1,2,3], Xiao Zheng[1,2,3], Yingting Liu[1,2,3], Haifeng Deng[1,2,3], Ming Liu[1,2,3], Xiubao Ren[4], Jianchuan Xia[5], Xiangyin Kong[6], Tao Huang[7] and Jingting Jiang[1,2,3]**

[1] Tumor Biological Diagnosis and Treatment Center, The Third Affiliated Hospital of Soochow University, Changzhou, China, [2] Jiangsu Engineering Research Center for Tumor Immunotherapy, Changzhou, China, [3] Institute of Cell Therapy, Soochow University, Changzhou, China, [4] Department of Immunology and Biotherapy, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China, [5] State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China, [6] CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China, [7] Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Xiao Chang,
Children's Hospital of Philadelphia, United States
Yi Zhang,
Fudan University, China

***Correspondence:**
Jingting Jiang
jiangjingting@suda.edu.cn
Tao Huang
tohuangtao@126.com

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *03 March 2021*
Accepted: *24 March 2021*
Published: *05 May 2021*

Citation:
Zhou Y, Xu B, Zhou Y, Liu J, Zheng X, Liu Y, Deng H, Liu M, Ren X, Xia J, Kong X, Huang T and Jiang J (2021) Identification of Key Genes With Differential Correlations in Lung Adenocarcinoma. Front. Cell Dev. Biol. 9:675438. doi: 10.3389/fcell.2021.675438

Background: With the advent of large-scale molecular profiling, an increasing number of oncogenic drivers contributing to precise medicine and reshaping classification of lung adenocarcinoma (LUAD) have been identified. However, only a minority of patients archived improved outcome under current standard therapies because of the dynamic mutational spectrum, which required expanding susceptible gene libraries. Accumulating evidence has witnessed that understanding gene regulatory networks as well as their changing processes was helpful in identifying core genes which acted as master regulators during carcinogenesis. The present study aimed at identifying key genes with differential correlations between normal and tumor status.

Methods: Weighted gene co-expression network analysis (WGCNA) was employed to build a gene interaction network using the expression profile of LUAD from The Cancer Genome Atlas (TCGA). R package DiffCorr was implemented for the identification of differential correlations between tumor and adjacent normal tissues. STRING and Cytoscape were used for the construction and visualization of biological networks.

Results: A total of 176 modules were detected in the network, among which yellow and medium orchid modules showed the most significant associations with LUAD. Then genes in these two modules were further chosen to evaluate their differential correlations. Finally, dozens of novel genes with opposite correlations including ATP13A4-AS1, HIGD1B, DAP3, and ISG20L2 were identified. Further biological and survival analyses highlighted their potential values in the diagnosis and treatment of

LUAD. Moreover, real-time qPCR confirmed the expression patterns of ATP13A4-AS1, HIGD1B, DAP3, and ISG20L2 in LUAD tissues and cell lines.

Conclusion: Our study provided new insights into the gene regulatory mechanisms during transition from normal to tumor, pioneering a network-based algorithm in the application of tumor etiology.

Keywords: WGCNA, differential correlation, switching mechanism, gene regulation, lung adenocarcinoma

INTRODUCTION

Lung cancer has been one of the most leading causes responsible for cancer mortality globally for several decades (Halliday et al., 2019). According to the Surveillance, Epidemiology, and End Results (SEER) Program, 228,820 new cases and 135,720 deaths of lung and bronchus cancer patients have been estimated in the United States in 2020 (Siegel et al., 2020). In China, estimated new cases and deaths from lung cancer were 730,000 and 610,000 in 2015, respectively, accounting for over 30% of the world's total (Chen et al., 2016). Based on histological subtypes, approximately 85% are of non-small cell lung cancer (NSCLC), of which lung adenocarcinoma (LUAD) is the most prevalent and attributed to more than 50% of all lung cancer cases (Oberndorfer and Mullauer, 2018). Despite that conventional therapies including surgery, radiotherapy, chemotherapy, immunotherapy, and targeted drugs have been applied to LUAD treatment, the 5-year survival rate for patients remains less than 15%, mainly due to lack of early detection and intervention (Herbst et al., 2018). In addition to environmental exposures and tobacco smoking, genetic susceptibility has been recognized as the most important risk factor associated with LUAD. Recently, a new conception called "oncogene addiction" referring to dependence of cancer cells on the activation of specific oncogenes has attracted increasing attention and underlined the significance of oncogenic drivers in LUAD (Calvayrac et al., 2017). At present, patients with LUAD harboring gene aberrations in EGFR (Thomas et al., 2019), KRAS (Kim et al., 2019), ALK (Shaw et al., 2019), BRAF (Dankner et al., 2018), HER2 (Mazieres et al., 2013), and MET (Drilon et al., 2017) benefit the most from targeted therapies, and those genes have been considered as oncogenic drivers in the field of thoracic oncology. However, on account of the dynamic mutational spectrum comprised of a vast number of hidden drivers, only a minority of patients archived improved outcome under current standard therapies (Oberndorfer and Mullauer, 2018). Thus, it is imperative to identify new risk genes for elucidating lung carcinogenetic mechanisms in order to guide researchers to develop new therapeutic strategies and physicians to tailor the treatment options.

Exponential advances in the high-throughput sequencing technologies and informatics have generated large-scale omics data which promotes a paradigm shift in the study of biomedical sciences and is available for deciphering molecular characteristics of oncogenesis that could be translated into clinical practice (Manzoni et al., 2018; Cheng, 2020). Among them, transcriptome data is quite informative because of the discovery of the significantly altered abundance of cellular components or pathways between disease and healthy tissues or two disease states (Kwon et al., 2019). In recent years, comparative analysis of transcriptome data has helped to discover prognostic markers and signatures by looking for differentially expressed genes in a variety of cancer types (Jiang et al., 2015; Zhang et al., 2020). However, a fact that could not be ignored is that most of the biological activities require an orchestrated action of multiple genes, whose dysregulation could lead to the occurrence of cancer. Therefore, gene correlation approaches have been intensively used for transcriptional profiling, providing preliminary steps toward genetic interaction networks and offering clues about the function of unknown genes.

Complementary to gene correlation analysis, changes in correlation patterns under different conditions, referred to as "differential correlations," are also attractive as for the reconstruction of genome architecture and identification of regulators or marker genes (Ideker and Krogan, 2012; Li et al., 2015). For example, one transcription factor defined as a master gene regulates a number of downstream targets. Also, these targets express in an ordered module where the regulatory mechanism is functional. However, in disease tissues where the regulatory mechanism is malfunctional, the gene expression module may be disordered or random. In this case, correlation changes can be detected by differential correlations rather than differential expression analysis. Recently, such alterations in network structures and variance in gene expression levels have been observed in cellular differentiation and cancer cell formation (Bockmayr et al., 2013; Ando et al., 2015), contributing to understanding the expression patterns between two states. However, differential correlations in LUAD are poorly determined, so it is imperative to unravel network dynamics that can be used to identify new candidate genes. Hitherto, an increasing number of systematic approaches have been developed for identifying such changes in network structures (Ideker and Krogan, 2012; Yu et al., 2015). Application of these methods to LUAD and adjacent normal tissues may reveal variance in gene expression levels during carcinogenesis.

Weighted gene co-expression network analysis (WGCNA) is a systematic method widely utilized in oncology research that aims at finding co-expressed genes through calculating gene connectivity (Chang et al., 2013; Yang et al., 2018). The soft thresholding of the Pearson correlation matrix is integrated into analysis for determining the connection strengths within gene pairs (Zhang and Horvath, 2005; Ghazalpour et al., 2006). Based on Fisher's z-test, R package DiffCorr is a useful tool in identifying differential correlations between disease states and providing a list of differentially correlated gene pairs

(Fukushima, 2013). In this study, we developed an *in silico* framework to identify key genes with differential correlations (**Figure 1**). First, we employed WGCNA to build a gene interaction network using the expression profile of LUAD from The Cancer Genome Atlas (TCGA). A total of 176 modules were detected in the network, among which yellow and medium orchid modules showed the most significant associations with LUAD. Then, we calculated differential correlations of genes in these two modules and identified significant differences between tumor and adjacent normal tissues using DiffCorr. Finally, dozens of new genes including ATP13A4-AS1, HIGD1B, DAP3, and ISG20L2, were identified and their expression patterns were confirmed by real-time qPCR. Further biological and survival analysis yielded their valuable effects on the progression of LUAD.

MATERIALS AND METHODS

LUAD RNA-Sequencing Datasets

The RNA-sequencing data of LUAD was downloaded from the TCGA database[1], including 468 tumor samples and 58 normal samples. As previously described, the gene expression levels were quantified as FPKM (fragments per kilobase per million mapped reads) using TopHat and HTSeq-count (Kim et al., 2013; Anders et al., 2015). The TCGA sample information was listed in **Supplementary Table 1**. Within the 468 tumor samples, 3 samples did not have stage information, 165 samples were in stage I, 247 samples were in stage II, 39 samples were in stage III, and 14 samples were in stage IV. The sample distribution also stressed the importance of identification of new genes for early diagnosis.

Co-expression Network Analysis

Hub gene screening and co-expression of gene pair detection were performed by an R package WGCNA, which was designed for analyzing a weighted correlation network (Langfelder and Horvath, 2008). The premise of the network construction was that elements in the gene co-expression matrix were the weighted values of the correlation coefficient between gene pairs. The selection criterion of the weight was to make the connections among genes conform to the scale-free network distribution; that is, for the number of connections i, the probability $p(i)$ was inverse to i^n. Practically, a weighted coefficient was selected to approximate the scale-free topology, which needs to satisfy the following condition: the log of the number of connected nodes $log(i)$ was negatively correlated with the log of occurrence probability of nodes $log(p(i))$ (the correlation coefficient should be at least 0.8). At the same time, the average connection degree of genes in different modules should be quite high. Associated genes were clustered based on dissimilarity of the unsigned topological overlap matrix (TOM). Finally, network modules and the genes within them were identified.

Functional Enrichment Analysis of Module Genes

R package clusterProfiler was used to perform functional enrichment analysis on clustered genes in yellow and medium orchid modules. A hypergeometric distribution test was applied to detect enrichment terms, and P values were adjusted by false discovery rate (FDR) method with a cutoff FDR < 0.05 (Yu et al., 2012).

Differential Correlation Analysis

R package DiffCorr was implemented for the identification and visualization of differential correlations in biological networks, and details were described in Fukushima (2013). Briefly, the DiffCorr package mainly contained three functions. First is calculation of differential correlations. The correlation coefficients for each of the two conditions (herein referred to tumor and normal), r_A and r_B, were transformed into Z_A and Z_B, respectively, based on Fisher's transformation: $Z = \frac{1}{2}\log\frac{1+r}{1-r}$. Differences between the two correlations could be tested using the equation $Z = \frac{Z_A - Z_B}{\sqrt{\frac{1}{n_A-3}+\frac{1}{n_B-3}}}$, where n_A and n_B represent the sample size for each biomolecule pair in each condition. Then, the local false discovery rate (lfdr) derived from the fdrtool package was used for controlling true estimates and identifying significant differential correlations. Second is identifying eigen-molecules. Eigen-molecules or "eigengenes" in the network were calculated based on the first principal component of a data matrix of a module which was extracted from a hierarchical cluster analysis. In addition to pair-wise differential correlations between molecules, differential correlations between modules were tested by using these eigen-molecule modules. Third is scaling and clustering. Different pretreatment methods with downstream correlation analyses were integrated, including auto-scaling, range scaling, Pareto scaling, vast scaling, level scaling, and power transformation.

Visualization and Functional Analysis

Construction and analysis of networks were carried out using STRING (11.0)[2] (Szklarczyk et al., 2019). Visualization of networks was realized using Cytoscape (3.6.0) (Doncheva et al., 2019).

The Survival Analysis of Key Genes With Differential Correlations

There were 468 LUAD patients with survival information. We performed survival analysis using the Cox proportional hazard regression model on these samples (Andersen and Gill, 1982). Each gene expression level was divided into two groups and repeated for 90 times based on the cutoff value from 5 to 95% of its expression level. A repeated log-rank test based on each cutoff value was processed, and a cutoff value with the lowest P-value was selected for subsequent univariate survival analysis. The Kaplan–Meier plot was used to describe the survival curves of these two groups of patients. The significance of the survival

[1]https://portal.gdc.cancer.gov/

[2]https://string-db.org/

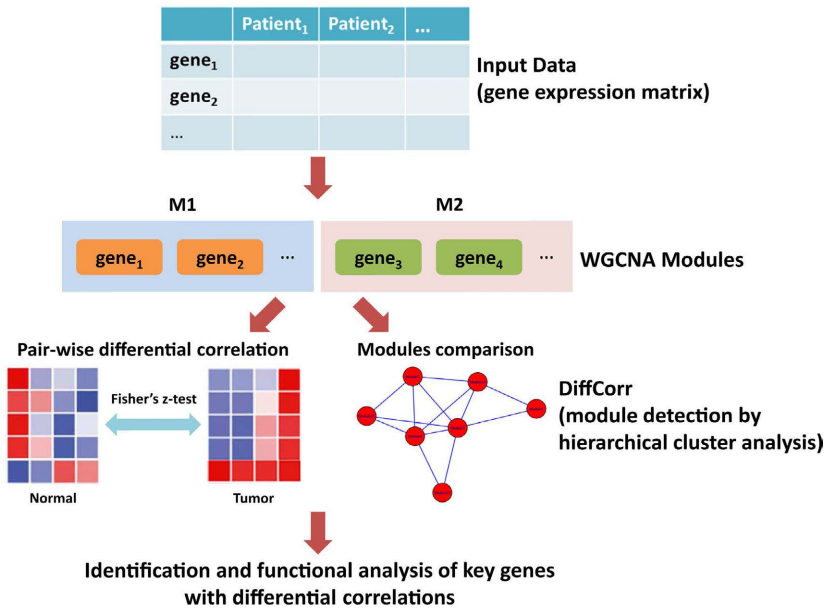


FIGURE 1 | The workflow of this study. First, the gene expression matrix of LUAD patients was obtained. Second, the WGCNA network was constructed and genes were clustered into modules. Third, differential correlations between genes were calculated and significant differences between tumor and adjacent normal tissues were identified using DiffCorr. Finally, functional analysis of key genes with differential correlations.

difference between these two patient groups was evaluated by the log-rank test *P* value. If the *P* value was less than 0.05, its survival was considered as significantly different. The R package survival was used to perform the survival analysis.

Cell Lines and Cell Culture

The LUAD cell lines (A549, SPCA1, PC9, H1299, and H1975) and the normal human bronchial epithelial cell line 16HBE were from the Cell Bank of Type Culture Collection of the Chinese Academy of Sciences, Shanghai Institute of Cell Biology (Shanghai, China). A549, H1975, and H1299 cells were grown in RPMI-1640 (Thermo, United States) containing 10% FBS, and 16HBE, SPCA1, and PC9 cells were grown in a DMEM medium (Thermo, United States) containing 10% fetal bovine serum (FBS). In addition, all cell media were supplemented with penicillin (100 U/mL) and streptomycin (100 U/mL) at 37°C in a 5% CO_2 incubator.

LUAD Patients and Tissue Specimens

A total of 15 pairs of LUAD tumor and corresponding adjacent normal tissues were collected from the Third Affiliated Hospital of Soochow University between October 2019 and June 2020. Samples were snap-frozen and stored at −80°C until use in real-time qPCR (RT-qPCR) experiments. In addition, we conducted immunohistochemical staining of formalin-fixed paraffin-embedded LUAD patients and normal control samples. The Research Ethics Committee of the Third Affiliated Hospital of Soochow University approved this study, which was consistent with the Declaration of Helsinki. All patients provided written informed consent.

RNA Isolation, cDNA Synthesis, and RT-qPCR

We isolated total RNA with TRIzol reagent (Thermo, United States). First-strand cDNA was synthesized from 1 μg total RNA using a ReverTra Ace qPCR RT Kit (Toyobo, Osaka, Japan). The Fast SYBR Green Master Mix was used for RT-qPCR (Applied Biosystems Inc., CA, United States). The cycling conditions were 30 s of polymerase activation at 95°C followed by 40 cycles at 95°C for 5 s and 60°C for 30 s. GAPDH was used as an internal loading control. The relative level was calculated by the relative quantification 2-ΔΔCT method. All primer sequences were listed in **Supplementary Table 2**.

Statistical Analysis

Student's *t*-test was applied to identify genes differentially expressed between normal and tumor samples. *P* values were adjusted by the Benjamini–Hochberg method (Hochberg and Benjamini, 1990). Differentially expressed genes were defined as adjusted *P* value < 0.05. Fisher's *z*-test was employed to evaluate differential correlations of gene pairs between normal and tumor samples. Moreover, lfdr < 0.05 was defined as significant differential correlations.

RESULTS

Construction of the Co-expression Network

The weighted gene co-expression network was constructed from 58,387 coding and noncoding RNAs through the WGCNA

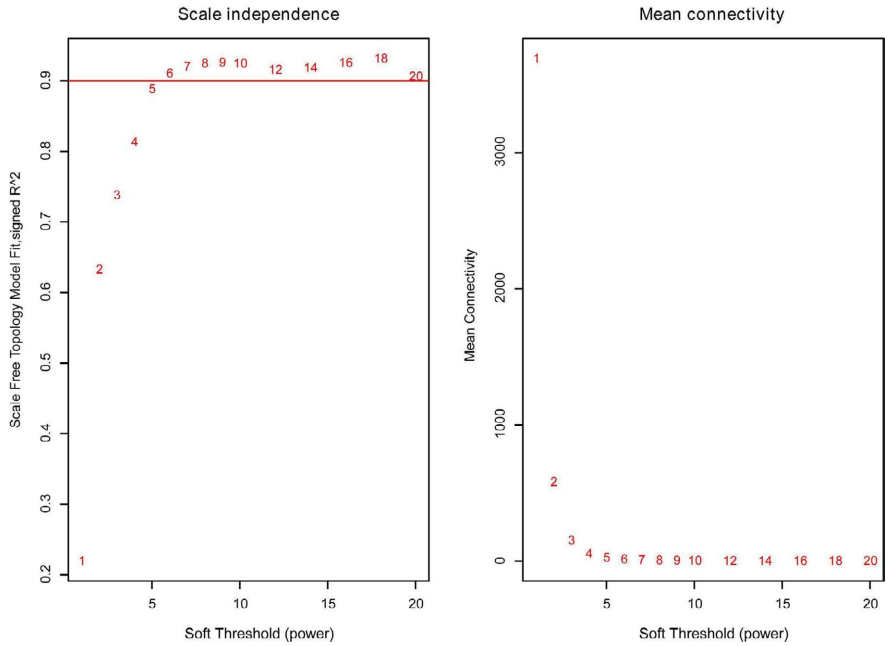


FIGURE 2 | The relationship between soft threshold (power) and network properties. Left panel: The relationship between soft-threshold (power) and scale-free topology. Right panel: The relationship between soft threshold (power) and mean connectivity. When the soft threshold (power) was six, the scale-free topology (R^2) was 0.91 and mean connectivity became stable. Therefore, we set the soft threshold (power) to be six.

approach. Here, the soft-thresholding power was set to be six to satisfy the scale-free topology of the network (**Figure 2**), in which R^2 was used to check how well the network fit the scale freeness. When the soft-thresholding power was set to be six, the R^2 was 0.91. Furthermore, we detected 176 modules in this network, whose relationship was shown in a cluster dendrogram (**Figure 3A**). The number of members in different modules varied widely. The members of each module were listed in **Supplementary Table 3**. Besides the gray module which comprised many unclassified members, the turquoise module contained a maximum of 4,594 genes, while a minimum of 30 genes were included in the dark sea green module.

Each module represented a group of genes with similar expression profiles across samples. Next, we quantified module-trait associations (**Supplementary Figure 1**), among which the yellow and medium orchid modules showed the most significant associations with LUAD. The corresponding correlation coefficients of yellow and medium orchid modules were 0.83 ($P = 410^{-133}$) and −0.58 ($P = 510^{-49}$), respectively. Clearly, Gene Significance (GS) and Module Membership (MM) analysis illustrated that genes highly significantly associated with LUAD were also the most important elements of modules associated with LUAD (**Supplementary Figure 2**). We also performed Gene Ontology (GO) enrichment analysis of these two modules. As presented in **Figure 3B**, genes in the yellow module were significantly enriched in the negative regulation of growth and vasculogenesis, while genes in the medium orchid module were enriched in the cellular metabolic process (geranyl diphosphate metabolic process, geranyl diphosphate biosynthetic process, farnesyl diphosphate biosynthetic process, and farnesyl diphosphate metabolic process) with an adjusted *P* value smaller than 0.05, conferring the importance of these biological functions on LUAD development.

Identification of Differential Correlations

The genes in the yellow and medium orchid modules were further chosen to evaluate their differential correlations. Using R package DiffCorr, these genes were grouped based on their expression patterns in each subtype (normal or tumor) using the *cluster.molecule* function. We used the one-correlation coefficient as a distance measure (the cutoff of the coefficient was 0.6) according to the *cutree* function. The *get.eigen.molecule* and *get.eigen.molecule.graph* functions were used for visualization of the module network (**Figure 4**). The *comp.2.cc.fdr* function provided the resulting pair-wise differential correlations in the yellow (**Supplementary Table 4**) and medium orchid modules (**Supplementary Table 5**).

The DiffCorr package also detected oppositely correlated pairs where, for example, two genes exhibited a positive correlation in normal samples and a negative correlation in tumor samples, or *vice versa*, a condition referred to as a "switching mechanism" (Kayano et al., 2011). These switched gene pairs, which especially show a differential expression at the same time, were worth noticing for their critical roles in understanding cellular functions in the development of LUAD. In total, we obtained 42 oppositely correlated gene pairs with a differential expression simultaneously from the yellow module and 62 from the medium orchid module (**Supplementary Table 6**). Their interaction networks are shown in **Figure 5**. The top 10 significant switching mechanisms of gene expression between normal and tumor

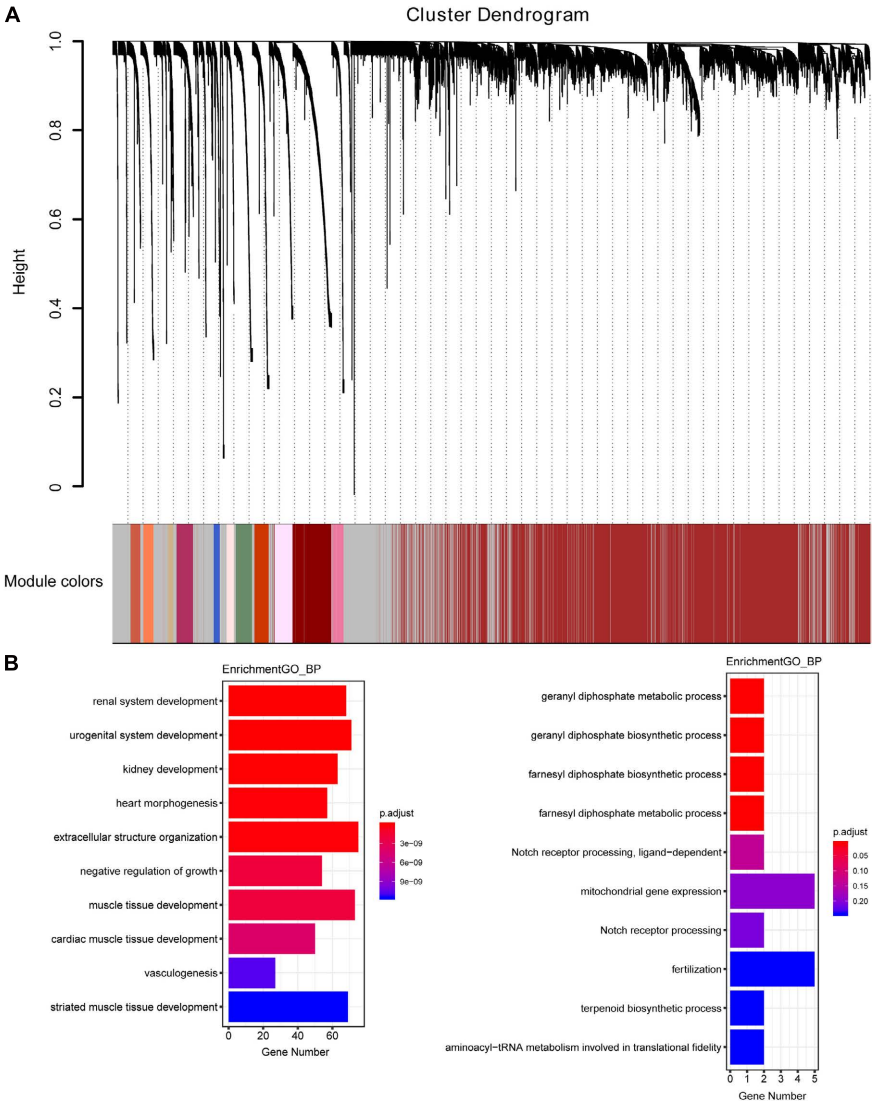


FIGURE 3 | The cluster dendrogram of the WGCNA co-expression network and functional enrichment of module genes. **(A)** Total genes were clustered in 176 modules. Each module was marked with one color. Except for the gray module, which included many unclassified members, the turquoise module contained a maximum of 4,594 members, while a minimum of 30 members were included in the dark sea green module. **(B)** GO analysis showed the top 10 enriched biological processes in the yellow (left panel) and medium orchid (right panel) modules.

samples from the yellow and medium orchid modules are shown in **Table 1**.

Functional Analysis of Key Genes With Differential Correlations

Next, we focused on several key genes that acted as master regulators for their occupying most connections with other genes. In the yellow module, DUOX1, DUOXA1, HIGD1B, and ATP13A4-AS1 stood out in the network (**Figure 5A**). ATP13A4 antisense RNA 1 (ATP13A4-AS1) is an antisense long noncoding RNA (lncRNA) derived from ATP13A4 with an unknown function in cancer. Here, we can infer the roles of ATP13A4-AS1 from its interacted genes advanced glycosylation end-product specific receptor (AGER) and angiopoietin-like 7 (ANGPTL7) (**Figure 5A**). As a member of the immunoglobulin superfamily, cell surface receptor AGER is involved in multiple inflammatory responses whose abnormal expression has been reported to be closely associated with carcinogenesis (Bongarzone et al., 2017). Accumulating evidence has shown the downregulation of AGER in lung cancer, leading to enhanced proliferation, invasion and migration abilities, and decreased apoptosis of cells (Zhang et al., 2018; Wang et al., 2020). In addition, genetic polymorphisms of AGER have been reported to increase risks of lung cancer and breast cancer (Yin et al., 2015; Liu et al., 2019). Angiopoietin-like 7 (ANGPTL7) belongs to a family of secreted angiopoietin-like proteins and plays vital roles in the modulation of hematopoietic stem cell maintenance, angiogenesis, and lipid

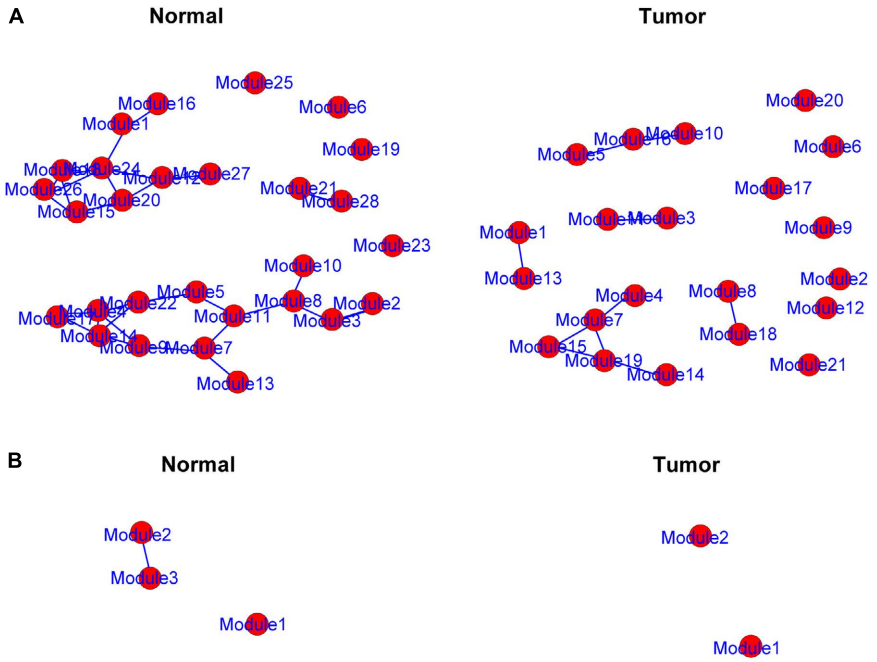


FIGURE 4 | Representation of the module networks. Images of yellow **(A)** and medium orchid **(B)** module networks from the TCGA dataset were shown. Each node represented one module, and each edge represented the module correlation.

metabolism (Zhang et al., 2006; Hato et al., 2008). Prior studies have revealed striking ANGPTL7 underexpression in various cancers such as colorectal cancer and breast cancer. Besides, upregulation of ANGPTL7 has been found in cancer cells after myeloid cell depletion, which affected liver metastasis by diminishing cell growth and vascular density, implying that ANGPTL7 could act as a mediator of metastatic progression and as a promising intervention target (Lim et al., 2015). As AGER and ANGPTL7 both show strong links to cancer, it is plausible that their interacted gene ATP13A4-AS1 also engages in the progression of lung cancer by exerting similar impacts on cellular activities. Furthermore, the Kaplan–Meier plot exhibited that lower levels of ATP13A4-AS1 correlated with poor patient outcome (**Figure 6A**), corresponding to its lower expression in LUAD than normal samples confirmed by our RT-qPCR (**Figure 7A**). Hence, our study revealed, for the first time, the roles of ATP13A4-AS1 in carcinogenesis, providing experimental clues for further investigations.

Dual oxidase 1 (DUOX1), an oxidant-generating enzyme within the airway epithelium, has been previously reported to take part in innate airway host defense and epithelial homeostasis, the activation of which stimulated cell migration dependent on Src family tyrosine kinases and epidermal growth factor receptor (EGFR) signaling pathways (Yoo et al., 2012; Gorissen et al., 2013; Sham et al., 2013; Hristova et al., 2016). DUOXA1, as a maturation factor of DUOX1, is transcriptionally and functionally linked to DUOX1 (Grasberger and Refetoff, 2006). Recent evidence has indicated that DUOX1 and DUOXA1 were frequently silenced due to promoter hypermethylation in various epithelial cancers including lung cancer (Luxen et al., 2008; Ling et al., 2014; Little et al., 2016). Consistent with pan-cancer analysis, DUOX1 and DUOXA1 were also downregulated in LUAD from our TCGA dataset, proving their potential tumor suppression roles.

HIG1 domain family member 1B (HIGD1B) is localized to the cell membrane whose expression is induced by hypoxia and glucose deprivation (Denko et al., 2000). Moreover, an increased expression of HIGD1B has been observed in cervical cancer and plurihormonal pituitary adenomas (Denko et al., 2000; Jiang et al., 2012). In contrast with previous research, HIGD1B showed lower expression levels in LUAD compared with normal samples in TCGA datasets, which was also observed in our LUAD patients and cell lines (**Figure 7B**), probably due to different cancer characteristics, implying its dual function in carcinogenesis. Moreover, its interaction with DUOX1 and DUOXA1 emphasized similar contributions to inhibit the malignant progress of cancer. The survival analysis also demonstrated that the low expression was associated with high risk (**Figure 6B**). Thus, our study expanded the roles of HIGD1B in cancer biology which need further validations.

In the medium orchid module, PIP5K1A, CRTC2, DAP3, and ISG20L2 deserved much attention for their central positions in the network (**Figure 5B**). As a lipid kinase, the basic function of PIP5K1A is to phosphorylate PI4P to synthesize important signaling phospholipid $PI(4,5)P_2$, which serves as the substrate for phosphoinositide 3-kinase (PI3K) for conversion into $PI(3,4,5)P_3$ to promote cell proliferation and survival (Adhikari and Counter, 2018). Intriguingly, PIP5K1A has been recently shown to interact directly with mutant KRAS and TP53, which were the most prevalent drivers in lung cancer,

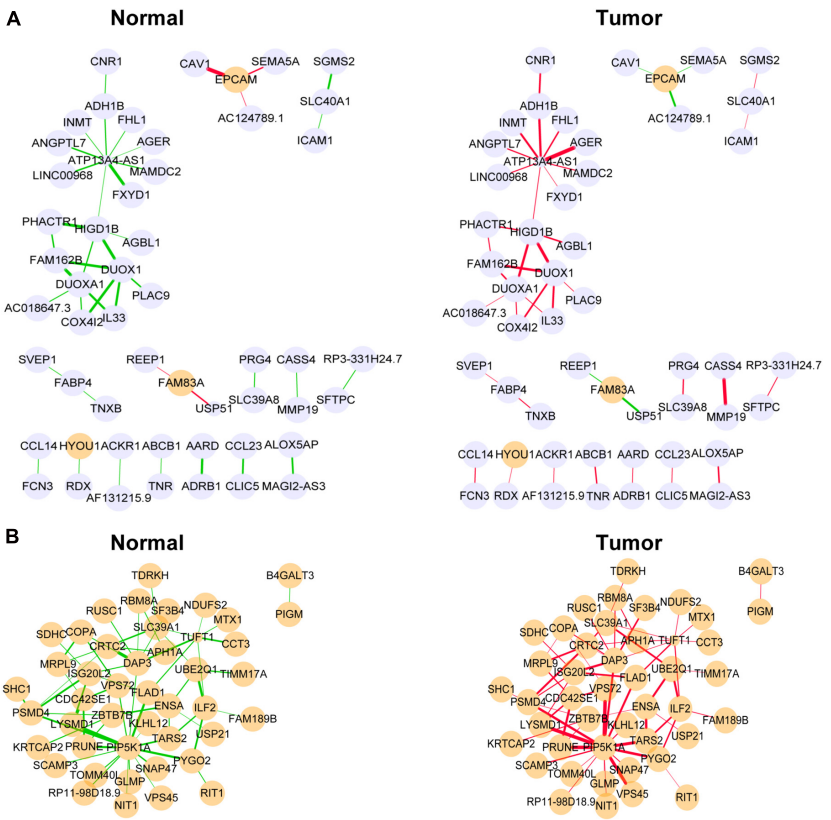


FIGURE 5 | Differentially co-expressed gene networks in the yellow **(A)** and medium orchid **(B)** modules from the TCGA dataset. Each node represented a gene, with lavender-filled color denoting downregulation and orange-filled color upregulation. The larger size of node represented smaller adjusted *P*-value. The edge represented connection between two genes. The green edge represented negative correlation and red positive correlation. The thicker part of the edge represented a stronger correlation coefficient.

and facilitated downstream oncogenic signaling (Adhikari and Counter, 2018; Choi et al., 2019).

CRTC2, belonging to the CREB-regulated transcription coactivator (CRTC) family, can bind the leucine zipper DNA-binding region of CREB which results in the enhancement of CREB transcriptional activity (Koo et al., 2005; Dentin et al., 2008). Recent studies have identified novel mutations and a high expression of CRTC2 in non-small cell lung cancer patients, leading to reinforced migration and invasion abilities of cancer cells (Shi et al., 2018; Rodón et al., 2019). Current evidence has pointed out oncogenic roles of CRTC2 in lung cancer, consistent with its elevated expression levels in our study.

As a molecule involved in controlling apoptosis and anoikis, death-associated protein 3 (DAP3) has been highly implicated in the context of carcinogenesis. However, controversy exists with regard to its roles in human cancer (Wazir et al., 2015). Increased expression levels of DAP3 have been observed in invasive glioblastoma as well as glioma cells with induced migratory phenotype (Mariani et al., 2001), while an inverse association between DAP3 expression and clinical outcome has been found in breast cancer, corresponding to the pro-apoptotic function of DAP3 (Wazir et al., 2012). A recent study has identified differentially expressed DAP3 for classification between LUAD and normal samples (Hsu et al., 2015), indicating its crucial roles in carcinogenesis. Consistently, our Kaplan–Meier analysis showed better survival in the low transcription group approaching significance (**Figure 6C**), supporting tumor promotion influences of DAP3 on LUAD. At the same time, our experiments also witnessed the upregulation of DAP3 in LUAD patients and cell lines (**Figure 7C**). Therefore, our study illustrated that, for the first time, DAP3 occupied the core position in the regulatory network of LUAD, explaining its mechanism of action to some extent.

Interferon-stimulated 20-kDa exonuclease-like 2 (ISG20L2) is a novel vertebrate nucleolar exoribonuclease and involved in ribosome biogenesis (Couté et al., 2008; Simabuco et al., 2012). As one of target genes regulated by miR-139-3p, ISG20L2 has recently been related to hepatocellular carcinoma prognosis (Zhu et al., 2019), consistent with an immunogenomic landscape study which adds the roles of ISG20L2 in reflecting levels of infiltration in diverse immune cells (Xiao et al., 2021). Additionally, high expression of ISG20L2 has been found in several cancers and affected cancer occurrence in a variety of aspects (Xu et al., 2020; Xiao et al., 2021). Nevertheless, there is scant literature with regard to its role in LUAD. Here, we proposed for the first time the roles of ISG20L2 in promoting carcinogenesis and

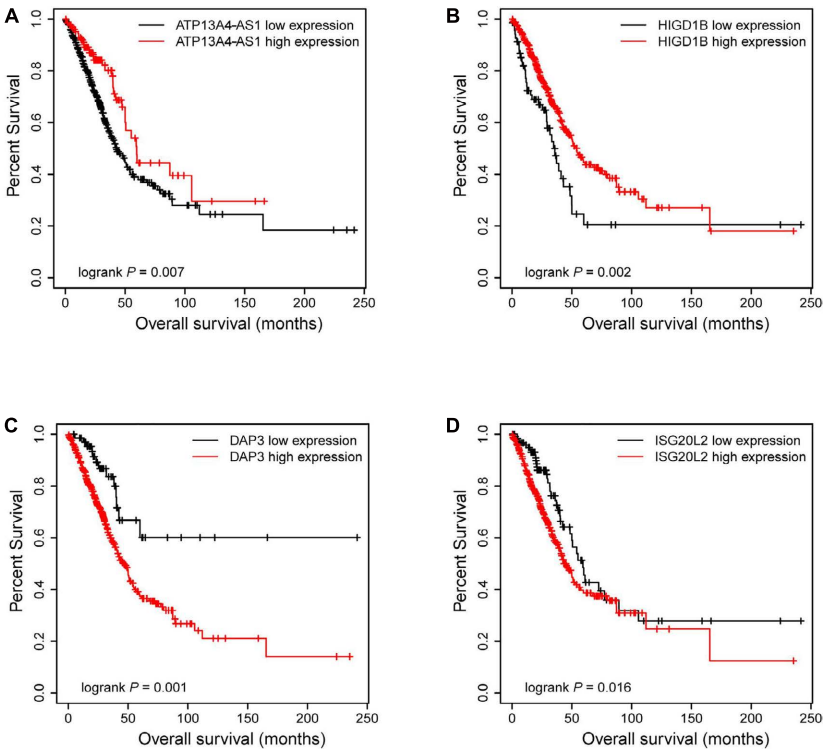


FIGURE 6 | The Kaplan–Meier plot of ATP13A4-AS1, HIGD1B, DAP3, and ISG20L2. The low expressions of ATP13A4-AS1 **(A)** and HIGD1B **(B)** were associated with high risk. The high expressions of DAP3 **(C)** and ISG20L2 **(D)** were associated with high risk.

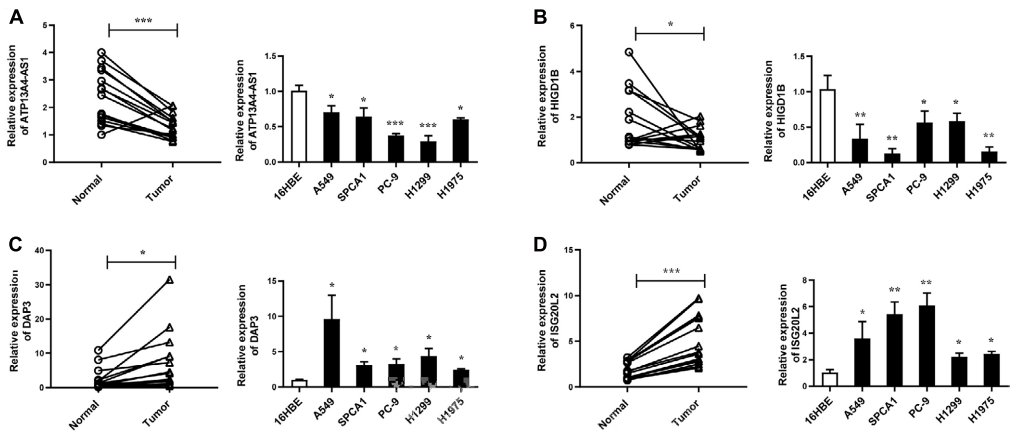


FIGURE 7 | The expression levels of ATP13A4-AS1 **(A)**, HIGD1B **(B)**, DAP3 **(C)**, and ISG20L2 **(D)** in LUAD tissues and cell lines detected by RT-qPCR. Human bronchial epithelial cell line: 16BE. LUAD cell lines: A549, SPCA1, PC9, H1299, and H1975. Data were means ± SEM. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Experiments were repeated three times.

identified its interacted genes in the regulatory network which contributed to understanding further functional mechanisms. Accordingly, survival time was significantly higher in patients with low ISG20L2 expression, compared with the high ISG20L2 expression group (**Figure 6D**). Meanwhile, remarkable elevated levels of ISG20L2 were also verified by RT-qPCR in LUAD patients and cell lines (**Figure 7D**).

DISCUSSION

Although multiple studies have continuously explored gene interaction networks which were constructed from a series of correlated gene pairs, little is known about the differential correlations between normal and tumor status which contributed to elucidating deep mechanisms of gene regulation and

TABLE 1 | Top 10 correlated gene pairs changed to the opposite direction from the yellow and medium orchid modules between normal and LUAD samples.

Molecule X	Molecule Y	r1 (normal)	r2 (tumor)	lfdr[a] (difference)	Module color
HIGD1B	DUOX1	−0.58	0.53	0	Yellow
FAM162B	DUOX1	−0.56	0.51	0	Yellow
EPCAM	CAV1	0.65	−0.40	0	Yellow
FAM162B	DUOXA1	−0.59	0.46	4.67E-12	Yellow
PHACTR1	HIGD1B	−0.58	0.45	1.47E-11	Yellow
DUOX1	COX4I2	−0.55	0.47	3.16E-11	Yellow
IL33	DUOX1	−0.53	0.48	8.28E-11	Yellow
FXYD1	ATP13A4-AS1	−0.58	0.41	1.29E-10	Yellow
MMP19	CASS4	−0.43	0.55	2.30E-10	Yellow
ATP13A4-AS1	AGER	−0.40	0.58	2.51E-10	Yellow
VPS72	PIP5K1A	−0.47	0.73	0	Mediumorchid
VPS45	PIP5K1A	−0.41	0.69	0	Mediumorchid
UBE2Q1	ILF2	−0.60	0.58	0	Mediumorchid
TARS2	PIP5K1A	−0.54	0.66	0	Mediumorchid
TARS2	ENSA	−0.46	0.62	0	Mediumorchid
PYGO2	PIP5K1A	−0.59	0.62	0	Mediumorchid
PSMD4	PIP5K1A	−0.55	0.60	0	Mediumorchid
PRUNE	PIP5K1A	−0.58	0.70	0	Mediumorchid
PRUNE	CDC42SE1	−0.55	0.59	0	Mediumorchid
PIP5K1A	LYSMD1	−0.74	0.71	0	Mediumorchid

[a]*Local FDR.*

identifying master genes. Benefiting from the availability of high-throughput data and different kinds of algorithms, detecting correlation changes became feasible. In this study, we developed an *in silico* framework to identify key genes with differential correlations in LUAD. Both candidate genes and their targets from gene regulatory networks could be further used for experimental investigations of biological functions in order to guide the diagnosis and treatment of patients. As transcriptomic data only represents a single layer of genome and is not sufficient to depict the whole-cell atlas, multidimensional molecular data including proteomic and metabolomic data as well as genome-wide association studies are required to be integrated for elaborating transition mechanisms from normal to tumor. Meanwhile, a statistical evaluation should be more strongly emphasized in future studies of differential correlations.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Third Affiliated Hospital of Soochow University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YoZ conceived the study. YiZ and JL conducted the database search and analysis. BX, TH, and XZ designed the algorithm and network construction. YL and HD conducted the experimental validations. YoZ and JJ wrote the manuscript. All authors contributed to biological analysis, interpretation of the results, read, and approved the final version of the manuscript.

FUNDING

This project was supported by grants from the National Natural Science Foundation of China (31701111, 31701151, and 81902386), Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), National Key R&D Program of China (2018YFC0910403 and 2017YFC1201200), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), Young Talent Development Plan of Changzhou Health Commission (CZQM2020008), and Youth Talent Science and Technology Project of Changzhou Health Commission (QN202014).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.675438/full#supplementary-material

Supplementary Figure 1 | Module-trait associations. Each row corresponded to a module eigengene, column to a trait. Each cell contained the corresponding

correlation and *P*-value. The table was color-coded by correlation according to the color legend.

Supplementary Figure 2 | A scatterplot of Gene Significance (GS) for Disease vs. Module Membership (MM) in the yellow **(A)** and mediumorchid **(B)** modules. There was a highly significant correlation between GS and MM in the module.

Supplementary Table 1 | The TCGA sample IDs and their clinical information.

Supplementary Table 2 | Primer sequences of GAPDH, ATP13A4-AS1, HIGD1B, DAP3, and ISG20L2.

Supplementary Table 3 | The members of each module.

Supplementary Table 4 | Pair-wise differential correlations in the yellow module.

Supplementary Table 5 | Pair-wise differential correlations in the mediumorchid module.

Supplementary Table 6 | Oppositely correlated gene pairs with differentially expression simultaneously from the yellow and mediumorchid modules.

REFERENCES

Adhikari, H., and Counter, C. M. (2018). Interrogating the protein interactomes of RAS isoforms identifies PIP5K1A as a KRAS-specific vulnerability. *Nat. Commun.* 9:3646. doi: 10.1038/s41467-018-05692-6

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Andersen, P. K., and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* 10, 1100–1120.

Ando, T., Kato, R., and Honda, H. (2015). Differential variability and correlation of gene expression identifies key genes involved in neuronal differentiation. *BMC Syst. Biol.* 9:82. doi: 10.1186/s12918-015-0231-6

Bockmayr, M., Klauschen, F., Györffy, B., Denkert, C., and Budczies, J. (2013). New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Syst. Biol.* 7:78. doi: 10.1186/1752-0509-7-78

Bongarzone, S., Savickas, V., Luzi, F., and Gee, A. D. (2017). Targeting the receptor for advanced glycation endproducts (RAGE): a medicinal chemistry perspective. *J. Med. Chem.* 60, 7213–7232. doi: 10.1021/acs.jmedchem.7b00058

Calvayrac, O., Pradines, A., Pons, E., Mazieres, J., and Guibert, N. (2017). Molecular biomarkers for lung adenocarcinoma. *Eur. Respir. J.* 49:1601734. doi: 10.1183/13993003.01734-2016

Chang, X., Shi, L., Gao, F., Russin, J., Zeng, L., He, S., et al. (2013). Genomic and transcriptome analysis revealing an oncogenic functional module in meningiomas. *Neurosurg. Focus* 35:E3. doi: 10.3171/2013.10.focus13326

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer statistics in China, 2015. *CA Cancer J. Clin.* 66, 115–132. doi: 10.3322/caac.21338

Cheng, L. (2020). Omics data and artificial intelligence: new challenges for gene therapy. *Curr. Gene Ther.* 20:1. doi: 10.2174/1566523220012006041500 41

Choi, S., Chen, M., Cryns, V. L., and Anderson, R. A. (2019). A nuclear phosphoinositide kinase complex regulates p53. *Nat. Cell Biol.* 21, 462–475. doi: 10.1038/s41556-019-0297-2

Couté, Y., Kindbeiter, K., Belin, S., Dieckmann, R., Duret, L., Bezin, L., et al. (2008). ISG20L2, a novel vertebrate nucleolar exoribonuclease involved in ribosome biogenesis. *Mol. Cell. Proteomics* 7, 546–559. doi: 10.1074/mcp.M700510-MCP200

Dankner, M., Rose, A. A. N., Rajkumar, S., Siegel, P. M., and Watson, I. R. (2018). Classifying BRAF alterations in cancer: new rational therapeutic strategies for actionable mutations. *Oncogene* 37, 3183–3199. doi: 10.1038/s41388-018-0171-x

Denko, N., Schindler, C., Koong, A., Laderoute, K., Green, C., and Giaccia, A. (2000). Epigenetic regulation of gene expression in cervical cancer cells by the tumor microenvironment. *Clin. Cancer Res.* 6, 480–487.

Dentin, R., Hedrick, S., Xie, J., Yates, J. III, and Montminy, M. (2008). Hepatic glucose sensing via the CREB coactivator CRTC2. *Science* 319, 1402–1405. doi: 10.1126/science.1151363

Doncheva, N. T., Morris, J. H., Gorodkin, J., and Jensen, L. J. (2019). Cytoscape StringApp: network analysis and visualization of proteomics data. *J. Proteome Res.* 18, 623–632. doi: 10.1021/acs.jproteome.8b00702

Drilon, A., Cappuzzo, F., Ou, S. I., and Camidge, D. R. (2017). Targeting MET in lung cancer: will expectations finally be MET? *J. Thorac. Oncol.* 12, 15–26. doi: 10.1016/j.jtho.2016.10.014

Fukushima, A. (2013). DiffCorr: an R package to analyze and visualize differential correlations in biological networks. *Gene* 518, 209–214. doi: 10.1016/j.gene.2012.11.028

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* 2:e130. doi: 10.1371/journal.pgen.0020130

Gorissen, S. H., Hristova, M., Habibovic, A., Sipsey, L. M., Spiess, P. C., Janssen-Heininger, Y. M., et al. (2013). Dual oxidase-1 is required for airway epithelial cell migration and bronchiolar reepithelialization after injury. *Am. J. Respir. Cell. Mol. Biol.* 48, 337–345. doi: 10.1165/rcmb.2012-0393OC

Grasberger, H., and Refetoff, S. (2006). Identification of the maturation factor for dual oxidase. Evolution of an eukaryotic operon equivalent. *J. Biol. Chem.* 281, 18269–18272. doi: 10.1074/jbc.C600095200

Halliday, P. R., Blakely, C. M., and Bivona, T. G. (2019). Emerging targeted therapies for the treatment of non-small cell lung cancer. *Curr. Oncol. Rep.* 21:21. doi: 10.1007/s11912-019-0770-x

Hato, T., Tabata, M., and Oike, Y. (2008). The role of angiopoietin-like proteins in angiogenesis and metabolism. *Trends Cardiovasc. Med.* 18, 6–14. doi: 10.1016/j.tcm.2007.10.003

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature* 553:446. doi: 10.1038/nature25183

Hochberg, Y., and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat. Med.* 9, 811–818. doi: 10.1002/sim.4780090710

Hristova, M., Habibovic, A., Veith, C., Janssen-Heininger, Y. M., Dixon, A. E., Geiszt, M., et al. (2016). Airway epithelial dual oxidase 1 mediates allergen-induced IL-33 secretion and activation of type 2 immune responses. *J. Allergy Clin. Immunol.* 137, 1545–1556. doi: 10.1016/j.jaci.2015.10.003

Hsu, M. K., Wu, I. C., Cheng, C. C., Su, J. L., Hsieh, C. H., Lin, Y. S., et al. (2015). Triple-layer dissection of the lung adenocarcinoma transcriptome: regulation at the gene, transcript, and exon levels. *Oncotarget* 6, 28755–28773. doi: 10.18632/oncotarget.4810

Ideker, T., and Krogan, N. J. (2012). Differential network biology. *Mol. Syst. Biol.* 8:565. doi: 10.1038/msb.2011.99

Jiang, Z., Zhou, X., Li, R., Michal, J. J., Zhang, S., Dodson, M. V., et al. (2015). Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. *Cell. Mol. Life Sci.* 72, 3425–3439. doi: 10.1007/s00018-015-1934-y

Jiang, Z., Gui, S., and Zhang, Y. (2012). Analysis of differential gene expression in plurihormonal pituitary adenomas using bead-based fiber-optic arrays. *J. Neurooncol.* 108, 341–348. doi: 10.1007/s11060-011-0792-1

Kayano, M., Takigawa, I., Shiga, M., Tsuda, K., and Mamitsuka, H. (2011). ROS-DET: robust detector of switching mechanisms in gene expression. *Nucleic Acids Res.* 39:e74. doi: 10.1093/nar/gkr130

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Kim, Y. J., Baek, D. S., Lee, S., Park, D., Kang, H. N., Cho, B. C., et al. (2019). Dual-targeting of EGFR and Neuropilin-1 attenuates resistance to EGFR-targeted antibody therapy in KRAS-mutant non-small cell lung cancer. *Cancer Lett.* 466, 23–34. doi: 10.1016/j.canlet.2019.09.00

Koo, S. H., Flechner, L., Qi, L., Zhang, X., Screaton, R. A., Jeffries, S., et al. (2005). The CREB coactivator TORC2 is a key regulator of fasting glucose metabolism. *Nature* 437, 1109–1111. doi: 10.1038/nature03967

Kwon, O. S., Kim, W., Cha, H. J., and Lee, H. (2019). In silico drug repositioning: from large-scale transcriptome data to therapeutics. *Arch. Pharm. Res.* 42, 879–889. doi: 10.1007/s12272-019-01176-3

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Li, Y., Jin, S., Lei, L., Pan, Z., and Zou, X. (2015). Deciphering deterioration mechanisms of complex diseases based on the construction of dynamic networks and systems analysis. *Sci. Rep.* 5:9283. doi: 10.1038/srep09283

Lim, S. Y., Gordon-Weeks, A., Allen, D., Kersemans, V., Beech, J., Smart, S., et al. (2015). Cd11b(+) myeloid cells support hepatic metastasis through down-regulation of angiopoietin-like 7 in cancer cells. *Hepatology* 62, 521–533. doi: 10.1002/hep.27838

Ling, Q., Shi, W., Huang, C., Zheng, J., Cheng, Q., Yu, K., et al. (2014). Epigenetic silencing of dual oxidase 1 by promoter hypermethylation in human hepatocellular carcinoma. *Am. J. Cancer Res.* 4, 508–517.

Little, A. C., Sham, D., Hristova, M., Danyal, K., Heppner, D. E., Bauer, R. A., et al. (2016). DUOX1 silencing in lung cancer promotes EMT, cancer stem cell characteristics and invasive properties. *Oncogenesis* 5:e261. doi: 10.1038/oncsis.2016.61

Liu, W., Ouyang, S., Zhou, Z., Wang, M., Wang, T., Qi, Y., et al. (2019). Identification of genes associated with cancer progression and prognosis in lung adenocarcinoma: Analyses based on microarray from Oncomine and The Cancer Genome Atlas databases. *Mol. Genet. Genomic Med.* 7:e00528. doi: 10.1002/mgg3.528

Luxen, S., Belinsky, S. A., and Knaus, U. G. (2008). Silencing of DUOX NADPH oxidases by promoter hypermethylation in lung cancer. *Cancer Res.* 68, 1037–1045. doi: 10.1158/0008-5472.can-07-5782

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., et al. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* 19, 286–302. doi: 10.1093/bib/bbw114

Mariani, L., Beaudry, C., McDonough, W. S., Hoelzinger, D. B., Kaczmarek, E., Ponce, F., et al. (2001). Death-associated protein 3 (Dap-3) is overexpressed in invasive glioblastoma cells in vivo and in glioma cell lines with induced motility phenotype in vitro. *Clin. Cancer Res.* 7, 2480–2489.

Mazieres, J., Peters, S., Lepage, B., Cortot, A. B., Barlesi, F., Beau-Faller, M., et al. (2013). Lung cancer that harbors an HER2 mutation: epidemiologic characteristics and therapeutic perspectives. *J. Clin. Oncol.* 31, 1997–2003. doi: 10.1200/JCO.2012.45.6095

Oberndorfer, F., and Mullauer, L. (2018). Molecular pathology of lung cancer: current status and perspectives. *Curr. Opin. Oncol.* 30, 69–76. doi: 10.1097/CCO.0000000000000429

Rodón, L., Svensson, R. U., Wiater, E., Chun, M. G. H., Tsai, W. W., Eichner, L. J., et al. (2019). The CREB coactivator CRTC2 promotes oncogenesis in LKB1-mutant non-small cell lung cancer. *Sci. Adv.* 5:eaaw6455. doi: 10.1126/sciadv.aaw6455

Sham, D., Wesley, U. V., Hristova, M., and van der Vliet, A. (2013). ATP-mediated transactivation of the epidermal growth factor receptor in airway epithelial cells involves DUOX1-dependent oxidation of Src and ADAM17. *PLoS One* 8:e54391. doi: 10.1371/journal.pone.0054391

Shaw, A. T., Solomon, B. J., Besse, B., Bauer, T. M., Lin, C. C., Soo, R. A., et al. (2019). ALK resistance mutations and efficacy of lorlatinib in advanced anaplastic lymphoma kinase-positive non-small-cell lung cancer. *J. Clin. Oncol.* 37, 1370–1379. doi: 10.1200/JCO.18.02236

Shi, S., Luo, W., Zhang, R., Wang, C., Zheng, Y., Song, Y., et al. (2018). CRTC2 promotes non-small cell lung cancer A549 migration and invasion in vitro. *Thorac. Cancer* 9, 136–141. doi: 10.1111/1759-7714.12550

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA A Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590

Simabuco, F. M., Morello, L. G., Aragão, A. Z., Paes Leme, A. F., and Zanchin, N. I. (2012). Proteomic characterization of the human FTSJ3 preribosomal complexes. *J. Proteome Res.* 11, 3112–3126. doi: 10.1021/pr201106n

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Thomas, R., Srivastava, S., Katreddy, R. R., Sobieski, J., and Weihua, Z. (2019). Kinase-Inactivated EGFR is required for the survival of wild-type EGFR-Expressing cancer cells treated with tyrosine kinase inhibitors. *Int. J. Mol. Sci.* 20:2515. doi: 10.3390/ijms20102515

Wang, Q., Zhu, W., Xiao, G., Ding, M., Chang, J., and Liao, H. (2020). Effect of AGER on the biological behavior of non-small cell lung cancer H1299 cells. *Mol. Med. Rep.* 22, 810–818. doi: 10.3892/mmr.2020.11176

Wazir, U., Orakzai, M. M., Khanzada, Z. S., Jiang, W. G., Sharma, A. K., Kasem, A., et al. (2015). The role of death-associated protein 3 in apoptosis, anoikis and human cancer. *Cancer Cell Int.* 15:39. doi: 10.1186/s12935-015-0187-z

Wazir, U., Jiang, W. G., Sharma, A. K., and Mokbel, K. (2012). The mRNA expression of DAP3 in human breast cancer: correlation with clinicopathological parameters. *Anticancer Res.* 32, 671–674.

Xiao, H., Wang, B., Xiong, H. X., Guan, J. F., Wang, J., Tan, T., et al. (2021). A novel prognostic index of hepatocellular carcinoma based on immunogenomic landscape analysis. *J Cell Physiol* 236, 2572–2591. doi: 10.1002/jcp.30015

Xu, T., Ruan, H., Gao, S., Liu, J., Liu, Y., Song, Z., et al. (2020). ISG20 serves as a potential biomarker and drives tumor progression in clear cell renal cell carcinoma. *Aging (Albany NY)* 12, 1808–1827. doi: 10.18632/aging.102714

Yang, L., Li, Y., Wei, Z., and Chang, X. (2018). Coexpression network analysis identifies transcriptional modules associated with genomic alterations in neuroblastoma. *Biochim. Biophys. Acta Mol. Basis Dis.* 1864, 2341–2348. doi: 10.1016/j.bbadis.2017.12.020

Yin, N. C., Lang, X. P., Wang, X. D., and Liu, W. (2015). AGER genetic polymorphisms increase risks of breast and lung cancers. *Genet. Mol. Res.* 14, 17776–17787. doi: 10.4238/2015.December.22.2

Yoo, S. K., Freisinger, C. M., LeBert, D. C., and Huttenlocher, A. (2012). Early redox, Src family kinase, and calcium signaling integrate wound responses and tissue regeneration in zebrafish. *J. Cell. Biol.* 199, 225–234. doi: 10.1083/jcb.201203154

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, X., Zeng, T., Wang, X., Li, G., and Chen, L. (2015). Unravelling personalized dysfunctional gene network of complex diseases based on differential network model. *J. Transl. Med.* 13:189. doi: 10.1186/s12967-015-0546-5

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:Article17. doi: 10.2202/1544-6115.1128

Zhang, C. C., Kaba, M., Ge, G., Xie, K., Tong, W., Hug, C., et al. (2006). Angiopoietin-like proteins stimulate ex vivo expansion of hematopoietic stem cells. *Nat. Med.* 12, 240–245. doi: 10.1038/nm1342

Zhang, H., Wang, S., and Huang, T. (2020). Identification of chronic hypersensitivity pneumonitis biomarkers with machine learning and differential co-expression analysis. *Curr. Gene Ther.* doi: 10.2174/1566523220666201208093325 [Epub ahead of print].

Zhang, W., Fan, J., Chen, Q., Lei, C., Qiao, B., and Liu, Q. (2018). SPP1 and AGER as potential prognostic biomarkers for lung adenocarcinoma. *Oncol. Lett.* 15, 7028–7036. doi: 10.3892/ol.2018.8235

Zhu, Y., Zhou, C., and He, Q. (2019). High miR-139-3p expression predicts a better prognosis for hepatocellular carcinoma: a pooled analysis. *J. Int. Med. Res.* 47, 383–390. doi: 10.1177/0300060518802727

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Zhou, Xu, Zhou, Liu, Zheng, Liu, Deng, Liu, Ren, Xia, Kong, Huang and Jiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 04 June 2021
doi: 10.3389/fcell.2021.675978

Pan-Cancer Metastasis Prediction Based on Graph Deep Learning Method

*Yining Xu, Xinran Cui and Yadong Wang**

Department of Computer Science, Harbin Institute of Technology, Harbin, China

Tumor metastasis is the major cause of mortality from cancer. From this perspective, detecting cancer gene expression and transcriptome changes is important for exploring tumor metastasis molecular mechanisms and cellular events. Precisely estimating a patient's cancer state and prognosis is the key challenge to develop a patient's therapeutic schedule. In the recent years, a variety of machine learning techniques widely contributed to analyzing real-world gene expression data and predicting tumor outcomes. In this area, data mining and machine learning techniques have widely contributed to gene expression data analysis by supplying computational models to support decision-making on real-world data. Nevertheless, limitation of real-world data extremely restricted model predictive performance, and the complexity of data makes it difficult to extract vital features. Besides these, the efficacy of standard machine learning pipelines is far from being satisfactory despite the fact that diverse feature selection strategy had been applied. To address these problems, we developed directed relation-graph convolutional network to provide an advanced feature extraction strategy. We first constructed gene regulation network and extracted gene expression features based on relational graph convolutional network method. The high-dimensional features of each sample were regarded as an image pixel, and convolutional neural network was implemented to predict the risk of metastasis for each patient. Ten cross-validations on 1,779 cases from The Cancer Genome Atlas show that our model's performance (area under the curve, AUC = 0.837; area under precision recall curve, AUPRC = 0.717) outstands that of an existing network-based method (AUC = 0.707, AUPRC = 0.555).

Keywords: pan-cancer analysis, cancer metastasis, machine learning method, GCN, CNN

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Haixiu Yang,
Harbin Medical University, China
Hui Ding,
University of Electronic Science and Technology of China, China

***Correspondence:**
Yadong Wang
ydwang@hit.edu.cn

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *04 March 2021*
Accepted: *12 April 2021*
Published: *04 June 2021*

Citation:
Xu Y, Cui X and Wang Y (2021) Pan-Cancer Metastasis Prediction Based on Graph Deep Learning Method. Front. Cell Dev. Biol. 9:675978. doi: 10.3389/fcell.2021.675978

INTRODUCTION

For years, tumor metastasis remains the leading cause of death from malignancies. Researchers have explored the causes of cancer metastasis based on a variety of biological processes (Spano et al., 2012; Fares et al., 2020; Wang et al., 2019). Gene expression data shows tumor status directly and is easily achieved. With its high sensitivity, wide detection range, and low cost, RNA-seq gene expression data is suitable for the analysis of biological samples from very limited sources like metastasizing cancer cells. For these reasons, RNA-seq gene expression data is widely used in cancer prognosis analysis and treatment plan formulation. Besides this, large-scale cancer databases like The Cancer Genome Atlas (TCGA), TCGA Pan-cancer Clinical Data Resource (TCGA-CDR)

(Liu et al., 2018), and Cancer Cell Line E (Ghandi et al., 2019) are becoming more sophisticated, providing long-term-tracked reliable clinical data as well as the corresponding gene expression data and making it possible to conduct an in-depth study on the relationship between gene expression and clinical phenotype. Such genomic feature with cancer phenotype can effectively improve cancer prognosis outcome over the current clinical measures for risk assessment of patients (Byron et al., 2016; Grimes et al., 2018; Zhao et al., 2020a). The majority of cancer metastasis prediction methods based on RNA-seq gene expression data are feature processing followed by machine learning models.

Cancer Metastasis Prediction Methods Based on Fold-Change Feature Selection

The commonly used fold-change feature selection methods include: *filters*, in which data is independent of the sequential machine learning model, and the method's evaluation is to judge the relationship between the one-dimensional features and the target variables (i.e., Pearson correlation coefficient, Gini coefficient, information gain, variance check, and similarity measurement); *wrapper*s, in which feature selection is wrapped with the classifier, and the result of the classifier is used as the evaluation method of whether the feature is filtered or not (i.e., recursive feature elimination, stability selection); *embedded*, in which feature selection is carried out by the characteristics of the classifier (i.e., L1-regularization, L2-regularization, mean decrease impurity or mean decrease accuracy). Landemaine et al. (2008) selected six target genes according to gene expression fold-change and predicted breast cancer lung metastasis. Chibon et al. (2010) selected gene expression features according to *z*-scores and *P*-values and predicted sarcoma prognosis. GV Glinsky (2006) filtered gene expression features according to fold-change and predicted prostate cancer outcome. These methods ignored the biological significance in the data, leading to limited explanatory ability.

Cancer Metastasis Prediction Methods Based on Priori Knowledge Feature Selection

Feature selection depends more on prior knowledge with the development of biological information database, including pathway database (i.e., KEGG, BioCyc, and Reactome) and gene enrichment database (i.e., MSigDB). Priori-knowledge-based cancer metastasis prediction methods mainly consist of two key steps: feature filtering based on priori-knowledge database or fold-change feature selection or both, then machine learning modeling (Kamps et al., 2017; Chaurasia et al., 2018; Ideta et al., 2021). These methods took gene pathway or enrichment knowledge into consideration but still ignored gene–gene regulation knowledge, from which vital information could be extracted. Besides this, it had been proven that regulation network information could enhance a machine learning model's performance (Zhao et al., 2020b, 2021b). In conclusion, prior-knowledge-based feature extraction can promote a machine learning model's performance albeit limitations still exist. Information from gene–gene regulation network should be effectively developed and efficiently combined with existing methods in future studies to achieve more convincing results.

Cancer Metastasis Prediction Methods Based on Network Feature Extraction

Network information efficiently helps improve cancer outcome prediction and is always accompanied by prior knowledge feature selection. Protein–protein regulation information and gene regulation network could significantly improve a model's prediction accuracy. In the study of Roy et al. (2014), features were first filtered by transcript factors' prior knowledge, then ranked by protein–protein interaction network and gene regulation network, and finally predicted 13 cancer cases' prognoses by support vector machine model. Protein interaction network is involved in the study of HY Chuang et al. (2007) and achieved 72.2% accuracy in prediction result on breast cancer metastasis by logistic regression classifier. In the study of Z Wang et al. (2020), application of a gene–gene regulation network-based feature extraction method helped raise the model's area under the curve (AUC) from 0.623 to 0.707 on breast cancer outcome prediction. In conclusion, the network helped to increase the machine learning model's performance. Meanwhile, the feature extracting method needs to be optimized.

Machine Learning Models Used for Cancer Metastasis Prediction

Various machine learning models have been applied to cancer metastasis prediction. Tapak et al. (2015) predicted breast cancer survival and metastasis by multi-machine learning methods including naive Bayes, random forest, AdaBoost, support vector machine (SVM), least square SVM and Adabag, logistic regression, and linear discriminant analysis. They found that all these models performed well in cancer survival prediction but had a limited effect in cancer metastasis prediction. Nicolò et al. (2020) used random survival forest analysis to predict early-stage breast cancer metastatic relapse [area under receiver operating characteristic curve (AUROC) = 0.73]. Advanced machine learning is also applied to cancer metastasis prediction studies. Wang et al. (2020) classified a breast cancer patient's metastasis status by support vector machine (AUROC = 0.71) and deep neural network (AUROC = 0.61). Among these studies, SVM outstands other models. The prediction results of advanced machine learning models like deep neural network (DNN) and convolutional neural network (CNN) were not satisfying. Until recently, DNN and CNN are used more commonly in combination with relational network prior knowledge. The research of Zhao et al. (2020c, 2021a) has proven the advanced machine learning models' prediction performance.

Application of Graph Convolutional Network in Network Feature Extraction

Graph convolutional network (GCN) is commonly used to handle topological data, including recovery of missing links or classification of ungrouped nodes. It is suitable for processing

relation data including gene–gene regulation network, protein–protein interaction network, gene–disease relation network, *etc*. The research of Zhao et al. (2020c) indicated GCN's prediction power on drug–target interaction networks. It is thus speculated that GCN is with high proficiency to handle gene–gene regulation network with improvement to directed network.

Our Aims

To address the limitations of the above-mentioned approaches, we propose directed regulation graph convolutional network (DR-GCN), in which improvements for directed graph are added to GCN. In our study, gene–gene regulation network, gene–cancer correlation, and gene expression are integrated. Then, with the advanced feature we got from DR-GCN, we trained a CNN model to predict pan-cancer metastasis.

The major contributions of this research are as follows:

- We integrated multiple features by DR-GCN and reserved and extracted cancer-related information from gene expression data.
- The results of our CNN model show that DR-GCN feature extraction and CNN performed remarkably in pan-cancer outcome prediction.

MATERIALS AND METHODS

In this section, we first introduce the data preparation and workflow of the whole progress, then introduce the transfer formula of D-GCN-based feature extraction network (DR-GCN as well as its improvement on directed graph), and finally introduce the CNN model's structure designed for pan-cancer outcome prediction.

Data Preparation

In this study, we choose gene expression data (mRNA expression sequenced by second-generation technique) as our model's input feature, which were downloaded from TCGA. To label the data, we analyzed clinical data downloaded from TCGA-CDR (Liu et al., 2018), which is the official clinical supplement for TCGA.

We choose three key clinical indicators from TCGA-CDR clinical file, which are tumor_status, new_tumor_event_type, new_tumor_event_site, and PFI.time; tumor_status tells whether one patient is in a state of with tumor, tumor-free, or not clear; new_tumor_event_type tells whether one patient's new tumor event is metastasis, recurrence, new primary, or none; new_tumor_event_site tells whether one patient has a tumor discovered in a certain site; and PFI.time tells one patient's last record time point or last tumor event time point, whichever is shorter. Firstly, three types of patients were labeled "metastasis" (1): cases that have a clear tumor_status metastasis, cases that have no clear tumor_status but have clear new_tumor_event_type metastasis, and cases that have no clear tumor_status or new_tumor_event_type label but have a clear new_tumor_event telling the metastasis site. Secondly, among the residual cases, we removed the cases in which the tumor_status is "not clear" or "NA" and then preliminarily label the remaining cases "non-metastasis" (0); among these "non-metastasis" cases, we kept those with new_tumor_event "NA" as well as tumor_status "tumor free." Finally, we considered the PFI.time; some of the non-metastasis cases could be temporary due to the short clinical test interval. According to this consideration, we kept 70% of the longest non-metastasis cases and all the metastasis cases.

Considering the sample size, for a relatively balanced proportion of positive and negative cases, we choose six cancer types from TCGA for this study, which are as follows: breast invasive carcinoma, stomach adenocarcinoma, lung squamous cell carcinoma, lung adenocarcinoma, pancreatic adenocarcinoma, and skin cutaneous melanoma as pan-cancer input. These six cancer types have the highest rate of clearly described metastasis status in TCGA-CDR. Besides that, they have a moderate proportion of positive and negative cases (near 1:1). They are also among the most common human cancers. The number of positive and negative cases for each cancer is shown in **Table 1**.

Each case's gene expression feature is listed in a feature matrix, in which f_{ij} presents $gene_j$'s expression quantity in $case_i$ (**Figure 1A**). The gene regulation network here is downloaded from humannet V2 (Ideta et al., 2021); an adjacency matrix of the gene regulatory network was obtained, A total of 17,926 genes were included in this network, in which $a_{ij} = 1$ presents $gene_i$ regulates $gene_j$ (**Figure 1A**). The initial feature of each gene, which is gene's correlation with certain cancer types, is downloaded from DisGenet (Glinsky, 2006), and then we made a gene–cancer correlation table in which $w_{ik} = 1$ presents $gene_i$ and $cancer_k$ is related (**Figure 1A**. Then, we kept the intersection genes and filtered the data (**Figure 1A**).

Workflow

After the data preparation process, we extracted advanced feature from pan-cancer gene expression data by relational graph convolutional network (R-GCN)-based model DR-GCN with improved model's ability for directed graph. In this step, data's dimension is increased, the whole dataset is shaped (m,1,8526,6), in which m presents m cases included in our study, 1 presents one-layer graph structure, and 8526,6 presents the two-dimension feature extracted by DR-GCN (**Figure 1B**). Finally, based on the feature and label prepared, we designed and trained a CNN model and made prediction of pan-cancer outcomes. The whole workflow is shown in **Figure 1**.

TABLE 1 | Number of positive and negative cases for each cancer type.

	Non-metastasis cases (0)	Metastasis cases (1)
Breast invasive carcinoma	589	102
Lung adenocarcinoma	152	122
Lung squamous cell carcinoma	157	76
Stomach adenocarcinoma	151	99
Skin cutaneous melanoma	75	153
Pancreatic adenocarcinoma	21	82
Total	1,145	634

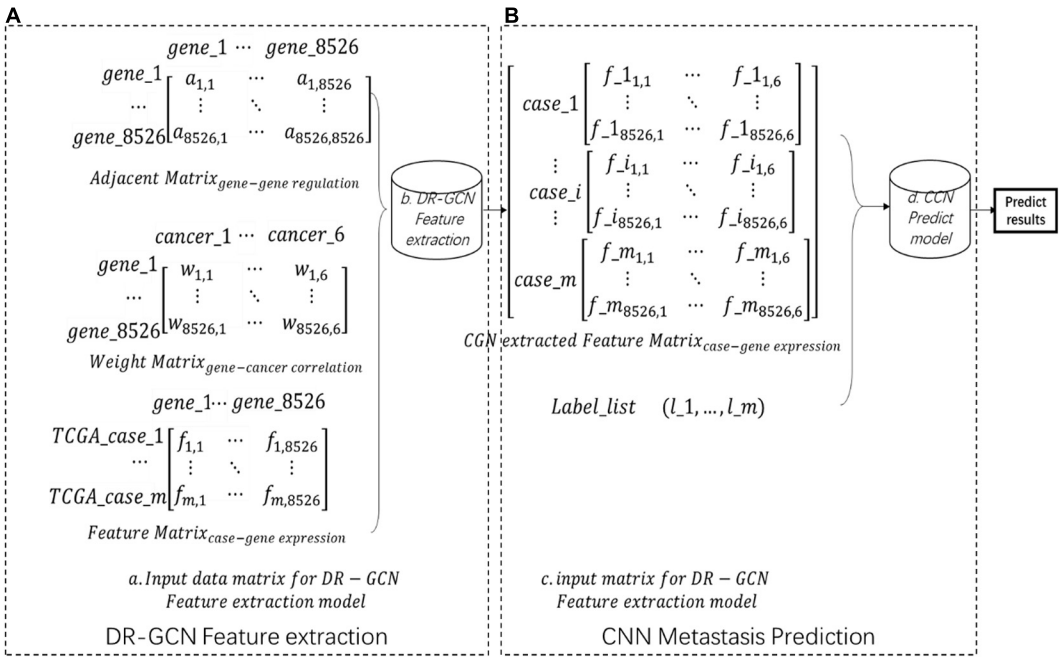


FIGURE 1 | Workflow **(A)** showing the input data matrix for directed relation-graph convolutional network (DR-GCN). Then, data were put into a DR-GCN feature extraction model. After feature extraction, we get a graph–structure data matrix as shown in panel **(B)**. Then, this data matrix combined label were put into a convolutional neural network prediction model. Finally, we got the prediction result.

R-GCN-Based Gene Expression Feature Extraction

R-GCN was introduced and applied to network link prediction (recovery of missing facts) and entity classification (recovery of missing entity attributes) (Kipf and Welling, 2016; Schlichtkrull et al., 2018). In this study, R-GCN is applied to handle gene–gene regulation network. More than ever, gene–gene regulation network is a directed network, and R-GCN was generally applied to undirected graph. To address this issue, we apply DR-GCN, a novel directed-graph processing method that has been added in this study.

DR-GCN Model Architecture

For gene–gene regulation network G = (V, E), we use binary adjacency matrix A, as described in data preparation. The adjacency matrix shows the associative property of the graph; to add the nodes' self-connection feature, we add identity matrix to the adjacency matrix:

$$\tilde{A} = A + I \tag{1}$$

A graph convolution layer can be written as such a nonlinear function:

$$H^{l+1} = f(H^l, \tilde{A}) \tag{2}$$

in which $H_0 = A$. Graph convolution problems can be abstracted as solution of $f\left(H^l, A\right) = \sigma\left(AHW^{(l)}\right)$, in which $W^{(l)}$ is the graph feature, H is weight, and σ is an activation function, $H_0 = \tilde{A}$.

Here we consider our model with the following layer-wise propagation rule:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1}\tilde{A}H^{(l)}W^{(l)}) \tag{3}$$

We consider that all genes have equal weight; then, we have:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1}\tilde{A}H^{(l)}) \tag{4}$$

$H^{(l)}$ denotes the feature vector of each gene as described in "Data Preparation." $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$. We have $\tilde{D} = \sum_j \tilde{A}_{ij}$. $L^{RW} = \tilde{D}^{-1}\tilde{A} = I_n - D^{-1}A$ is Laplacian.

After the DR-GCN feature extraction, a new feature matrix contains all gene–cancer information as well as gene–gene regulation network information. In the following section, we show that the form of this propagation rule can be applied on directed graphs.

Directed-Graph Processing Method

For the directed graph, we use each nodes' out-degree to initialize the degree matrix D. In order to make this matrix full rank, we have a strategy as follows:

$$D_{ji} = \begin{cases} \frac{1}{\sum Din} & if\ D_{ij} \neq 0 \text{ and } Din_{ij} \neq 0 \\ -\frac{1}{\sum D} & if\ D_{ij} \neq 0 \text{ and } Din_{ij} = 0 \\ 0 & \text{otherwise} \end{cases}$$

$\frac{1}{\sum Din}$ is the normalization operator for directed graph. We use in-degree in this study; $-\frac{1}{\sum D}$ is minimal value added

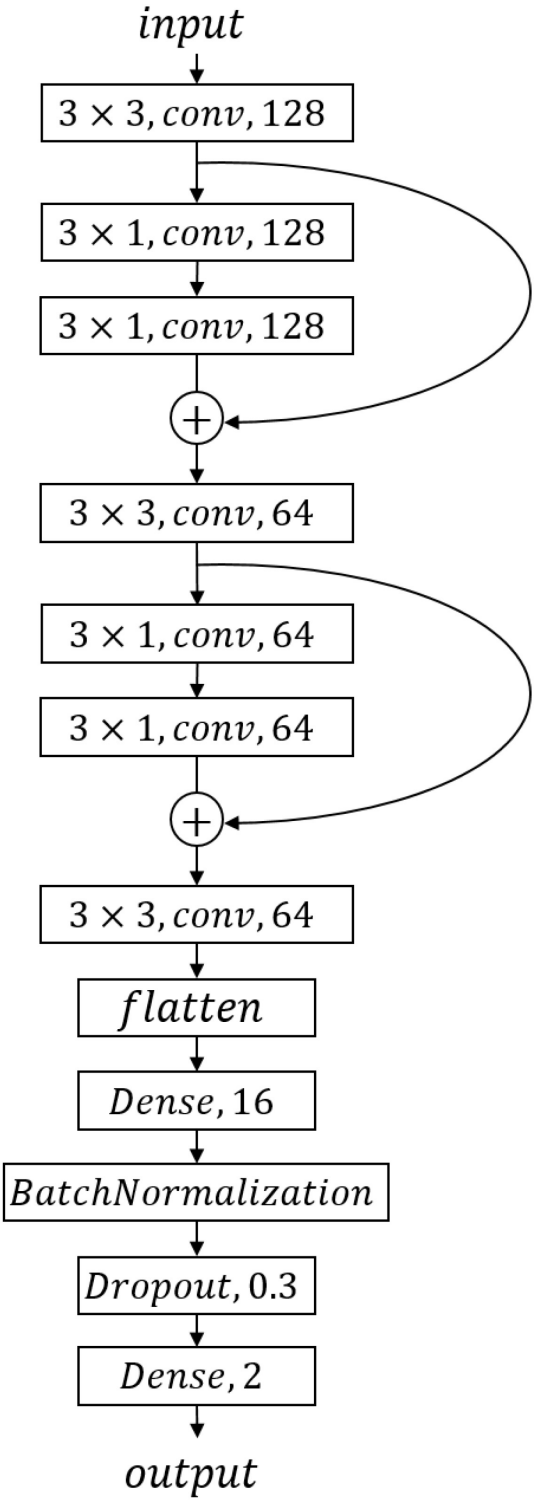


FIGURE 2 | Convolutional neural network structure and parameters.

to the degree matrix. In this way, we keep the directivity of the graph and also guarantee the invertibility of the Laplace matrix.

CNN-Based Classification Model

After the feature extraction by DR-GCN, a CNN model is designed and used as a supervised machine learning model to classify TCGA pan-cancer as benign or malignant.

We designed a CNN model, the structure of which is shown in **Figure 2**. We got one input layer, seven convolution layers followed by activation layers, one flattened layer, one batch-normalized layer, and two dense layers with one drop layer in CNN. LeakyReLU is chosen as each convolution layer's activation function to avoid neuronal death; the parameter of LeakyReLU is 0.1. We converted the two-category result to one-hot format and used softmax activation function for the last dense layer to achieve a two-category prediction result. Each layer's parameters are shown in **Figure 2**.

We choose categorical_crossentropy as loss function and optimizer "adam." This pair worked better in our study than the more commonly used binary_cross_entropy and "RMSProp" optimizer in our two-category data.

EXPERIMENTS

In this section, we first generated a real-world gene–gene regulation network, set the initial weight feature of each gene as the gene–cancer correlation, then put them into DR-GCN, and run the feature extraction module. After DR-GCN, we got a gene–gene regulation that masked the gene–cancer correlation weight table, and then we weighed the pan-cancer gene expression data from TCGA by the gene–cancer correlation weight table. Finally, we designed and trained a CNN model to identify the malignancy of the tumor.

Experiments on Feature Extraction

The gene–gene regulation network used here was downloaded from humannetV2 (Hwang et al., 2019), and the initial weight feature of each gene is downloaded from DisGenet (Piñero et al., 2020). With a raw network gene list, we converted all gene names to gene symbol ID, and then an intersection was made between raw TCGA FPKM gene list and raw network gene list. Finally, 8,526 overlapping unique genes derived from the two gene lists were included in this study.

The patients were classified as non-metastasis and metastasis groups based on the clinical information from the TCGA-CDR. After data processing, we got 1,145 non-metastasis cases and 634 metastasis cases from six different cancer types (shown in **Table 1**). All the cases were mixed randomly, and we generated our final dataset.

We next input gene–gene interaction network (8526,8526) and gene–cancer (8526,6) initial weight into DR-GCN. After

TABLE 2 | Comparison of prediction models on pan-cancer data.

	AUROC	AUPRC
DR-GCN-CNN	0.8365	0.7164
NetML-SVM	0.6122	0.4837
NetSML	0.6396	0.6331

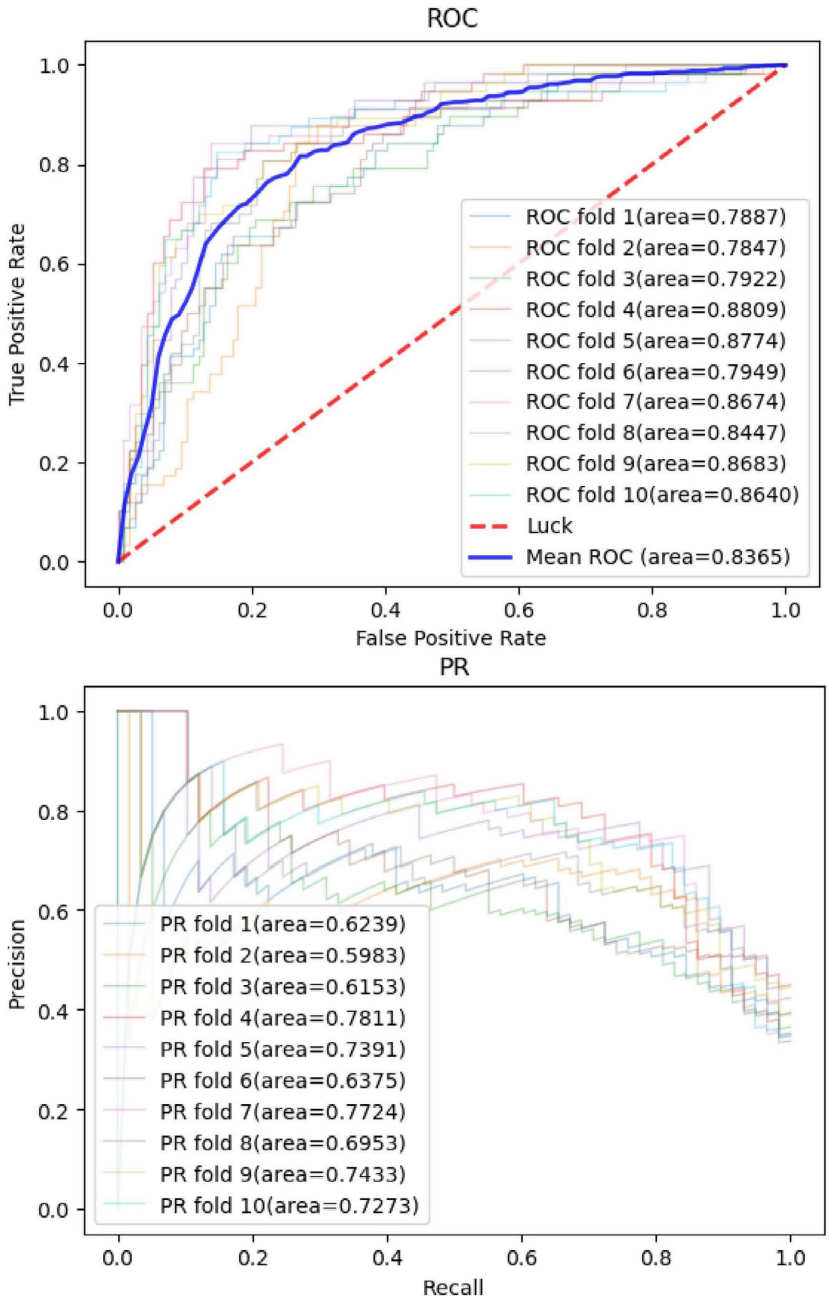


FIGURE 3 | Convolutional neural network model's receiver operating characteristic curve and precision recall curve in 10-fold cross-validations.

DR-GCN feature extraction, we got the gene–gene interaction network that masked the gene–cancer relation weight (8526,6). Then, we multiplied pan-cancer's gene expression data by this weight as CNN's input in the following experiment.

Experiments Across Pan-Cancer Datasets

We applied the CNN model to predict the cancer outcome on pan-cancer data. In the training progress, we choose categorical_crossentropy as loss function and optimizer "adam." This pair worked better in our study than the more commonly used binary_cross_entropy and "RMSProp" optimizer in our two-category data. In modeling progress, we adjusted the model's parameter by AUROC as performance measure. After parameter adjustment and model structure adjustment, we got a CNN model structure as well as each layer's parameters as shown in **Figure 2**.

To evaluate the prediction power of the workflow and the models described above, we performed a 10-fold cross-validation (CV) on pan-cancer data with ninefold data for

training and onefold data for test. The classification performance was evaluated on the test data in each dataset separately.

Then, we applied the same pan-cancer data and experimental process on the network-based model of Wang et al. (2020) to demonstrate our model's predictive ability compared with those of other network-based models. We report 10CVs' mean AUROC and the area under precision recall curve (AUPRC) scores of each model mentioned above in **Table 2**. Our CNN model's 10-CV receiver operating characteristic curve and precision recall results are shown in **Figure 3**.

DISCUSSION AND CONCLUSION

For years, researchers made efforts to enhance the cancer metastasis predicting models' performance. Many methods have been raised for different cancers' metastasis prediction, but the prediction results of the models were not satisfying, especially in pan-cancer metastasis studies; few of them took gene–gene regulation network into consideration. We believe that gene–gene regulation network information provides vital information in a pan-cancer metastasis study. According to this consideration, we applied DR-GCN to extract feature from gene–gene regulation net and achieved high AUC and AUPR scores in a CNN predicting model.

In most studies, researchers used prior knowledge like gene co-expression information and pathway analysis in feature selection to enhance the cancer metastasis predicting models' performance. In contrast, our work applied an effective R-GCN method, DR-GCN, specially for gene–gene regulation network information extraction. We efficiently added information to regular gene expression data and thereby achieved a better CNN prediction result, but limitations still exist in DR-GCN feature extraction and similar researches. DR-GCN essentially highlighted nodes that could be vital according to information network; its reliability highly depends on the network's reliability. Due to DR-GCN's character mentioned above, as well as the gene–gene regulation network which remained to be improved, we believe that DR-GCN's performance is far from optimal. In subsequent studies, R-GCN can be applied to gene–gene regulation focusing on unknown regulation relation detection.

In cancer metastasis predictions, the most commonly used predicting model is SVM, and it always worked better than advanced models like DNN and CNN because SVM is more suitable for linear gene expression data. DR-GCN feature extraction provided a higher feature dimension and therefore enhanced CNN model's prediction performance.

In conclusion, DR-GCN feature extraction distinctly improved the CNN model's prediction ability compared with other cancer metastasis prediction methods. Ten-fold cross-validation confirmed the high AUC and AUPR of CNN. The code and results of DR-GCN and CNN model are uploaded on Github[1], which will allow researchers to apply it to other pan-cancer datasets.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

YX designed workflow, improved feature extraction algorithm and structured predict model, then processed main part of experiment, and composed main part of this thesis. XC participated predict model's parameter debugging and thesis composing. YW advised the whole thesis work. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The results published or shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

[1] https://github.com/YININGXU512/GCN-CNN

REFERENCES

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., and Craig, D. W. (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* 17, 257–271. doi: 10.1038/nrg.2016.10

Chaurasia, V., Pal, S., and Tiwari, B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *J. Algorithms & Comput. Technol.* 12, 119–126. doi: 10.1177/1748301818756225

Chibon, F., Lagarde, P., Salas, S., Pérot, G., Brouste, V., Tirode, F., et al. (2010). Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat. Med.* 16, 781–787. doi: 10.1038/nm.2174

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network–based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3:140. doi: 10.1038/msb4100180

Fares, J., Fares, M. Y., Khachfe, H., Salhab, H. A., and Fares, Y. (2020). Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct. Target. Ther.* 5, 1–17.

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R. III, et al. (2019). Next-generation characterization of the cancer cell line encyclopedia. *Nature* 569, 503–508.

Glinsky, G. V. (2006). Genomic models of metastatic cancer: functional analysis of death-from-cancer signature genes reveals aneuploid, anoikis-resistant, metastasis-enabling phenotype with altered cell cycle control and activated pcg protein chromatin silencing pathway. *Cell Cycle* 5, 1208–1216. doi: 10.4161/cc.5.11.2796

Grimes, T., Walker, A. R., Datta, S., and Datta, S. (2018). Predicting survival times for neuroblastoma patients using RNA-seq expression profiles. *Biol. Direct* 13, 1–15.

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2019). HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* 47, D573–D580.

Ideta, Y., Tagawa, T., Hayashi, Y., Baba, J., Takahashi, K., Mitsudo, K., et al. (2021). Transcriptomic profiling predicts multiple pathways and molecules associated with the metastatic phenotype of oral cancer cells. *Cancer Genomics Proteomics* 18, 17–27. doi: 10.21873/cgp.20238

Kamps, R., Brandão, R. D., Bosch, B. J., Paulussen, A. D., Xanthoulea, S., Blok, M. J., et al. (2017). Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. *Int. J. Mol. Sci.* 18:308. doi: 10.3390/ijms18020308

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv* [preprint] arXiv:1609.02907,

Landemaine, T., Jackson, A., Bellahcene, A., Rucci, N., Sin, S., Abad, B. M., et al. (2008). A six-gene signature predicting breast cancer lung metastasis. *Cancer Res.* 68, 6092–6099. doi: 10.1158/0008-5472.can-08-0436

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400.–416.

Nicolò, C., Périer, C., Prague, M., Bellera, C., MacGrogan, G., Saut, O., et al. (2020). Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. *JCO Clin. Cancer Inf.* 4, 259–274. doi: 10.1200/cci.19.00133

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855.

Roy, J., Winter, C., Isik, Z., and Schroeder, M. (2014). Network information improves cancer outcome prediction. *Brief. Bioinform.* 15, 612–625. doi: 10.1093/bib/bbs083

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. (2018). "Modeling relational data with graph convolutional networks," in *Proceedings of the European Semantic Web Conference*, (Cham: Springer), 593–607. doi: 10.1007/978-3-319-93417-4_38

Spano, D., Heck, C., De Antonellis, P., Christofori, G., and Zollo, M. (2012). Molecular networks that regulate cancer metastasis. *Seminars Cancer Biol.* 22, 234–249. doi: 10.1016/j.semcancer.2012.03.006

Tapak, L., Saidijam, M., Sadeghifar, M., Poorolajal, J., and Mahjub, H. (2015). Competing risks data analysis with high-dimensional covariates: an application in bladder cancer. *Genomics Proteomics Bioinformatics* 13, 169–176. doi: 10.1016/j.gpb.2015.04.001

Wang, J., Dean, D. C., Hornicek, F. J., Shi, H., and Duan, Z. (2019). RNA sequencing (RNA-Seq) and its application in ovarian cancer. *Gynecol. Oncol.* 152, 194–201. doi: 10.1016/j.ygyno.2018.10.002

Wang, Z., He, Z., Shah, M., Zhang, T., Fan, D., and Zhang, W. (2020). Network-based multi-task learning models for biomarker selection and cancer outcome prediction. *Bioinformatics* 36, 1814–1822. doi: 10.1093/bioinformatics/btz809

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* bbaa212. [Epub ahead of print].

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Valsdottir, L. R., Zang, T., and Peng, J. (2020c). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22, 2141–2150. doi: 10.1093/bib/bbaa044

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021a). Prediction and collection of protein–metabolite interactions. *Brief. Bioinform.* bbab014. [Epub ahead of print].

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Xu, Cui and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 07 June 2021
doi: 10.3389/fcell.2021.671359

Identification and Validation of a Novel Prognosis Prediction Model in Adrenocortical Carcinoma by Integrative Bioinformatics Analysis, Statistics, and Machine Learning

Xin Yan[1,2], Zi-Xin Guo[2], Dong-Hu Yu[2], Chen Chen[1,3], Xiao-Ping Liu[2], Zhi-Wei Yang[2], Tong-Zu Liu[2] and Sheng Li[1,2,3]**

[1] Department of Biological Repositories, Zhongnan Hospital, Wuhan University, Wuhan, China, [2] Department of Urology, Zhongnan Hospital, Wuhan University, Wuhan, China, [3] Human Genetics Resource Preservation Center of Hubei Province, Wuhan, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Jungang Chen,
University of Arkansas for Medical Sciences, United States
Ling-Ling Zheng,
Sun Yat-sen University, China

***Correspondence:**
Tong-Zu Liu
liutongzu@163.com
Sheng Li
lisheng-znyy@whu.edu.cn

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *23 February 2021*
Accepted: *23 April 2021*
Published: *07 June 2021*

Citation:
Yan X, Guo Z-X, Yu D-H, Chen C, Liu X-P, Yang Z-W, Liu T-Z and Li S (2021) Identification and Validation of a Novel Prognosis Prediction Model in Adrenocortical Carcinoma by Integrative Bioinformatics Analysis, Statistics, and Machine Learning. Front. Cell Dev. Biol. 9:671359. doi: 10.3389/fcell.2021.671359

Adrenocortical carcinoma (ACC) is a rare malignancy with poor prognosis. Thus, we aimed to establish a potential gene model for prognosis prediction of patients with ACC. First, weighted gene co-expression network (WGCNA) was constructed to screen two key modules (blue: P = 5e-05, R^2 = 0.65; red: P = 4e-06, R^2 = −0.71). Second, 93 survival-associated genes were identified. Third, 11 potential prognosis models were constructed, and two models were further selected. Survival analysis, receiver operating characteristic curve (ROC), Cox regression analysis, and calibrate curve were performed to identify the best model with great prognostic value. Model 2 was further identified as the best model [training set: $P < 0.0001$; the area under curve (AUC) value was higher than in any other models showed]. We further explored the prognostic values of genes in the best model by analyzing their mutations and copy number variations (CNVs) and found that MKI67 altered the most (12%). CNVs of the 14 genes could significantly affect the relative mRNA expression levels and were associated with survival of ACC patients. Three independent analyses indicated that all the 14 genes were significantly associated with the prognosis of patients with ACC. Six hub genes were further analyzed by constructing a PPI network and validated by AUC and concordance index (C-index) calculation. In summary, we constructed and validated a prognostic multi-gene model and found six prognostic biomarkers, which may be useful for predicting the prognosis of ACC patients.

Keywords: adrenocortical carcinoma, bioinformatics, machine learning, hub genes, prognosis prediction model

INTRODUCTION

Adrenal cortical carcinoma (ACC) is a relatively rare malignant tumor (Lam, 2021). Forty percent of the cases had distant metastasis at the time of diagnosis (Ayala-Ramirez et al., 2013; Guillaume et al., 2014). Worst of all, most patients with ACC have to face poor prognosis. The morbidity age of ACC showed bimodal distribution in kids and in 40- to 50-year-old people (Rodriguez-Galindo et al., 2021). As for the prognosis of ACC patients, the 5-year survival rate for ACC patients

was as reported as 50% according to recent data from the American Society of Clinical Oncology (ASCO). At present, many studies identified molecular biomarkers associated with ACCs to solve this problem (He et al., 2017). It means that predicting survival and prognosis of patients with the help of biomarkers has great application prospects. Thus, we aimed to construct a multi-gene signature and some prognostic biomarkers, which perhaps can help in predicting the prognosis of patients with ACC.

Weighted gene co-expression network (Langfelder and Horvath, 2008) has been widely used for potential biomarker selection (Lv et al., 2021). In the present study, by using GSE76019 [downloaded from Gene Expression Omnibus (GEO) database], a co-expression network was constructed to select two key modules related to survival state. Then we screened out hub genes in a co-expression network and further identified survival-associated genes among them by using GEPIA (Gene Expression Profiling Interactive Analysis) (Tang et al., 2017). Considering that Ridge, Elastic Net, and Lasso (least absolute shrinkage and selection operator) methods were widely used data mining methods especially for regression model construction (Ye et al., 2021), based on the survival-associated genes, all the three methods were included for constructing multiple-gene signatures by using the data from TCGA database. Two prognostic models were constructed, which could significantly predict the OS (overall survival) of patients with ACC. Moreover, we constructed two nomograms based on the two multiple-gene signatures, which were verified to predict the probability of the overall survival of patients with ACC well. The calibration curve indicated that nomogram of model 2 showed better prognostic value compared with model 1, which was regarded as the best prognostic model. Moreover, six genes were further screened out by protein–protein interaction (PPI) network construction, which were regarded as hub genes in this study. In conclusion, we constructed a multi-gene signature that might be an independent survival and prognostic biomarker for ACCs. The nomogram based on the best model for survival rate of ACC patient prediction might give more beneficial guidance for clinical practice, and six hub genes were also screened out, which might be novel prognostic biomarkers of patients with ACC.

MATERIALS AND METHODS

Data Collection

Supplementary Figure 1 showed the flowchart for the present research. In total, four independent datasets with complete clinical information were retrieved from public databases in the present study. For the WGCNA analysis, dataset GSE76019 (Pinto et al., 2016) was retrieved from the GEO database[1] to conduct WGCNA. GSE76019, performed on Affymetrix HT HG-U133 + PM array plate, included 34 ACCs. To search for the best regression model by using the Ridge ($\alpha = 0$), Elastic Net (with α varying from 0.1 to 0.9), and Lasso ($\alpha = 1$) methods, we downloaded the mRNA-seq data and clinical information of ACCs from The Cancer Genome Atlas (TCGA) database. After weeding out incomplete samples, 79 ACCs were included in this study. We randomly split them into two groups with a ratio of 2:1. The training set, including 52 samples, was selected to construct the regression models. Meanwhile, the rest of the 27 samples were used to preliminarily screen the best model and further validate our results (as an internal verification set). In order to conduct external validation for our results, two independent datasets were further retrieved. GSE19750 (Demeure et al., 2013) that performed on Affymetrix Human Genome U133 Plus 2.0 array included 44 ACCs and four normal adrenal glands. Dataset GSE76021, performed on Affymetrix Human Genome U133A Array, included 29 ACC samples. **Supplementary Table 1** shows a detailed information including gender, age, stage, grade, and survival data of these datasets. The clinical characteristics of the training dataset and the internal validation dataset are shown in **Supplementary Table 2**. Further correlation analysis indicated that there was no significant difference between the two datasets (P of age: 0.870, P of gender: 0.660, P of stage: 0.051, and P of status: 0.816), which demonstrated that our grouping method was appropriate and reasonable.

To explore mutations and copy number variations (CNVs) of genes in the best model, we also downloaded the CNV data of the ACC samples from TCGA database.

Preprocessing of Collected Data

For the microarray data of GSE19750, GSE76019, and GSE76021 from Affymetrix, we first downloaded the raw "CEL" data from the GEO database and further applied robust multiarray averaging (RMA) method for background adjustment and quantile normalization, by using the "affy" package in R software. For TCGA–ACC data, we first downloaded the count number expression profiles. The normalization and log2 transformation were next obtained via "DEseq.2" (Love et al., 2014) in R. Batch effects from non-biological technical biases were corrected using ComBat algorithm based on R package "sva" (Leek et al., 2012).

For WGCNA, 34 adrenocortical carcinoma samples from GSE76019 were included for WGCNA. Probe sets were filtered by their variances across all samples; only probes with variances ranked in the top 2,500 were selected for subsequent analyses. Microarray quality was assessed by sample clustering according to the distance between different samples in Pearson's correlation matrices and average linkage. The expression values from TCGA–ACC data of the genes we identified in WGCNA were extracted to construct the regression models by using the Ridge, Elastic Net, and Lasso methods.

Co-expression Network Constructing and Key Module Screening

First, gsg (goodSamplesGenes) and sample network methods were used to check the expression data profile of the top 2,500 genes from GSE76019. A standard of Z.Ku ≥ -2.5 was set to pick out qualified ACC samples for constructing a co-expression network. Based on "WGCNA" (Langfelder and Horvath, 2008) in R software, we constructed a co-expression network. We used three branch cutting methods used for

[1] http://www.ncbi.nlm.nih.gov/geo/

classification of genes and construction of modules [manual (interactive) branch cutting approach, automatic single block analysis, and two block analysis]. Some essential parameters for branch splitting are shown here: minClusterSize = 30 and deepSplit = 2. Furthermore, a cutline for merging high-associated models was set through calculating the dissimilarity of module eigengenes (MEs). In order to identify key modules that correlated to survival state (the trait that interested us the most), we first calculated the gene significance (GS) with the aim of quantifying the correlation between genes and trait. Then we further defined the module significance (MS), which was the average GS of all the genes in a module. Based on the above analyses, the most positive correlation module and the most negative correlation module were regarded as key modules.

Exploration of Potential Functions of Genes

GO enrichment (Ashburner et al., 2000) and KEGG pathway analyses (Kanehisa and Goto, 2000) were conducted *via* "clusterProfiler" (Yu et al., 2012) in R software. For the GO part, we only identified the biological process (BP). A gene set or KEGG signaling pathway with $P < 0.05$ was thought to be strong and enriched.

Survival-Associated Gene Identification

In the present study, we chose genes that reached the standards of |cor.geneModuleMembership| > 0.8 and |cor.geneTraitSignificance| > 0.2 as hub genes. In order to find out genes associated with survival, we further performed overall survival (OS) analysis and diseases-free survival (DFS) analysis for these hub genes by using GEPIA (Tang et al., 2017). Genes were considered to be survival-associated genes when they showed significant *P*-value in both OS and DFS.

Regression Model Construction and the Best Model Screening

Based on the training set, the expression values of survival-associated genes were extracted to construct models by using the Ridge, Elastic Net, and Lasso methods. Elastic Net regression utilizes a combination of Lasso (L1-norm) and Ridge (L2-norm) regularization penalties, and reduces overfitting by limiting coefficient magnitudes (Zou and Hastie, 2005). The regression equation can be seen in a previous study (Zou and Hastie, 2005). All of the three methods were performed by using R package "glmnet" (Friedman et al., 2010). A 10-fold cross-validation was performed for tuning parameter selection in the models. The 10-fold cross-validation was based on 1-SE (standard error) criteria. Samples in the training set were randomly divided into 10 parts, nine of which were trained in turn, while the remaining one was tested, and the average of the results of the 10 tests was used to estimate the accuracy of the algorithm. A total of 11 models would be constructed. We further got a gene signature from each model that contained the most useful prognostic biomarkers that the model was thought to have. To identify the best model among the 11 models, we calculated the risk score (RS) of every sample in the internal validation set based on these models. In each model, we calculated the RS by using the formula as follows:

$$\text{Risk score} = \sum_{i=1}^{n} Coef_i \times Exp_i$$

In which Coef represents the relative regression coefficient and Exp represents the expression value of each prognostic gene identified by the Ridge, Elastic Net, or Lasso methods. In the internal validation set, we used the risk score of each model to represent the model and further worked out the AUC to distinguish live ACCs from dead ACCs by performing receiver operating characteristic curve (ROC) analysis using the R package "plotROC" (Sachs, 2015). The model with the highest AUC was regarded as the best model, which was included for further analysis.

Survival Analysis and Time-Dependent Receiver Operating Characteristic Analysis

We further worked out the risk score of each sample in all datasets based on the best model. ACCs in the training set, internal validation set, entire set, GSE19750, and GSE76021 were divided into high- and low-risk groups according to the median risk score we calculated. R package "survival" (Therneau, 2015) was used to perform survival analysis. In order to investigate the performance of the signature, R package "timeROC" (Blanche et al., 2013) was used to perform time-dependent ROC analysis. In this study, AUC values of 1, 3, and 5 years were explored, accurately, which could affect predictive accuracy.

Cox Proportional Hazards Regression Analysis

To verify the prognostic value of the gene signature (best model), we included the risk score of the best model and some essential clinicopathological features (gender, age, tumor-, T-, N-, and M-stages, and Weiss score) for univariable Cox analysis of OS via TCGA-ACC dataset. Features with *P*-value < 0.05 were immediately chosen for multivariate Cox analysis. It would be clear if the gene signature was irrelevant to other clinical features for predicting OS of ACCs after Cox regression analysis. Visualization was applied by using "forestplot" (Aut and Aut, 2016) in R software.

Gene Set Enrichment Analysis

In order to understand the lurking functions of the best multi-gene signature, we evaluated the median risk score based on GSE76019 (the dataset which was used for WGCNA analysis). After that, a total of 34 ACCs were split into two groups accurately (high- and low-risk groups). "c2.cp.kegg.v7.2.symbols.gmt," "c5.all.v7.4.symbols.gmt," and "c6.all.v7.4.symbols.gmt" were set as the reference gene sets, respectively. GSEA (Subramanian et al., 2005) was conducted between the two groups. In this study, KEGG signaling pathways reached the standards (nominal $P < 0.05$, $|ES| > 0.6$, gene size ≥ 100 and FDR < 25%) and were significantly enriched.

Nomogram Construction and Validation

Cross-validation could deal with the overfitting of the model. Thus, we performed cross-validation before nomogram construction. Based on TCGA-ACC data, "rms" in R was used to construct the nomogram (Yizhou et al., 2013). Calibration curve was plotted to test the nomogram. The calibration curve was also used for visualization and the 45° line in the curve represents the best prediction. Besides, in order to evaluate the prediction effectiveness of the nomograms, the consistency index (C-index) (Michael and Ralph, 2010) between actual probability and predicted probability was measured. ROC curves were also plotted using R package "pROC" (Robin et al., 2011). In addition, in order to check the stability of the nomogram with and without ARG signature, we performed time-dependent (1-, 3-, and 5-year) ROC analysis. Beyond ROC, we also used R package "rmda" to perform decision curve analysis (DCA) (Vickers and Elkin, 2006) to make sure if the signature was of great value for predicting 1-, 3-, and 5-year survival probability by using TCGA–ACC data and GSE19750. Moreover, survival analysis was performed to explore the difference in survival between different nomogram point groups.

Mutations and Copy Number Variations of Genes in the Best Model

To explore mutations and CNVs of genes in the best model, we included all the ACC samples with CNV data from TCGA database for this analysis. A webtool named CBio Cancer Genomics Portal[2] was chosen for exploration of genetic alterations. Furthermore, we identified the association between genetic alterations and the clinical features of patients with ACC. We used chi-square or Mann–Whitney *U*-test to analyze the statistical significance of the result. By using the relative mRNA expression levels of genes in the best model, we also explored the relationship between CNVs and mRNA expression levels of these genes. ANOVA or Kruslal–Wallis test was chosen to test the result. Unpaired t-test was used to test the expression differences between shallow deletion and gain groups. Moreover, the association between CNVs of genes in the best model and survival of patients with ACC was also identified by performing survival analysis based on R package "survival."

Linear Discriminant Analysis, K-Nearest Neighbor, and Support Vector Machine

To verify the prognostic value of genes in the best model, we regarded genes as variables and relative mRNA expression values as variable values. Then we performed LDA, KNN, and SVM analyses. We used package "MASS" (Venables and Ripley, 2002) in R software for LDA. For KNN algorithm, we chose the best K parameter by adopting cross validation method based on R package "caret" (Kuhn, 2015). Then we performed KNN algorithm by using R packages "class" (Venables and Ripley, 2002) and "kknn" (Hechenbichler and Schliep, 2004). For SVM analysis, four kinds of algorithms were used in the present study based on "e1071" (Meyer, 2015) in R software [linear-SVM, polynomial-SVM, radial basis function (RBF) SVM, and sigmoid-kernel SVM]. R package "kernlab" (Karatzoglou et al., 2004) was used to assist with SVM feature selection. We used the data from TCGA as training set to first construct classifier in each analysis and further verify the value of the classifier by calculating the classification rate based on TCGA–ACC data, GSE19750, GSE76019, and GSE76021. In this study, we regarded genes in the best model as meaningful prognostic biomarkers when the accuracy of classification ≥ 0.50.

Protein–Protein Interaction Network Construction to Screen Hub Genes in the Best Model

To find out some relative genes in the best model, we first constructed a PPI network of genes formed from this model by using the STRING (Szklarczyk et al., 2019). Then the PPI network was imported into the Cytoscape software[3]. We further calculated the degree of connectivity of each gene. In this study, genes with the highest degree were regarded as hub genes. We also analyzed and visualized the correlation among genes in the best model by performing Spearman analysis using R package "corrplot" (Wei and Simko, 2013) and "PerformanceAnalytics" (Peterson et al., 2014). We further performed ROC and calculated AUC and C-index to validate these hub genes.

RESULTS

Constructing a Co-expression Network to Identify Key Modules

We first included 33 ACC samples to perform WGCNA analysis after weeding out the outlier samples (**Supplementary Figure 2**). According to the result, the beta (β) = 5 (scale free R2 = 0.89) was further set as the soft-thresholding value for further adjacency calculation (**Supplementary Figure 3**). Eventually, as shown in **Supplementary Figure 4**, in total, eight modules were identified. Genes not included in any other significant modules were included in the gray module, which were removed from subsequent analysis. Among the eight modules, the blue module was positively associated with survival status the most (P = 5e−05, R2 = 0.65). Meanwhile, the red module was the most negatively associated module (P = 4e−06, R2 = −0.71) (**Figure 1A**). As shown in **Figure 1B**, the MS of the two modules were also higher than the MS of any other module. The relationships between MM and GS in the blue module (cor = 0.64) and red module (cor = 0.74) were also significant as suggested by **Figures 1C,D**. Thus, we considered blue and red modules to be the key modules in this study. As shown in **Supplementary Figures 5A,B**, we also created a network heatmap and a classical MDS plot.

Genes in Key Module-Associated Biological Pathways

According to GO biological process analysis, genes in the red module were enriched in four BPs including leukocyte

[2]http://www.cbioportal.org/

[3]https://cytoscape.org/

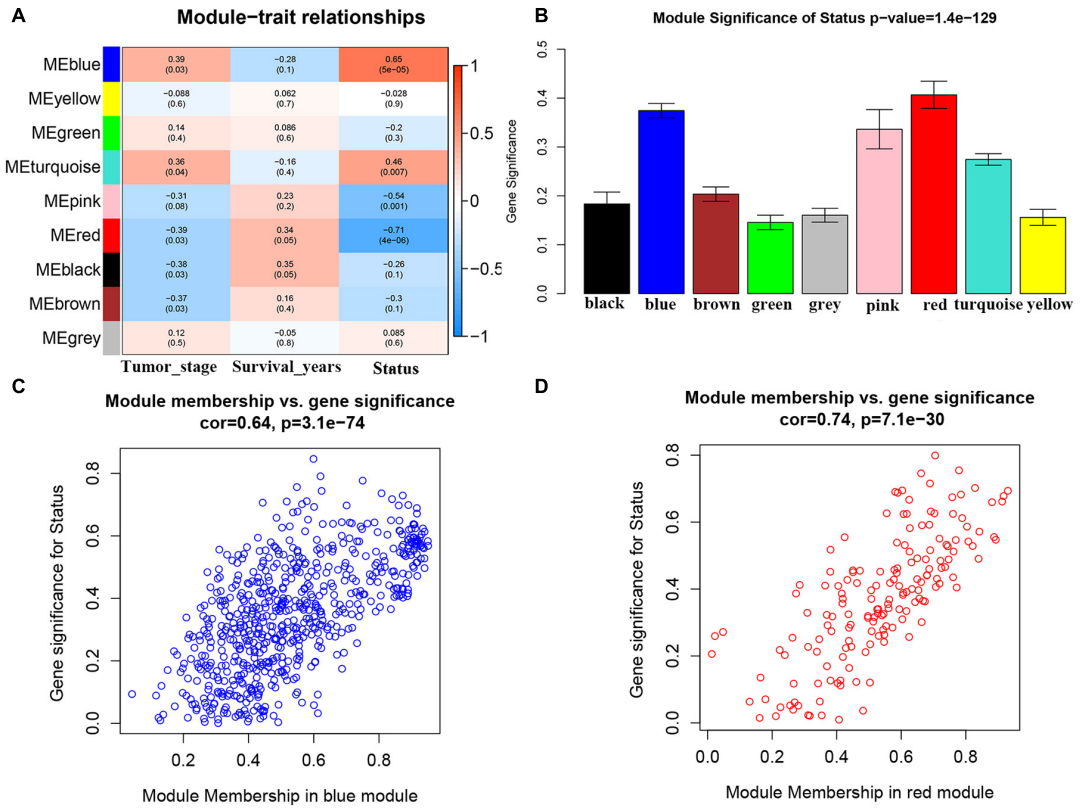


FIGURE 1 | Identification of modules associated with the clinical traits of adrenocortical carcinoma (ACC). **(A)** Heatmap of the correlation between module eigengenes and clinical traits of ACC. **(B)** Distribution of average gene significance and errors in the modules associated with survival status of ACC. **(C)** Scatter plot of module eigengenes related to survival status in the blue module. **(D)** Scatter plot of module eigengenes related to survival status in the red module.

migration, positive regulation of cell migration, regulation of G-protein-coupled receptor protein signaling pathway, and chemokine-mediated signaling pathway (**Supplementary Figure 6A**). As for genes in the blue module, in total, they were significantly correlated with 197 BPs (**Supplementary Table 3**). The top 10 enrichment BPs were organelle fission, nuclear division, chromosome segregation, nuclear chromosome segregation, mitotic nuclear division, sister chromatid segregation, mitotic sister chromatid segregation, microtubule cytoskeleton organization involved in mitosis, regulation of chromosome segregation, and mitotic spindle organization (**Supplementary Figure 6B**). When talking about KEGG pathway analysis, genes in the blue module were enriched in only three KEGG pathways including cell cycle, oocyte meiosis, and p53 signaling pathway (**Supplementary Figure 6C**).

Survival-Associated Gene Identification

With the cutoff criterion (| cor.geneModuleMembership| > 0.8 and | cor.geneTraitSignificance| > 0.2), 107 genes (12 genes in the red module and 95 genes in the blue module) were identified in the co-expression network in total. Then we performed OS and DFS analysis for these 107 genes by using GEPIA. According to the results, high expression levels of 95 genes were obviously associated with poor OS for ACCs. Meanwhile, 100 genes showed significant *P*-values in DFS analysis. Ninety-three genes were common among them, which were regarded as survival-associated genes (**Supplementary Table 4**).

Establishment of Two Multi-Gene Signatures for Predicting Overall Survival

We identified 93 biomarkers significantly related to survival of ACCs by using WGCNA analysis. With the aim of setting up a multiple-gene signature for prognosis in ACC patient prediction, we calculated relative regression coefficient of each gene using the Ridge, Elastic Net, and Lasso methods. The most powerful prognostic markers of the 93 biomarkers were screened out in each model as shown in **Supplementary Figure 7**. After that, 11 prognostic models were constructed, and the relative regression coefficients of the most powerful prognostic markers are shown in **Supplementary Table 5**. Combining the relative expression levels of the mRNA in the classifier and corresponding Lasso coefficients, we first worked out the risk scores of samples in the internal verification set. The distribution of risk scores of Ridge (**Supplementary Figure 8A**), Elastic Net with α varying from 0.1 to 0.9 (**Supplementary Figures 8E–J**), and Lasso model (**Supplementary Figure 8K**) of ACC patients based on TCGA–ACC data are shown. In all the models, it was obvious that the number of patients who died in the high-risk group was

high, when compared with the low-risk group (**Supplementary Figure 8**). Furthermore, we worked out the AUC based on the risk score (**Table 1**). Two models, namely, model 1 (AUC = 0.676; α = 0.1) and model 2 (AUC = 0.676; α = 0.6) were finally identified, which were considered as the best candidate models.

In model 1 and model 2, ACCs in the training set were further split into two groups (low- and high-risk) via a median-risk score (5.0406 in model 1 and 16.5329 in model 2). In both models, ACC patients in the high-risk group occupied worse OS (Model 1: $P < 0.0001$, **Figure 2A**; model 2: $P < 0.0001$, **Figure 2D**). We also performed time-dependent ROC analysis. The results suggested that the AUC values of model 1 in the training set were 0.95 at 1 year, 0.98 at 3 years, and 0.96 at 5 years (**Figure 2A**). Meanwhile the AUC values of model 2 in the training set were 0.92 at 1 year, 0.91 at 3 years, and 0.95 at 5 years (**Figure 2D**).

Validation of Model 1 and Model 2

In the present study, we used three validation sets (intervalidation set, GSE19750, and GSE76021) to validate the results we got from the training set. With the same method we mentioned in the training set, we calculated the risk score for each sample in these three validation sets based on model 1 and model 2. Samples in internal validation set, GSE19750, and GSE76021 were split into two groups relying on the median-risk score in each set. In model 1, according to the survival analysis, the high-risk group had lower survival rate compared with the low-risk group in all the validation sets (internal validation set: $P = 3E{-}04$; GSE19750: $P = 0.011$; GSE76021: $P = 0.012$) as the training set suggested (**Figures 2B**, **3A,B**). The same conclusion was reached in the entire TCGA-ACC dataset (**Figure 2C**). As for the results of time-dependent ROC analysis, the prognostic accuracy of model 1 in the internal validation set was 0.73 at 1 year, 0.82 at 3 years, and 0.77 at 5 years (**Figure 2B**); the prognostic accuracy of model 1 in the entire TCGA-ACC dataset was 0.86 at 1 year, 0.90 at 3 years, and 0.96 at 5 years (**Figure 2C**); the prognostic accuracy of model 1 in GSE19750 was 0.61 at 1 year, 0.88 at 3 years, and 0.91 at 5 years (**Figure 3A**); the prognostic accuracy of model 1 in GSE76021 was 0.84 at 1 year, 0.78 at 3 years, and 0.76 at 5 years (**Figure 3B**). As for model 2, the high-risk group was associated with poorer survival compared with the low-risk group (internal validation set: $P = 0.018$; entire TCGA-ACC dataset: $P < 0.0001$; GSE19750: $P = 0.0035$; GSE76021: $P = 0.012$; **Figures 2E,F, 3C,D**). Moreover, the prognostic accuracy of model 2 in the internal validation set was 0.78 at 1 year, 0.74 at 3 years, and 0.76 at 5 years (**Figure 2E**); the prognostic accuracy of model 2 in the entire TCGA-ACC dataset was 0.92 at 1 year, 0.91 at 3 years, and 0.95 at 5 years (**Figure 2F**); the prognostic accuracy of model 2 in GSE19750 was 0.60 at 1 year, 0.89 at 3 years, and 0.89 at 5 years (**Figure 3C**); the prognostic accuracy of model 2 in GSE76021 was 0.90 at 1 year, 0.78 at 3 years, and 0.77 at 5 years (**Figure 3D**). Both the two models showed great prognostic values of patients with ACC.

TABLE 1 | Area under the curve (AUC) of risk scores calculated by the Ridge (α = 0), Elastic Net (with α varying from 0.1 to 0.9), and least absolute shrinkage and selection operator (Lasso; α = 1) methods by using internal validation set.

Models	Alive vs Dead	
	AUC	95% CI
Ridge (α = 0)	0.670	0.445-0.896
Elastic Net (α = 0.1)	0.676	0.447-0.905
Elastic Net (α = 0.2)	0.670	0.440-0.901
Elastic Net (α = 0.3)	0.636	0.404-0.869
Elastic Net (α = 0.4)	0.574	0.347-0.801
Elastic Net (α = 0.5)	0.574	0.349-0.798
Elastic Net (α = 0.6)	0.676	0.457-0.895
Elastic Net (α = 0.7)	0.563	0.333-0.792
Elastic Net (α = 0.8)	0.551	0.323-0.780
Elastic Net (α = 0.9)	0.551	0.326-0.776
Lasso (α = 1)	0.534	0.308-0.760

Prognostic Value of Model 1 and Model 2

Because we had identified the two best models before, we respectively included the risk score of model 1 (or model 2) and important factors we mentioned before for the Cox regression analysis. Risk score of model 1 (hazard ratio = 4.822, 95%CI of ratio: 2.636–8.821, $P < 0.001$), risk score of model 2 (hazard ratio = 4.465, 95% CI of ratio: 2.449–8.143, $P < 0.001$), tumor stage (hazard ratio = 6.097, 95% CI of ratio: 2.433–15.282, $P < 0.001$), and M-stage (hazard ratio = 5.806, 95% CI of ratio: 2.471–13.640, $P < 0.001$) were influence features of OS as suggested by univariate Cox analysis (**Figures 4A,C**). In model 1, the results of multivariate Cox analysis suggested that even being adjusted by other features, risk scores of model 1 (hazard ratio = 4.663, 95% CI of ratio: 2.238–9.713, $P < 0.001$) were still relevant to OS among patients with ACC as shown in **Figure 4B**. In model 2, the results of multivariate Cox analysis suggested that even being adjusted by other features, risk scores of model 2 (hazard ratio = 5.603, 95% CI of ratio: 2.410–13.024, $P < 0.001$) and tumor stage (hazard ratio = 3.803, 95%CI of ratio: 1.223–11.822, $P = 0.021$) were still relevant to OS among patients with ACC (**Figure 4D**).

Identification of Model 1 and Model 2-Associated Biological Pathways

With the cutoff criteria we set before, only one risk score-related KEGG signaling pathways were enriched in model 1 including cell cycle (nominal $P < 0.001$, $|ES| = 0.738$, gene size = 123, and FDR = 0.903%) as shown in **Supplementary Table 6**. In model 2, two risk score-related KEGG signaling pathways were significantly enriched including cell cycle (nominal $P < 0.001$, $|ES| = 0.722$, gene size = 123, and FDR = 0.884%) and oocyte meiosis (nominal $P < 0.001$, $|ES| = 0.604$, gene size = 111, and FDR = 0.913%; **Supplementary Table 6**). Interestingly, these two pathways were consistent with the results of KEGG analysis of genes in the blue module. We also increased the biological pathways related to C5 (ontology gene sets). The results indicated that model 1 was significantly associated with DNR replication-related and cell cycle-related signaling pathways (**Supplementary Table 6**). The top three enriched biological pathways were mitotic sister chromatid segregation (nominal $P < 0.001$, $|ES| = 0.829$, gene size = 158, and FDR = 0%), condensed chromosome

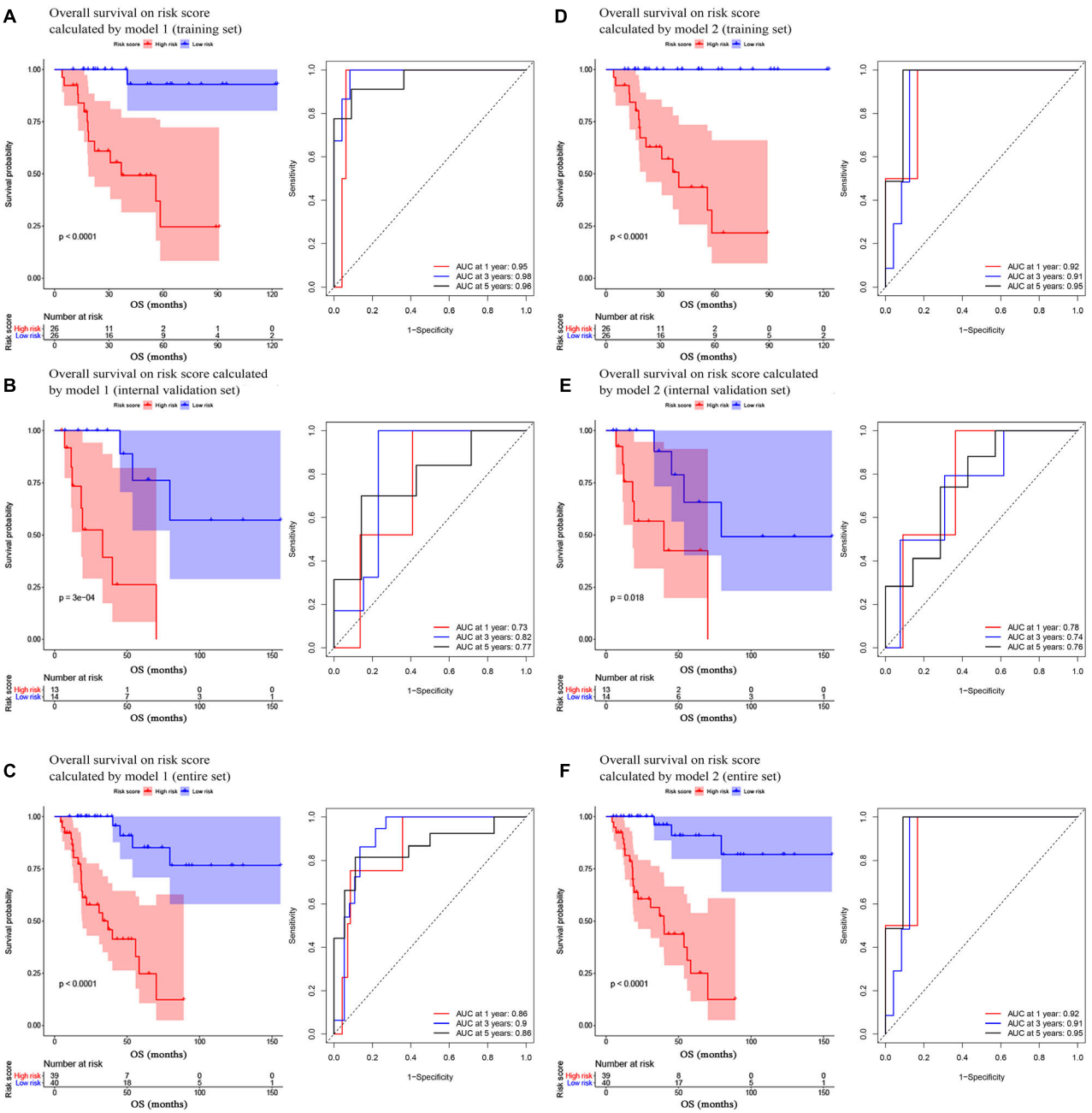


FIGURE 2 | Survival analysis of the association between risk score calculated by model 1 and overall survival time in ACC, and time-dependent receiver operating characterictic (ROC) analyses at 1, 3, and 5 years in panels **(A)** training set, **(B)** internal validation set, and **(C)** entire set. Survival analysis of the association between risk score calculated by model 2 and overall survival time in ACC, and time dependent ROC analyses at 1, 3, and 5 years in panels **(D)** training set, **(E)** internal validation set, and **(F)** entire set.

centromeric region (nominal P < 0.001, | ES| = 0.827, gene size = 108, and FDR = 0%), and kinetochore (nominal P < 0.001, | ES| = 0.809, gene size = 128, and FDR = 0%). In addition, model 2 was enriched in 43 biological pathways (**Supplementary Table 6**), and the top three were mitotic sister chromatid segregation (nominal P < 0.001, | ES| = 0.824, gene size = 158, and FDR = 0%), condensed chromosome centromeric region (nominal P < 0.001, | ES| = 0.811, gene size = 108, and FDR = 0%), and sister chromatid segregation (nominal P < 0.001, | ES| = 0.803, gene size = 187, and FDR = 0%). C6 (oncogenic signature gene sets)-related biological pathways were also explored. We found that model 1 was significantly related to seven biological pathways. Meanwhile, model 2 was associated with six pathways (**Supplementary Table 6**). Both the two models

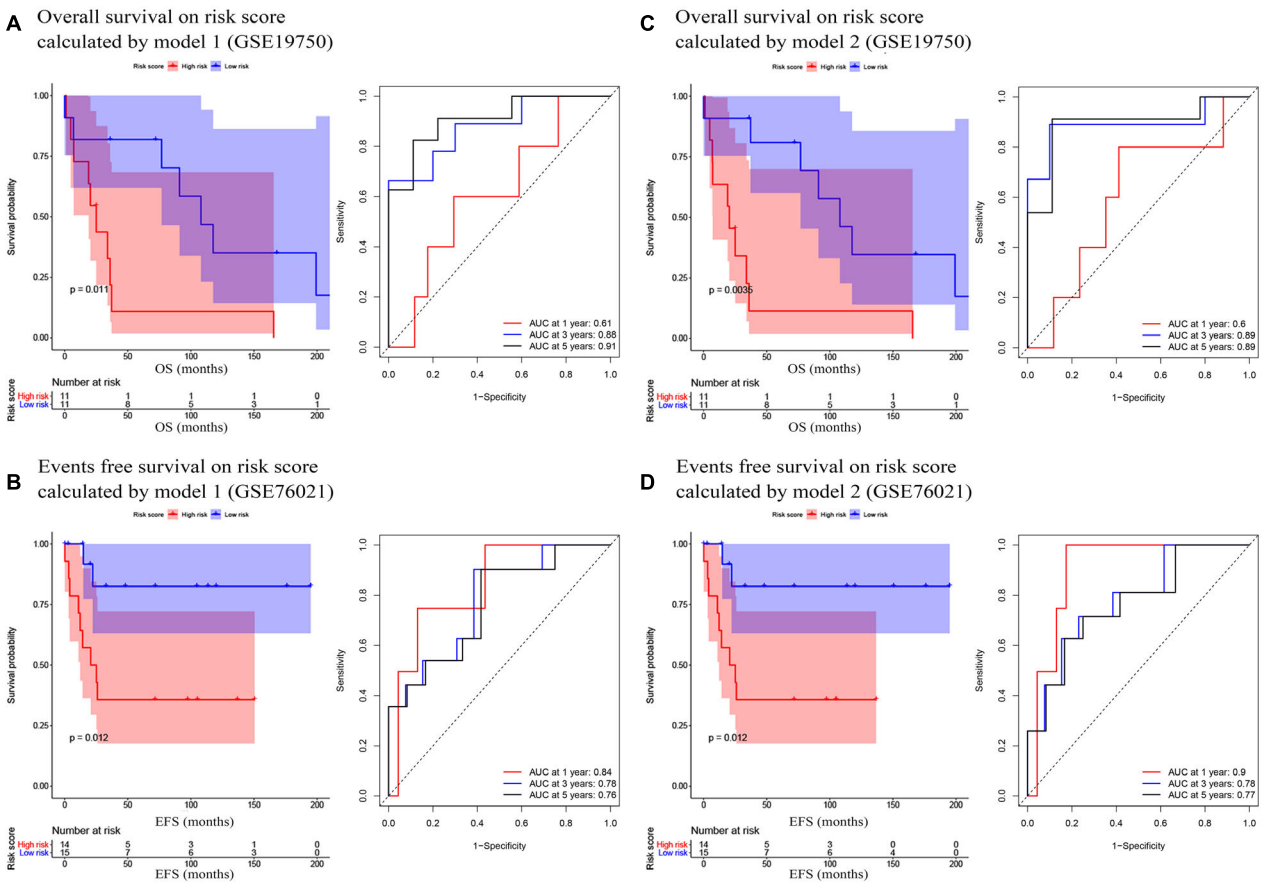


FIGURE 3 | (A) Survival analysis of the association between risk score calculated by model 1 and overall survival time in ACC, and time dependent ROC analyses at 1, 3, and 5 years in GSE19750; **(B)** Survival analysis of the association between risk score by model 1 and events-free survival time in ACC, and time dependent ROC analyses at 1, 3, and 5 years in GSE76021; **(C)** Survival analysis of the association between risk score calculated by model 2 and overall survival time in ACC, and time dependent ROC analyses at 1, 3, and 5 years in GSE19750; **(D)** Survival analysis of the association between risk score by model 2 and event-free survival time in ACC, and time dependent ROC analyses at 1, 3, and 5 years in GSE76021.

were associated with CSR LATE UP.V1 UP, RB P107 DN.V1 UP, E2F1 UP.V1 UP, RPS14 DN.V1 DN, GCNP SHH UP LATE.V1 UP, and VEGF A UP.V1 DN.

Model 2 Was Regarded as the Best Model With Better Clinical Utility

Both of the two candidate models showed great prognostic values of patients with ACC according to previous analyses. In order to distinguish the best prognostic model, we further constructed two nomograms separately according to risk scores of model 1 (**Supplementary Figure 9A**) and model 2 (**Figure 5A**) and other significant features in multivariate Cox analysis. Both of the two models showed good potential for clinical application. According to the result of the calibration curves (**Figures 5B–D** and **Supplementary Figures 9B-D**), the nomogram of model 2 showed better value compared with the ideal model and model 1. Especially for nomogram's 3-year and 5-year OS, the nomogram showed great performance almost as the same as the ideal model did (**Figures 5C,D**). Thus, we regarded model 2 ($\alpha = 0.6$) as the best prognostic model in the present study.

Validation of the Nomogram Assessed by the Best Model

In order to validate the predictive value of the nomogram based on the best model, we further performed DCA, C-index, and ROC analyses. DCA was performed to evaluate the clinical net benefit using the nomogram with and without risk score (calculated by the best model) for predicting 1-, 3-, and 5-year survival probability. As **Figure 5E** showed, there was a trend that the nomogram based on the best model occupied higher net benefit than the nomogram without risk score when predicting 1-year survival probability, when Pt ranged between 0 and 0.15. For the evaluation of the clinical net benefit using the nomogram based on the best model for predicting 3-year survival probability, the nomogram with the best model had higher net benefit compared with the nomogram without risk score, when Pt ranging from 0 to 0.50 (**Figure 5F**). The nomogram with the best model also obviously improved the net benefit for 5-year survival prediction. In detail, the nomogram with the best model had a higher net benefit than the simple nomogram of Pt between 0 to 0.20 (**Figure 5G**). Summarizing

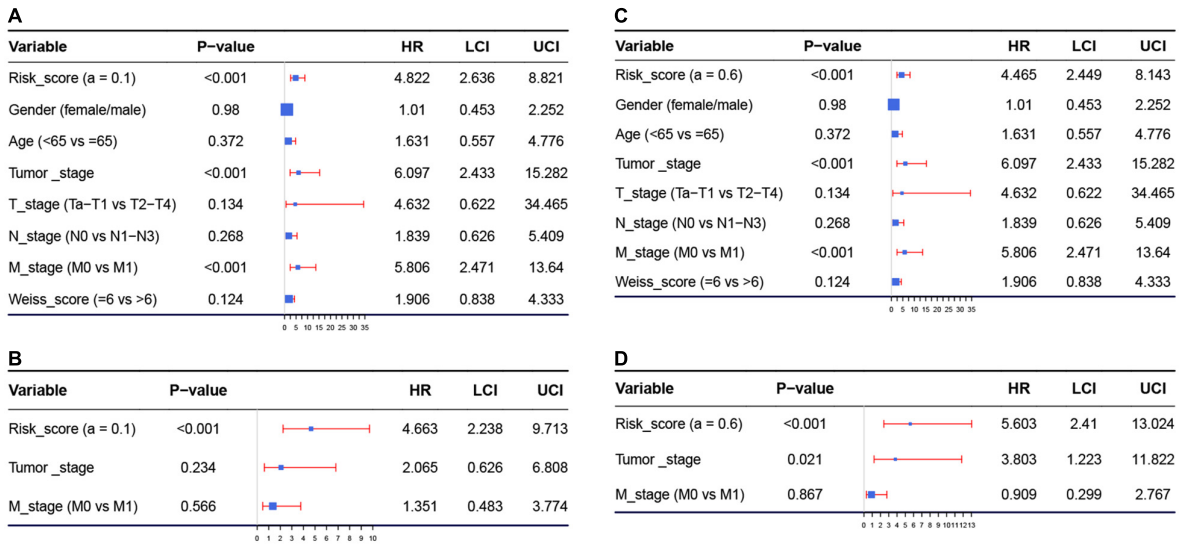


FIGURE 4 | (A) Forest plot summary of analyses of overall survival (OS), univariate analysis of risk score calculated by model 1, gender, age, tumor stage, TNM stage, and Weiss score by using The Cancer Genome Atlas (TCGA)–ACC data. **(B)** Forest plot summary of analyses of OS, univariate analysis of risk score calculated by model 1, tumor stage, and M stage by using TCGA–ACC data. **(C)** Forest plot summary of analyses of OS, univariate analysis of risk score calculated by model 2, gender, age, tumor stage, TNM stage, and Weiss score by using TCGA-ACC data. **(D)** Forest plot summary of analyses of OS, univariate analysis of risk score calculated by model 2, tumor stage, and M stage by using TCGA-ACC data. HR, hazard ratio; OS, overall survival.

above, the nomogram assessed by the best model showed high potential for clinical application, especially for 1- and 3-year survival prediction. ROC analysis using TCGA-ACC data indicated that the nomogram could predict OS of ACC patients effectively (AUC: 0.891; C-index: 0.872; **Figure 5H**). Time-dependent ROC curves demonstrated that this nomogram with the best model showed excellent stability over a period of 5 years (1-year AUC: 0.838, 3-year AUC: 0.942, 5-year AUC: 0.961, **Figure 5I**). We directly used the nomogram to divide the ACC patients into high- and low-point groups and survival analysis demonstrated that ACC patients with higher nomogram points had better OS compared with patients with lower nomogram points ($P < 0.0001$, **Figure 5J**). We also validated the predication value of this nomogram by using GSE19750, we got the similar conclusion that the nomogram with the best model could predict OS of ACC patients effectively (AUC: 0.864; C-index: 0.882; **Figure 5K**). Time-dependent ROC curves demonstrated that this nomogram based on the best model showed good stability over a period of 5 years (1-year AUC: 0.897, 3-year AUC: 0.882, 5-year AUC: 0.755, **Figure 5L**). Survival analysis also demonstrated that ACC patients in highnomogram point group had better OS compared with ACC patients in low-nomogram point group ($P = 0.029$, **Figure 5M**).

A Summary of Mutations and Copy Number Variations of Genes in the Best Model

The best model contained 14 most powerful prognostic markers [ASPM (abnormal spindle microtubule assembly), AURKA (aurora kinase A), CCNB2 (cyclin B2), CDC20 (cell division cycle 20), CENPA (centromere protein A), EXO1 (exonuclease 1), FBXO5 (F-box protein 5), HJURP (Holliday junction recognition protein), KIF2C (kinesin family member 2C), MKI67 (marker of proliferation Ki-67), NUF2 (NUF2 component of NDC80 kinetochore complex), PARPBP (PARP1 binding protein), TACC3 (transforming acidic coiled-coil containing protein 3), and TROAP (trophinin associated protein)], which were selected for further analysis. Among the 92 ACC patients with sequencing data, only in 10 independent samples contained mutations of genes as in the best model (**Table 2**). As for genetical alteration of the 14 genes, MKI67 altered the most (12%), and the main type was mRNA High (**Supplementary Figure 10**). Combined with the relative mRNA expression values of these genes, these genes seemed to highly express when there exist alterations in them (**Supplementary Figure 10**). Among the 90 ACC samples with CNV data, CNVs of the 14 genes constantly existed (**Figure 6A** and **Table 3**). Among the 14 genes, PARPBP was the most associated gene, which had the highest frequency of CNV events (77.78%, 70/90) (**Figure 6A** and **Table 3**). Moreover, gain of copy number was the most common CNV event (62.32%, 344/552, **Figure 6B**).

Relationship Between Alterations in Genes in the Best Model and the Clinical Features

Based on the data from TCGA database, six clinical features (gender, age, pathologic stage, T stage, N stage, and M stage) were collected for this analysis. As the result suggested (**Table 4**), alterations in genes were significantly associated with pathologic stage (Chi-square = 10.644, $P = 0.014$), T stage (Chi-square = 11.008, $P = 0.012$), and M stage (Chi-square = 10.687, $P = 0.001$).

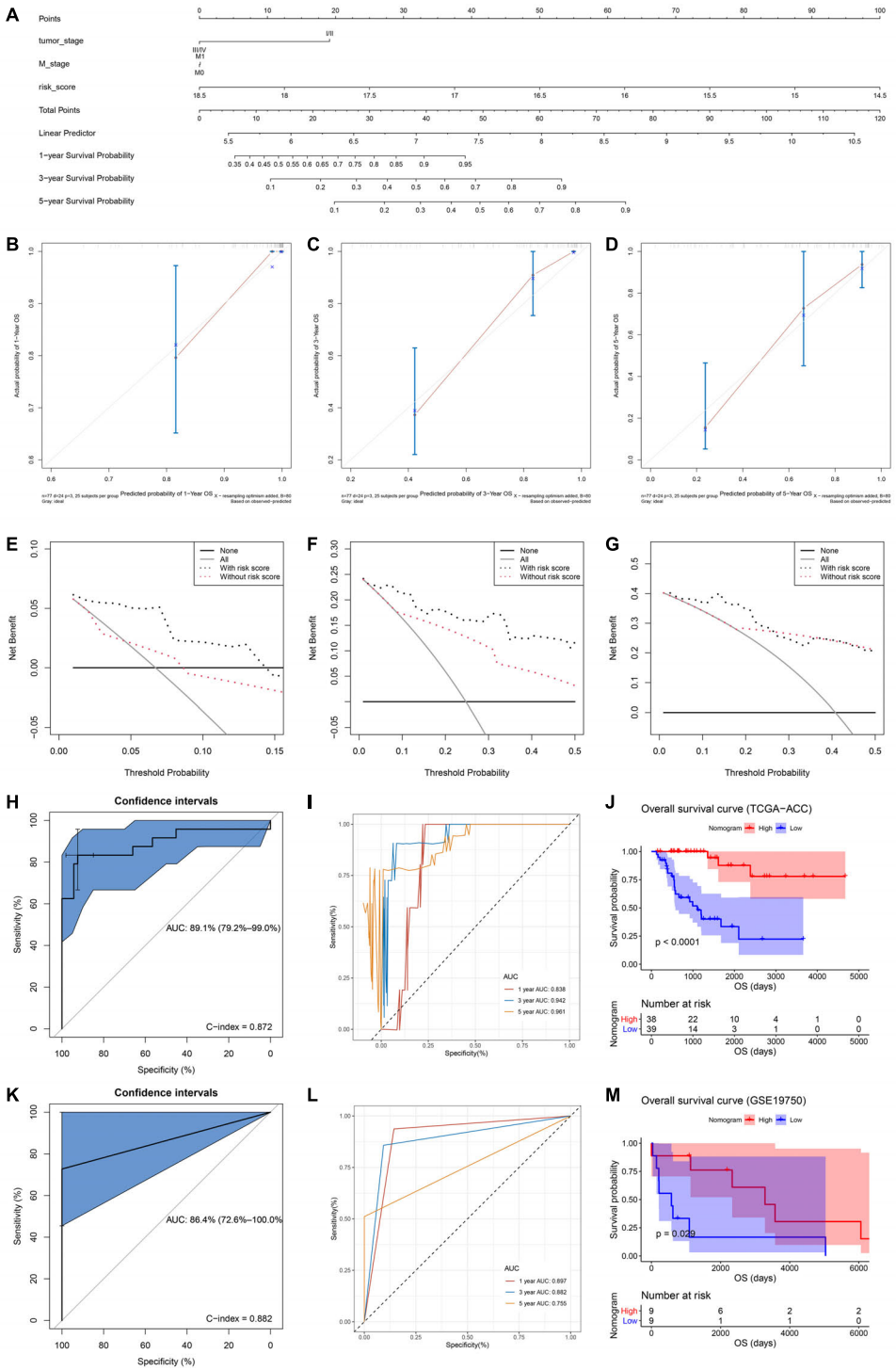


FIGURE 5 | (A) The nomogram constructed with model 2 for predicting proportion of patients with 1-, 3-, or 5-year OS. The calibration plots for predicting 1- **(B)**, 3- **(C)**, or 5- **(D)** year OS. Decision curve analysis (DCA) for assessment of the clinical utility for 1- **(E)**, 3- **(F)**, or 5- **(G)** year OS of model 2, the x-axis represents the percentage of threshold probability, and the y-axis represents the net benefit. OS, overall survival; DCA, decision curve analysis. **(H)** Receiver operating characteristic (ROC) curves and area under the curve (AUC) statistics to evaluate the diagnostic efficiency of the nomogram based on model 2 in TCGA–ACC data. **(I)** Time-dependent ROC curves indicating the predictive accuracy of the nomogram based on model 2 for 1-, 3-, or 5-year OS based on TCGA–ACC data. **(J)** Survival analysis of the association between risk score calculated by model 2 and overall survival time in ACC using TCGA-ACC data. **(K)** Receiver operating characteristic (ROC) curves and area under the curve (AUC) statistics to evaluate the diagnostic efficiency of the nomogram based on model 2 in GSE19750. **(L)** Time-dependent ROC curves indicating the predictive accuracy of the nomogram based on model 2 for 1-, 3-, or 5-year OS based on GSE19750. **(M)** Survival analysis of the association between risk score calculated by model 2 and overall survival time in ACC using GSE19750.

TABLE 2 | Mutations of genes in model 2 in adrenocortical carcinoma (ACC) patients from The Cancer Genome Atlas (TCGA) database.

ACC Sample ID	ASPM	AURKA	CDC20	EXO1	FBXO5	HJURP	KIF2C	MKI67	TACC3
TCGA-PK-A5HB	I1717*								
TCGA-OU-A5PI	Q2415H								
TCGA-OR-A5K4	R3198C		V361I						
TCGA-OR-A5JA		V27M		C695F				D2208Y	D709Y
TCGA-OR-A5JC					L127Ifs*6				
TCGA-OR-A5L3						P459L			
TCGA-OR-A5K9							T575I		
TCGA-OR-A5JP								R2786Q	
TCGA-OR-A5LP								T284N	
TCGA-OR-A5JG								P3145H	

*ACC, adrenocortical carcinoma. *a special mutation type of ASPM.*

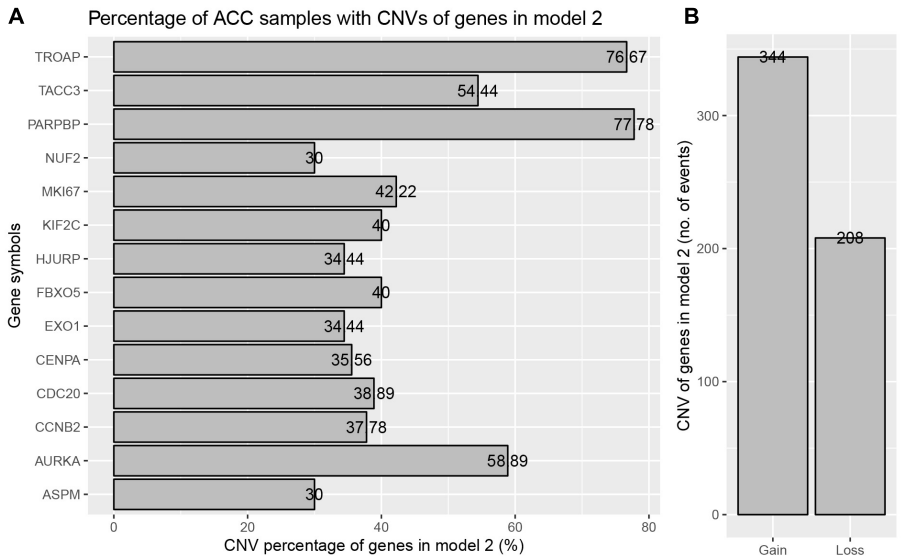


FIGURE 6 | Copy number variations (CNVs) of genes in model 2 in ACC. **(A)** Percentage of ACC samples with CNVs of genes in model 2 based on TCGA-ACC data. **(B)** Events of copy number gain or loss of genes in model 2 in ACC samples.

TABLE 3 | Different copy number variation (CNV) patterns occur in ACC samples from TCGA database.

	Diploid	Deep deletion	Shallow deletion	Copy number gain	Amplification	CNV sum	Percentage
ASPM	63	0	13	13	1	27	30.00%
AURKA	37	0	3	49	1	53	58.89%
CCNB2	56	0	22	11	1	34	37.78%
CDC20	55	0	32	2	1	35	38.89%
CENPA	58	0	18	14	0	32	35.56%
EXO1	59	0	15	16	0	31	34.44%
FBXO5	54	0	19	16	1	36	40.00%
HJURP	59	2	16	13	0	31	34.44%
KIF2C	54	0	32	3	1	36	40.00%
MKI67	52	0	10	28	0	38	42.22%
NUF2	63	0	11	14	2	27	30.00%
PARPBP	20	1	6	63	0	70	77.78%
TACC3	41	0	10	37	2	49	54.44%
TROAP	21	0	1	65	3	69	76.67%

TABLE 4 | Clinical pathological parameters of ACC patients with or without mutation/CNV of genes in model 2.

		With mutation and/or CNV	Without mutation and CNV	Total number	Chi-square	*P*-value
Age	≥65	6	8	14	0.200	0.655
	<65	24	50	74		
Gender	Female	19	40	59	0.284	0.594
	Male	11	18	29		
Pathologic stage	I	3	6	9	10.644	0.014
	II	10	33	43		
	III	5	13	18		
	IV	12	6	18		
T stage	T1	3	6	9	11.008	0.012
	T2	11	37	48		
	T3	3	8	11		
	T4	13	7	20		
N stage	N0	24	54	78	2.195	0.138
	N1	6	4	10		
	Nx	0	0	0		
M stage	M0	18	52	70	10.687	0.001
	M1	12	6	18		
	Mx	0	0	0		

With mutation or CNV: Cases have mutant or CNV or mutant + CNV, confirmed through TCGA database. Without mutation and CNV: Cases with neither mutant nor CNV, confirmed through TCGA database.

The Effects of Alterations in the 14 Genes on the Relative mRNA Expression Levels

According to the result, ASPM ($P < 0.0001$), CCNB2 ($P = 0.0437$), CDC20 ($P = 0.0451$), CENPA ($P < 0.0001$), EXO1 ($P = 0.0292$), HJURP ($P = 0.007$), KIF2C ($P = 0.0157$), MKI67 ($P = 0.0054$), NUF2 ($P = 0.0004$), PARPBP ($P = 0.0037$), TACC3 ($P < 0.0001$), and TROAP ($P = 0.0024$) were differently expressed among different types of CNVS. In detail, copy number gains of ASPM ($P < 0.001$, **Figure 7A**), CCNB2 ($P < 0.05$, **Figure 7C**), CENPA ($P < 0.0001$, **Figure 7E**), HJURP ($P < 0.01$, **Figure 7H**), PARPBP ($P < 0.01$, **Figure 7L**), and TACC3 ($P < 0.0001$, **Figure 7M**) were significantly associated with higher mRNA expression comparing with these in copy number shallow deletions (or deep deletions). As for AURKA (**Figure 7B**), EXO1 (**Figure 7F**), FBX05 (**Figure 7G**), MKI67 (**Figure 7J**), and NUF2 (**Figure 7K**), there was no significant expression difference between gains of copy number and shallow deletions of copy number. There was no copy number gain in CDC20, but samples with shallow deletion were significantly lower expressed compared with the diploid ones ($P = 0.0451$, **Figure 7D**). There were only one sample with copy number gain on KIF2C (**Figure 7I**) and one sample with copy number with shallow deletion of TROAP (**Figure 7N**), which perhaps meant that the result might not be convincing.

Correlation Between Copy Number Variations of the 14 Genes and Survival of Patients With Adrenocortical Carcinoma

In this step, survival analysis was performed to explore the prognostic value of CNVs in the 14 genes. According to the result, ACC patients with alterations in the 14 genes had poor OS ($P = 7.286E{-}4$, **Figure 8A**) and DFS ($P = 0.0131$, **Figure 8B**). Unfortunately, there was no association between patients with CNV or without CNV and OS ($P = 0.58$, **Figure 8C**). Then we separately explored the association between OS in each gene and CNVs. The result demonstrated that shallow deletions in ASPM ($P = 0.0170$, **Figure 8D**), CENPA ($P = 0.0071$, **Figure 8H**), EXO1 ($P = 0.0069$, **Figure 8I**), HJURP ($P = 0.0200$, **Figure 8K**), and NUF2 ($P = 0.0450$, **Figure 8N**) led to better OS of patients with ACC compared with those being affected by copy number gains, while patients with shallow deletions in AURKA ($P = 0.0150$, **Figure 8E**), MKI67 ($P = 0.0039$, **Figure 8M**), PARPBP ($P < 0.0001$, **Figure 8O**), and TACC3 ($P = 0.0016$, **Figure 8P**) had poorer OS. As for MKI67, patients affected by CNVs (shallow deletion or gain) had poorer OS compared with those affected by diploid ($P = 0.0039$, **Figure 8M**). As for CCNB2 ($P = 0.094$; **Figure 8F**), CDC20 ($P = 0.062$, **Figure 8G**), FBXO5 ($P = 0.190$, **Figure 8J**), and KIF2C ($P = 0.150$, **Figure 8L**), there were no significantly survival difference between different CNVs.

Validation of the 14 Genes in the Best Model

According to the result, all the classifiers constructed and verified by TCGA–ACC data showed good performance because the accuracy of classification was more than 0.70 (**Table 5**). Especially the polynomial-SVM-based classifier, its accuracy was 0.9870, and we validated all the classifiers by using GSE19750, GSE76019, and GSE76021. All the classifiers showed good performance in GSE76019 and GSE76021 (accuracy of classification ≥ 0.50, **Table 5**), but in GSE19750, KNN-based classifier and all the SVM-based classifiers did not play well as we expected (**Table 5**), perhaps because only 22 ACC samples from GSE19750 were

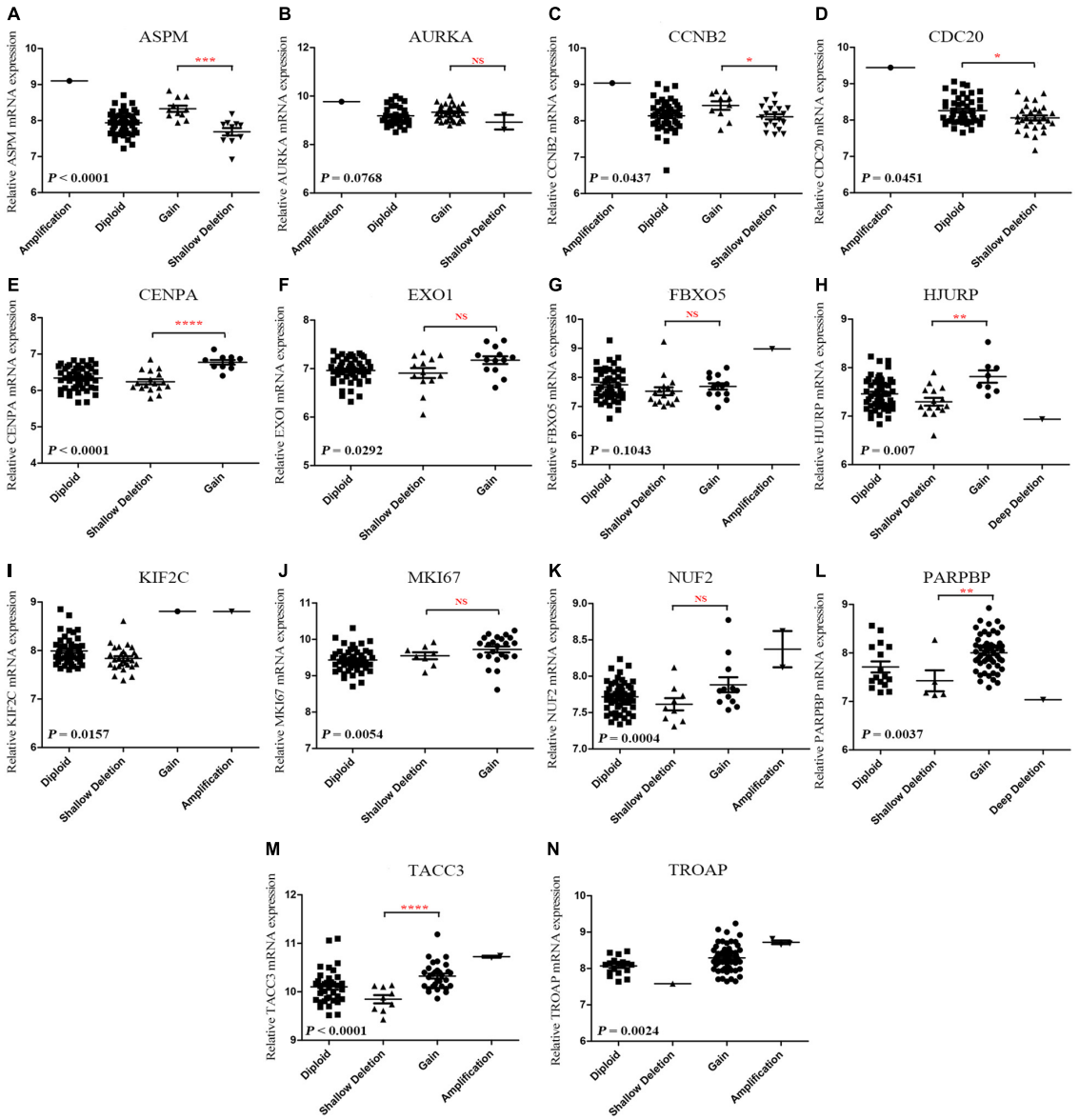


FIGURE 7 | (A-N) Correlation between different CNV patterns and mRNA expression levels of genes in model 2, respectively.

included for validation. All in all, all the classifiers constructed based on the 14 genes showed great performance to distinguish dead ACC samples from live ACC samples, which meant all the 14 genes were significantly associated with prognosis of patients with ACC.

Hub Genes in the Best Model

All the 14 genes from the best model were included in this PPI network (**Supplementary Figure 11A**). As shown in **Supplementary Figure 11B**, the degrees of six genes (ASPM, AURKA, CCNB2, CDC20, KIF2C, and NUF2) were higher than any other of the remaining eight genes (Degree = 13). Thus, the six genes were considered as hub genes in the best model. Expression of ASPM was significantly related to CCNB2, CDC20, FBXO5, HJURP, MKI67, NUF2, PARPBP, TACC3, and TROAP (**Supplementary Figure 11C**). Among them, MKI67 was the most related gene (**Supplementary Figure 11D**). Expression of AURKA was significantly associated with FBXO5, HJURP, MKI67, PARPBP, TACC3, and TROAP (**Supplementary Figure 11C**). Moreover, PARPBP was best associated with AURKA (**Supplementary Figure 11D**). Expression of CCNB2 was relevant to ASPM, CDC20, CENPA, FBXO5, HJURP, and PARPBP (**Supplementary Figure 11C**), and PARPBP was the most associated gene with CCNB2 (**Supplementary Figure 11D**). Expression of CDC20 was associated with CCNB2, ASPM, FBXO5, HJURP, KIF2C, MKI67, PARPBP, and TROAP (**Supplementary Figure 11C**), and the most related gene was HJURP (**Supplementary Figure 11D**). The expression of NUF2

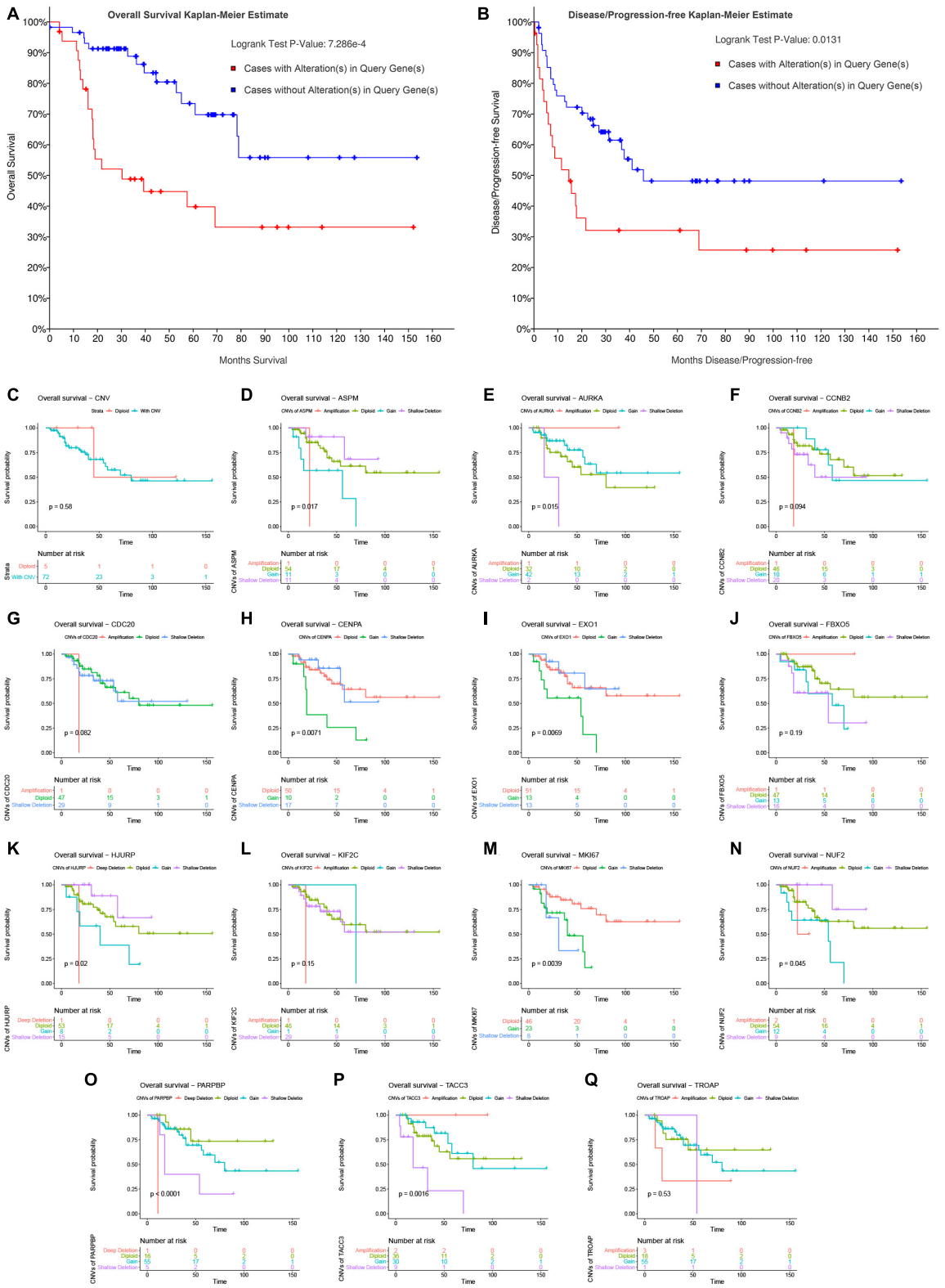


FIGURE 8 | Survival analysis of ACC patients with CNVs of genes in model 2 based on TCGA-ACC data. Overall survival **(A)** and disease-free survival **(B)** of patients with alterations in genes in model 2 or patients without alterations in genes in model 2. **(C)** Overall survival of patients with any CNVs of genes in model 2 or patients with diploid. Overall survival of patients with different CNV types of ASPM **(D)**, AURKA **(E)**, CCNB2 **(F)**, CDC20 **(G)**, CENPA **(H)**, EXO1 **(I)**, FBXO5 **(J)**, HJURP **(K)**, KIF2C **(L)**, MKI67 **(M)**, NUF2 **(N)**, PARPBP **(O)**, TACC3 **(P)**, and TROAP **(Q)**.

TABLE 5 | The accuracy of classification of linear discriminant analysis (LDA)-based classifier, K-nearest neighbor (KNN)-based classifier, linear-support vector machine (SVM)-based classifier, polynomial-SVM-based classifier, radial basis function (RBF)-SVM-based classifier, and sigmoid-kernel-SVM-based classifier.

		LDA	KNN	Kappa	Linear-SVM	Kappa	Polynomial-SVM	Kappa	RBF-SVM	Kappa	Sigmoid-kernel SVM	Kappa
TCGA-ACC	Training set	0.8701	0.8961	0.7463	0.8831	0.7111	0.9870	0.9694	0.9351	0.8469	0.8182	0.5664
GSE19750	Testing set 1	0.6364	0.2727	−0.0353	0.4091	0.0403	0.1818	0	0.2727	0.0435	0.4545	0.0704
GSE76019	Testing set 2	0.7647	0.7059	0.2056	0.7059	0.3561	0.6471	0	0.7353	0.3319	0.7647	0.5211
GSE76021	Testing set 3	0.6207	0.7241	0.7241	0.7931	0.5272	0.6897	0.2162	0.5862	−0.0235	0.7931	0.5272

LDA, Linear discriminant analysis; KNN, K-Nearest Neighbor; RBF, radial basis function; SVM, Support Vector Machine.

was significantly related to HJURP, FBXO5, CENPA, CDC20, AURKA, and ASPM (**Supplementary Figure 11C**). Among them, ASPM was the most associated gene (**Supplementary Figure 11D**).

External Validation of the Best Model/the Nomogram and Comparison With Other Biomarkers Using Gene Expression Data

In order to assess the prognostic value of the six hub genes, and compare the prognostic value of the best model/the nomogram with the prognostic biomarkers given by other studies, we collected several biomarkers from previous studies. In total, eight biomarkers including CTNNB1 (Assié et al., 2014), IGF2 (Catalano et al., 2021), TP53 (Assié et al., 2014), MKI67 (Beuschlein et al., 2015), SF1 (Tian et al., 2021), IPRPs model (Tian et al., 2021), m6A-related signature (Jin et al., 2021), and m6A-based signature (Shen et al., 2021) were used in the present study. According to the result (**Table 6**), all the hub genes could effectively predict the OS of ACC patients (ASPM: AUC = 0.840, C-index = 0.827; AURKA: AUC = 0.786, C-index = 0.794; CCNB2: AUC = 0.830, C-index = 0.821; CDC20: AUC = 0.833, C-index = 0.838; KIF2C: AUC = 0.819, C-index = 0.823; NUF2: AUC = 0.811, C-index = 0.843). Moreover, the risk score calculated by the best model (AUC = 0.844, C-index = 0.851) and the nomogram based on the best model (AUC = 0.891, C-index = 0.872) showed better ability for predicting prognosis than those biomarkers collected from previous studies. AUC and C-index of these biomarkers are shown in **Table 6**. To summarize the above, the results demonstrated that the best model and the nomogram based on the best model may act better than other biomarkers.

TABLE 6 | Comparison the best model vs. the nomogram with other biomarkers using TCGA–ACC data.

Source	Biomarker evaluation	AUC	C-index
	ASPM	0.840	0.827
	AURKA	0.786	0.794
	CCNB2	0.830	0.821
	CDC20	0.833	0.838
	KIF2C	0.819	0.823
	NUF2	0.811	0.843
	Risk score (a = 0.6)	0.844	0.851
	Nomogram	0.891	0.872
PMID: 24747642	CTNNB1	0.657	0.613
PMID: 33075426	IGF2	0.562	0.598
PMID: 24747642	TP53	0.565	0.603
PMID: 25559399	MKI67	0.818	0.859
PMID: 33626208	SF1	0.635	0.621
PMID: 33626208	IPRPs model	0.772	0.806
PMID: 33746977	m6A-related signature	0.818	0.778
PMID: 33692753	m6A based signature	0.744	0.696

DISCUSSION

Adrenocortical carcinoma is a relatively rare malignancy in the urinary system associated with a poor prognosis (Lam, 2021). 5-year survival rate for patients with ACC was 50% as ASCO reported. Similarly, it is not optimistic that 40% of the cases have distant metastasis when diagnosed (Ayala-Ramirez et al., 2013; Guillaume et al., 2014), which means that they have missed the best opportunity for treatment. Thus, there is of great need to develop novel molecular biomarkers for diagnosis and prognosis of ACC patients.

Over the past several years, the discoveries of novel and informative genes by using bioinformatics methods have provided valuable information in the diagnosis of malignancies (He et al., 2017; Meulekamp et al., 2017). At present, WGCNA has been widely used to screen out novel and effective molecular biomarkers in the bioinformatics field (Gu et al., 2019; Li et al., 2019; Lv et al., 2021). In our previous study, we have made great efforts to use WGCNA as the main method to identify prognosis biomarkers for bladder cancer (Yan et al., 2019). Thus, we first identified genes associated with survival of ACC in the present study, and 93 biomarkers significantly relevant to survival of ACC were further identified combined with overall survival analysis and disease-free survival analysis. Interestingly, some studies thought that a single biomarker might not predict the prognosis of patients with ACC well (Chen et al., 2019). Nowadays, more and more studies have paid their attention on developing a multi-gene signature for predicting prognosis in malignancies instead of just a single biomarker (Qiao et al., 2019; Wu et al., 2019). Given that the Ridge, Elastic Net, and Lasso methods have been widely used to construct a Cox proportional hazard regression model (Lu et al., 2019), 11 multiple-gene signatures were constructed for predicting prognosis of ACC patients. Two candidate models were further screened out and validated by calculating the AUC by using internal validation set. To test whether the prognostic value of the two signatures were independent of other clinical features (gender, age, tumor stage, T stage, N stage, M stage, and Weiss score), we included these factors for the Cox

regression analysis. The result indicated that tumor stage, risk score of model 1, and risk score of model 2 were significant prognostic factors for ACC patients. Moreover, risk score of model 1 and risk score of model 2 were independent of tumor stage. Tumor stage including TNM staging system reflects the internal characteristics of tumors, which is of great value for evaluation of the degree of tumor deterioration and prediction of prognosis according to previous findings (Veeratterapillay et al., 2012). In our study, model 1 and model 2 could not only discriminate between low- and high-risk groups like tumor stage did, but also showed more effective prognostic value compared with tumor stage, which made our results reliable. To better understand the biological role of the two signatures, we performed GSEA analysis. Two KEGG pathways associated with model 1 were identified including cell cycle and oocyte meiosis. Meanwhile, one KEGG pathway associated with model 2 was identified including cell cycle. Cell cycle was the basic process of cell proliferation. According to a study by Kaistha et al. (2015), the deregulation of cell cycle was obviously associated with tumor progression. As for oocyte meiosis, maturation and development of oocyte is a complex biologic process in mammals, which is the key event in the reproductive process (Jaffe and Egbert, 2017). Any errors in this process can lead to the oocyte developmental abnormalities or infertility (Jaffe and Egbert, 2017). Meiosis is specialized cell division with the reduction in genetic content, which guaranteed the stability of chromosome numbers of gamogenetic organisms (Jaffe and Egbert, 2017). Subsequent analysis demonstrated that model 2 showed better predictive value than model 1; thus, we regarded model 2 as the best model in the present study. Moreover, the best model was validated to show strong association with survival of ACC patients and high accuracy for OS probability prediction in patients with ACC.

By exploring the association between mutations and CNVs with genes in the best model, we found that high expression levels of genes in the best model was associated with gene alteration. Subsequent analysis indicated that shallow deletions in ASPM, CENPA, EXO1, HJURP, and NUF2 caused better OS of ACC patients. Meanwhile, shallow deletions in AURKA, MKI67, PARPBP, and TACC3 led to poorer OS. To summarize the above, we found that alterations in these biomarkers might have strong effects on the survival of ACC patients. A PPI network was constructed based on the 14 genes from the best model; six hub genes were further identified according to the degree of connectivity. All the hub genes could effectively predict the OS of ACC patients. In the past few years, some potential prognostic biomarkers for ACC were identified. Tian et al. constructed a robust prognostic model for ACC (Tian et al., 2021). Two m6A-related signatures, which might be used as independent prognostic factors in evaluating the prognosis of ACC patients, were also established (Jin et al., 2021; Shen et al., 2021). Among the six hub genes we identified, two, including ASPM (Yuan et al., 2018) and CCNB2 (Gao et al., 2019; Guo et al., 2020), were already regarded as prognostic biomarkers in previous studies. Yuan et al. screened out 12 hub genes including ASPM, which were associated with the progression and prognosis of ACC. Two recent studies proved that CCNB2 was associated with worse OS of ACC (Gao et al., 2019; Guo et al., 2020). As for the rest of the four genes, the relationship between them and prognosis of ACC has never been reported by previous studies, which might be novel prognostic biomarkers for ACC. Moreover, the risk score calculated by the best mode and the nomogram based on the best model showed better ability in predicting prognosis than these biomarkers collected from previous studies.

There still exist some limitations in our study. First, though our study was a relatively preliminary exploration of establishing a multi-gene signature by combing bioinformatics methods, statistics, and machine learning, there might be lack of novelty in methods. Therefore, we will explore more novel bioinformatics and machine learning methods to integrate prognostic gene signatures. Second, the best prognostic model was developed and validated by using data derived from TCGA and GEO databases. Although the signature could distinguish high- and low-risk groups well, it was not clear if it could show good performance as we expected in clinical trials. Thus, we will further apply and test this signature in clinical judgment for prognosis of ACCs. Third, according to the results of DCA, the best signature showed high potential for 5-year survival prediction when $0.00 < Pt < 0.20$, but when Pt equaled other values, this signature did not play well as we expected, perhaps due to the small size of these datasets. Therefore, we will further validate the multi-gene signature in larger and richer datasets.

To sum up, 93 hub genes in the co-expression network were identified as survival associated genes. Moreover, we constructed an Elastic Net model including 14 genes, which were validated to show good performance by using three independent datasets. This signature could act as an effective prediction tool for the prognosis of ACC patients independently. The best model-based nomogram was established based on risk score assessed by this risk signature to provide clinical doctors with a visual tool. Moreover, six hub genes were identified among the 14 genes from the best model, which might be novel prognostic biomarkers of prognosis of patients with ACC. However, the best model and six hub genes needs to be verified by using more novel bioinformatics methods, clinical trials, and larger datasets.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

SL and T-ZL conceived and designed the study. XY, Z-XG, D-HY, and X-PL performed the analysis procedures. XY, Z-XG, CC, and X-PL analyzed the results. SL, Z-WY, and T-ZL contributed analysis tools. XY contributed to the writing of the manuscript. All authors reviewed the manuscript.

FUNDING

This work was supported by the Medical Science and Technology Innovation Platform support project of Zhongnan Hospital of Wuhan University (PTXM2019006), the Young & Middle-aged Medical Key Talents Training Project of Wuhan (WHQG201901), the Science and Technology Department of Hubei Province Key Project (2018ACA159) Zhongnan Hospital of Wuhan University Science, Technology and Innovation Seed Fund (znpy2017050), the 351 Talent Project of Wuhan University (Luojia Young Scholars: SL), and the National Natural Science Foundation of China (81802541).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.671359/full#supplementary-material

Supplementary Figure 1 | Flow chart indicating the process used to select target genes included in the analysis.

Supplementary Figure 2 | (A) Samples clustering to detect outliers (GSE76019). **(B)** The clustering was based on the expression data of GSE76019. The top 2,500 genes with the highest SD values were used for the analysis by WGCNA. The color intensity was proportional to tumor stage, survival years (survival time), and survival status.

Supplementary Figure 3 | Determination of soft-thresholding power in the weighted gene co-expression network analysis (WGCNA). **(A)** Analysis of the scale-free fit index for various soft-thresholding powers (β). **(B)** Analysis of the mean connectivity for various soft-thresholding powers. **(C)** Histogram of connectivity distribution when $\beta = 5$. **(D)** Checking the scale free topology when $\beta = 5$.

Supplementary Figure 4 | The cluster dendrogram of genes in GSE76019. Each branch in the figure represents one gene, and every color below represents one co-expression module.

Supplementary Figure 5 | (A) Interaction relationship analysis of co-expression genes. Different colors of horizontal axis and vertical axis represent different modules. The brightness of yellow in the middle represents the degree of connectivity of different modules. There was no significant difference in interactions among different modules, indicating a high-scale independence degree among these modules. **(B)** Classical MDS plot whose input is the TOM dissimilarity. Each dot (gene) is colored by the module assignment.

Supplementary Figure 6 | Bioinformatics analysis of genes in the hub modules. **(A)** Biological process of genes in the red module. **(B)** Biological process of genes in the blue module. **(C)** KEGG pathway enrichment of genes in the blue module.

Supplementary Figure 7 | Ten-fold cross-validation for tuning parameter selection in the Ridge **(A)**, ELASTIC-NET with α varying from 0.1 to 0.9 **(E–J)**, and lasso model **(K)**. The solid vertical lines are the partial likelihood deviance SE. The dotted vertical lines are drawn at the optimal values by minimum criteria and 1-SE criteria.

Supplementary Figure 8 | Distribution of risk scores of Ridge **(A)**, ELASTIC-NET with α varying from 0.1 to 0.9 **(E–J)**, and lasso model **(K)** of ACC patients based on TCGA-ACC data. The number of survivors and non-survivors with different risk scores based on TCGA-ACC data; red represents the number of non-survivors, and blue represents the number of survivors.

Supplementary Figure 9 | (A) The nomogram constructed with model 1 for predicting proportion of patients with 1-, 3- or 5-year OS. The calibration plots for predicting 1- **(B)**, 3- **(C)** or 5- **(D)** year OS. OS: overall survival.

Supplementary Figure 10 | Genetic alterations associated with genes in model 2 and expression heatmap of genes in model 2 based on the data from TCGA.

Supplementary Figure 11 | Interaction among genes in model 2 and spearman correlation analysis of the genes in model 2. **(A)** Protein-protein interaction (PPI) network of genes in model 2. **(B)** Degree of connectivity of genes in model 2 according to the PPI network. **(C)** Spearman correlation analysis of the genes in model 2. **(D)** A visual summary for the spearman correlation analysis (scatter plot, bar plot of expression levels of genes in model 2, and spearman correlation coefficients).

Supplementary Table 1 | Clinical characteristics of patients with ACC in each study.

Supplementary Table 2 | GO biological processes of genes in blue module.

Supplementary Table 3 | Survival analysis of hub genes in co-expression network.

Supplementary Table 4 | The corresponding regression coefficients of each gene in models constructed by the Ridge, ELASTIC-NET, and LASSO methods.

Supplementary Table 5 | Gene set enrichment analysis of model 1 and model 2 by using GSE76019.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29.

Assié, G., Letouzé, E., Fassnacht, M., Jouinot, A., Luscap, W., Barreau, O., et al. (2014). Integrated genomic characterization of adrenocortical carcinoma. *Nat. Genet.* 46, 607–612. doi: 10.1038/ng.2953

Aut, M. G., and Aut, T. L. (2016). *Forestplot: Advanced Forest Plot using 'grid' Graphics*. Available online at: http://CRAN.R-project.org/package=forestplot (accessed December 12, 2020).

Ayala-Ramirez, M., Jasim, S., Feng, L., Ejaz, S., Deniz, F., Busaidy, N., et al. (2013). Adrenocortical carcinoma: clinical outcomes and prognosis of 330 patients at a tertiary care center. *Eur. J. Endocrinol.* 169, 891–899. doi: 10.1530/EJE-13-0519

Beuschlein, F., Weigel, J., Saeger, W., Kroiss, M., Wild, V., Daffara, F., et al. (2015). Major prognostic role of ki67 in localized adrenocortical carcinoma after complete resection. *J. Clin. Endocrinol. Metab.* 100, 841–849. doi: 10.1210/jc.2014-3182

Blanche, P., Dartigues, J. F., and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat. Med.* 32, 5381–5397. doi: 10.1002/sim.5958

Catalano, R., Giardino, E., Treppiedi, D., Mangili, F., Morelli, V., Elli, F. M., et al. (2021). The cytoskeleton actin binding protein filamin a impairs both igf2 mitogenic effects and the efficacy of igf1r inhibitors in adrenocortical cancer cells. *Cancer Lett.* 497, 77–88. doi: 10.1016/j.canlet.2020.10.022

Chen, L., Luo, Y., Wang, G., Qian, K., Qian, G., Wu, C. L., et al. (2019). Prognostic value of a gene signature in clear cell renal cell carcinoma. *J. Cell Physiol.* 234, 10324–10335. doi: 10.1002/jcp.27700

Demeure, M. J., Coan, K. E., Grant, C. S., Komorowski, R. A., Stephan, E., Sinari, S., et al. (2013). Pttg1 overexpression in adrenocortical cancer is associated with poor survival and represents a potential therapeutic target. *Surgery* 154:1416. doi: 10.1016/j.surg.2013.06.058

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Gao, Z., Man, X., Li, Z., Bi, J., Liu, X., Li, Z., et al. (2019). Expression profiles analysis identifies the values of carcinogenesis and the prognostic prediction of three genes in adrenocortical carcinoma. *Oncol. Rep.* 41, 2440–2452. doi: 10.3892/or.2019.7021

Gu, H. Y., Yang, M., Guo, J., Zhang, C., Lin, L. L., Liu, Y., et al. (2019). Identification of the biomarkers and pathological process of osteoarthritis: weighted gene co-expression network analysis. *Front. Physiol.* 10:275. doi: 10.3389/fphys.2019.00275

Guillaume, A., Eric, L., Martin, F., Anne, J., Windy, L., Barreau, O., et al. (2014). Integrated genomic characterization of adrenocortical carcinoma. *Nat. Genet.* 46, 607–612.

Guo, J., Gu, Y., Ma, X., Zhang, L., Li, H., Yan, Z., et al. (2020). Identification of hub genes and pathways in adrenocortical carcinoma by integrated bioinformatic analysis. *J. Cell Mol. Med.* 24, 4428–4438. doi: 10.1111/jcmm.15102

He, Z., Sun, M., Ke, Y., Lin, R., Xiao, Y., Zhou, S., et al. (2017). Identifying biomarkers of papillary renal cell carcinoma associated with pathological stage by weighted gene co-expression network analysis. *Oncotarget* 8, 27904–27914. doi: 10.18632/oncotarget.15842

Hechenbichler, K., and Schliep, K. (2004). *Weighted k-nearest-neighbor Techniques and Ordinal Classification*. Discussion Paper Sfb. 399: 386.

Jaffe, L. A., and Egbert, J. R. (2017). Regulation of mammalian oocyte meiosis by intercellular communication within the ovarian follicle. *Annu. Rev. Physiol.* 79, 237–260. doi: 10.1146/annurev-physiol-022516-034102

Jin, Y., Wang, Z., He, D., Zhu, Y., Hu, X., Gong, L., et al. (2021). Analysis of m6a-related signatures in the tumor immune microenvironment and identification of clinical prognostic regulators in adrenocortical carcinoma. *Front. Immunol.* 12:637933. doi: 10.3389/fimmu.2021.637933

Kaistha, B. P., Lorenz, H., Schmidt, H., Sipos, B., Pawlak, M., Gierke, B., et al. (2015). Plac8 localizes to the inner plasma membrane of pancreatic cancer cells and regulates cell growth and disease progression through critical cell-cycle regulatory pathways. *Cancer Res.* 76, 96–107. doi: 10.1158/0008-5472.can-15-0216

Kanehisa, M., and Goto, S. (2000). Kegg: kyoto encyclopaedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). Kernlab – an S4 package for Kernel methods in R. *J. Stat. Softw.* 11:9. doi: 10.18637/jss.v011.i09

Kuhn, M. (2015). Caret: classification and regression training. *Astrophysics Source Code Library* 129, 291–295.

Lam, A. K. (2021). Adrenocortical carcinoma: updates of clinical and pathological features after renewed world health organisation classification and pathology staging. *Biomedicines* 9:175. doi: 10.3390/biomedicines9020175

Langfelder, P., and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi: 10.1093/bioinformatics/bts034

Li, Q., Chen, W., Song, M., Chen, W., Yang, Z., and Yang, A. (2019). Weighted gene co-expression network analysis and prognostic analysis identifies hub genes and the molecular mechanism related to head and neck squamous cell carcinoma. *Cancer Biol. Ther.* 20, 750–759. doi: 10.1080/15384047.2018.1564560

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lu, H. X., Du, J., Wen, D. L., Sun, J. H., Chen, M. J., Zhang, A. Q., et al. (2019). Development and validation of a novel predictive score for sepsis risk among trauma patients. *World J. Emerg. Surg.* 14:11. doi: 10.1186/s13017-019-0231-8

Lv, D., Wu, X., Wang, M., Chen, W., Yang, S., Liu, Y., et al. (2021). Functional assessment of four novel immune-related biomarkers in the pathogenesis of clear cell renal cell carcinoma. *Front. Cell Dev. Biol.* 9:621618. doi: 10.3389/fcell.2021.621618

Meulekamp, M. Z., van der Wurff, P., van der Meer, A., and Lucas, C. (2017). Identifying prognostic factors for conservative treatment outcomes in servicemen with chronic exertional compartment syndrome treated at a rehabilitation center. *Mil. Med. Res.* 4:36. doi: 10.1186/s40779-017-0145-2

Meyer, D. (2015). Support vector machines: the interface to libsvm in package e1071. *R News* 1, 1–3.

Michael, J. P., and Ralph, B. D. (2010). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat. Med.* 23, 2109–2123. doi: 10.1002/sim.1802

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Zivot, E., et al. (2014). *Performanceanalytics: Econometric Tools for Performance and Risk Analysis*. Available online at: https://CRAN.R-project.org/package=PerformanceAnalytics (accessed February 6, 2020).

Pinto, E. M., Rodriguez-Galindo, C., Choi, J. K., Pounds, S., Liu, Z., Neale, G., et al. (2016). Prognostic significance of major histocompatibility complex class ii expression in pediatric adrenocortical tumors: a st. Jude and children's oncology group study. *Clin. Cancer Res.* 22, 6247–6255. doi: 10.1158/1078-0432.CCR-15-2738

Qiao, G. J., Chen, L., Wu, J. C., and Li, Z. R. (2019). Identification of an eight-gene signature for survival prediction for patients with hepatocellular carcinoma based on integrated bioinformatics analysis. *PeerJ* 7:e6548. doi: 10.7717/peerj.6548

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). Proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Rodriguez-Galindo, C., Krailo, M. D., Pinto, E. M., Pashankar, F., Weldon, C. B., Huang, L., et al. (2021). Treatment of pediatric adrenocortical carcinoma with surgery, retroperitoneal lymph node dissection, and chemotherapy: the children's oncology group arar0332 protocol. *J. Clin. Oncol.* doi: 10.1200/JCO.20.02871 [Epub ahead of print].

Sachs, M. C. (2015). *Plotroc: Generate Useful ROC Curve Charts for Print and Interactive Use*.

Shen, C., Liu, J., Yang, X., Jiao, W., and Wang, Y. (2021). Development and validation of an m6a rna methylation regulators-based signature for predicting the prognosis of adrenocortical carcinoma. *Front. Endocrinol. (Lausanne)* 12:568397. doi: 10.3389/fendo.2021.568397

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102:15545–15550 doi: 10.1073/pnas.0506580102

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., Zhang, Z., et al. (2017). Gepia: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102.

Therneau, T. M. (2015). Survival: survival analysis. *Technometrics* 46, 111–112.

Tian, X., Xu, W. H., Anwaier, A., Wang, H. K., Wan, F. N., Cao, D. L., et al. (2021). Construction of a robust prognostic model for adult adrenocortical carcinoma: results from bioinformatics and real-world data. *J. Cell Mol. Med.* 25, 3898–3911. doi: 10.1111/jcmm.16323

Veeratterapillay, R., Simren, R., El-Sherif, A., Johnson, M. I., Soomro, N., and Heer, R. (2012). Accuracy of the revised 2010 tnm classification in predicting the prognosis of patients treated for renal cell cancer in the north east of england. *J. Clin. Pathol.* 65, 367–371. doi: 10.1136/jclinpath-2011-200468

Venables, W. N., and Ripley, B. D. (2002). Modern applied statistics with s. *Stat. Comput.* 52, 704–705.

Vickers, A., and Elkin, E. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* 26:565–574. doi: 10.1177/0272989x06295361

Wei, T., and Simko, V. (2013). Corrplot: visualization of a correlation matrix. *Mmwr Morb. Mortal. Wkly. Rep.* 52, 145–151.

Wu, J., Jin, S., Gu, W., Wan, F., Zhang, H., Shi, G., et al. (2019). Construction and validation of a 9-gene signature for predicting prognosis in stage iii clear cell renal cell carcinoma. *Front. Oncol.* 9:152. doi: 10.3389/fonc.2019.00152

Yan, X., Guo, Z. X., Liu, X. P., Feng, Y. J., Zhao, Y. J., Liu, T. Z., et al. (2019). Four novel biomarkers for bladder cancer identified by weighted gene coexpression network analysis. *J. Cell Physiol.* 234, 19073–19087. doi: 10.1002/jcp.28546

Ye, Y., Dai, Q., Li, S., He, J., and Qi, H. (2021). A novel defined risk signature of the ferroptosis-related genes for predicting the prognosis of ovarian cancer. *Front. Mol. Biosci.* 8:645845. doi: 10.3389/fmolb.2021.645845

Yizhou, W., Jun, L., Yong, X., Renyan, G., Kui, W., Yan, Z., et al. (2013). Prognostic nomogram for intrahepatic cholangiocarcinoma after partial hepatectomy. *J. Clin. Oncol.* 31, 1188–1195. doi: 10.1200/jco.2012.41.5984

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). Clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yuan, L., Qian, G., Chen, L., Wu, C. L., Dan, H. C., Yu, X., et al. (2018). Co-expression network analysis of biomarkers for adrenocortical carcinoma. *Front. Genet.* 9:328. doi: 10.3389/fgene.2018.00328

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Yan, Guo, Yu, Chen, Liu, Yang, Liu and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 08 June 2021
doi: 10.3389/fcell.2021.670890

Identify Function of WASL in Prognosis of Cervical Cancer Based on Omics Data

Jinxuan Hou[1†], Chen Chen[2,3†], Yingying Hu[4†], Qing Gong[5], Lijuan Gan[6] and Yu Xu[7]*

[1] Department of Thyroid and Breast Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China, [2] Human Genetics Resource Preservation Center of Wuhan University, Wuhan, China, [3] Human Genetics Resource Preservation Center of Hubei Province, Wuhan, China, [4] Department of Obstetrics and Gynecology, The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu, China, [5] Department of Obstetrics and Gynecology, Zhongnan Hospital of Wuhan University, Wuhan, China, [6] Department of Gynecological Oncology, Zhongnan Hospital of Wuhan University, Hubei Key Laboratory of Tumor Biological Behaviors, Hubei Cancer Clinical Study Center, Wuhan, China, [7] Department of Radiation and Medical Oncology, Zhongnan Hospital of Wuhan University, Wuhan, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Bishuang Cai,
Icahn School of Medicine at Mount Sinai, United States
Xin Huang,
Columbia University Irving Medical Center, United States

****Correspondence:***
Yu Xu
xuyu@whu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *22 February 2021*
Accepted: *09 April 2021*
Published: *08 June 2021*

Citation:
Hou J, Chen C, Hu Y, Gong Q, Gan L and Xu Y (2021) Identify Function of WASL in Prognosis of Cervical Cancer Based on Omics Data. Front. Cell Dev. Biol. 9:670890. doi: 10.3389/fcell.2021.670890

Objective: To clarify the clinical relevance of WASP like actin nucleation promoting factor (WASL) in patients with cervical cancer and associated mechanisms.

Methods and Materials: We obtained high prediction accuracy and determined the correlation between the expression of WASL and the clinical characteristics of cervical cancer patients. Differentially expressed genes (DEGs) were identified using microarray. Gene ontology (GO) enrichment analysis and gene set enrichment analysis (GSEA) were performed to determine potentially relevant mechanisms related to the prognostication ability of WASL expression.

Results: Chi-square test and multivariable logistic regression analysis suggested that lower expression of WASL was associated with lower pathological stage (chi-square test: p = 0.022, chi-square = 9.613; logistic regression: OR = 0.869, 95% CI: 0.756–0.991, p = 0.041). Patients in the WASL high expression group have worse overall survival (OS) [hazard ratio (HR): 0.555, 95% CI: 0.348–0.884, log-rank p = 0.012] and recurrence-free survival (RFS) (HR = 0.449, 95% CI: 0.215–0.934, log-rank p = 0.028) compared with those in the WASL low expression group. Univariate and multivariable Cox proportional hazards regression model suggested that WASL expression was an independent prognostic factor for predicting OS and RFS in cervical cancer. DEGs were mostly enriched GO terms related to DNA replication or the proliferation of tumor cells. The results of GSEA suggested samples in the WASL knockdown group were enriched in glycolysis, TNF-α signaling via NFkB, mTORC1 signaling, and Wnt/β-catenin signaling.

Conclusions: WASL expression was associated with the pathological stage, and it might be an independent prognostication factor in patients with cervical cancer. Knockdown of WASL might be correlated with biological processes such as glycolysis, TNFα signaling, mTOR signaling, and Wnt/β-catenin signaling.

Keywords: WASL, cervical cancer, prognosis, microarray, omics

INTRODUCTION

Cervical cancer represents one of the most frequent malignancies in the female reproductive system (Small et al., 2017). Cervical cancer has the highest incidence in developing countries and represents the leading cause of cancer-related deaths in women in those countries (Pedersen et al., 2018). Meanwhile, there are approximately 500,000 new cases and 280,000 deaths of cervical cancer worldwide annually. Treatments of cervical cancer are tailored by stage (Mboumba Bouassa et al., 2017; Pedersen et al., 2018). For patients with early cervical cancer (carcinoma *in situ*, stages 1 and 2 cervical cancer), surgery to remove the tumor, the cervix, and some or all of the womb (conization, hysterectomy with or without bilateral salpingo-oophorectomy, radical trachelectomy) or radiotherapy (internal radiation therapy, radiation therapy with or without chemotherapy) or a combination of both is often recommended. For patients with advanced cervical cancer, radiation therapy as palliative therapy to relieve symptoms with or without chemotherapy, targeted therapy, and surgery to remove pelvic lymph nodes are often recommended (Li et al., 2016; Hu and Ma, 2018). Despite continuous improvements in surgical techniques, radiotherapy equipment, etc., the therapeutic effect on patients with cervical cancer has not improved fundamentally in the past 40 years. The overall 5-year survival rate of cervical cancer patients is about 40%, and it sharply decreases to about 16.5% for patients with advanced or metastatic cervical cancer (Mapanga et al., 2018; Alimena et al., 2019).

With the advancement of technology, many high-throughput sequencing data and phenotypic data based on cervical cancer have been published, and new algorithms have emerged (Liu et al., 2018; Yu et al., 2021), which has provided new possibilities for developing new molecular markers and personalized treatments for cervical cancer.

WASL, also known as WASP, like actin nucleation promoting factor, is a member of the Wiskott–Aldrich syndrome (WAS) protein family. Our previous study suggested that overexpression of WASL served as an oncogene in cervical cancer by promoting the invasion and migration of cervical cancer cells *in vitro* (Hou et al., 2017). However, the effect and associated mechanisms of WASL on the prognosis of patients with cervical cancer remains unclear.

Therefore, in the present study, we identified the prognostication role and potential mechanism of WASL in clinical settings.

MATERIALS AND METHODS

Cervical Cancer Gene Expression Studies

The Cancer Genome Atlas (TCGA) cervical cancer (TCGA-CESC), measured using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization center (Cancer Genome Atlas Research Network [CGARN], Albert Einstein College of Medicine [AECM], Analytical Biological Services [ABS], Barretos Cancer Hospital [BCH], Baylor College of Medicine [BCM], Beckman Research Institute of City of Hope [BRICH], et al., 2017), contained 290 cervical cancer samples. The level 3 data of TCGA-CESC and the associated clinical information were downloaded from University of California, Santa Cruz (UCSC) Xena[1]. The expression values of the cervical cancer samples were in $\log_2 (x + 1)$ transformed RNA-Seq by expectation-maximization (RSEM) normalized count. A total of 290 patients with cervical cancer in TCGA-CESC were included in this study, of which 149 patients were younger than or were 50 years old, and 141 patients were older than 50 years. There were 159 patients with stage 1, 64 patients with stage 2, 41 patients with stage 3, and 20 patients with stage 4 cancer. There are 18 people in Grade 1, 129 in Grade 2, and 117 in Grade 3.

Analyzing the Relationship Between the Expression of WASL and the Clinical Outcomes of Cervical Cancer Patients

Cervical patients in the TCGA-CESC cohorts were divided into WASL low expression group and WASL high expression group based on the optimal cutoff (see below). Then, chi-square test and logistic regression were performed to analyze the relationships between the expression of WASL and the age, pathological stage, grade, Eastern Cooperative Oncology Group (ECOG) score, and number of childbirths of cervical patients. Time-dependent receiver operating characteristic curve (ROC) (Kamarudin et al., 2017) was performed to identify the optimal cutoff point that divided cervical cancer patients into WASL low expression group and WASL high expression group. To clarify the prognostic role of WASL, Kaplan–Meier and univariate and multivariable Cox proportional hazards regression model were conducted on the overall survival (OS) and recurrence-free survival (RFS) of cervical cancer patients.

Lentiviral Vector Transfection

The recombinant lentivirus gene transfer vector targeting WASL (LV-WASL-RNAi) and the lentiviral vector LV-CON049-NC were synthesized by Shanghai Genechem Co. Ltd. The LV-WASL-RNAi and LV-CON049-NC titers were 5×10^8 and 2×10^8 infectious U/ml, respectively. Hela cells were infected with viral supernatant and considered to be knockdown group and normal control group. Fluorescence microscopy was used to detect the fluorescence signal. Each experiment was conducted three times. According to the manufacturer's protocol, qRT-PCR was conducted using iQTM SYBR® Green PCR supermix (1708880, Bio Red) to determine the transfection efficiency. The primers used were as follows: GAPDH: TGACTTCAACAGCGACACCCA (forward), CACCCTGTTGCTGTAGCCAA (reverse); WASL: GAACGAGTCCCTCTTCACTTTC (forward), GTTCCGATC TGCTGCATATAACT (reverse). As shown in

[1]https://xenabrowser.net/datapages/

Supplementary Figure 1, WASL gene knockdown efficiency reached 81.3% in the knockdown group (1 ± 0.033 vs 0.187 ± 0.020, $p < 0.001$)

Immunohistochemistry in Tissue Microarray

A tissue microarray of cervical cancer samples was acquired from the Shanghai Outdo Biotech Company (Shanghai China). The microarray is composed of 126 cervical cancer tissues and 42 paracancerous tissues. First, antigen retrieval was performed in citrate buffer (pH 6.0) and treated with 3% H_2O_2 after deparaffinization and rehydration. Then, the slides were incubated with primary antibody (ab126626; Abcam) and secondary antibody continuously at 4°C overnight until visualization with peroxidase and 3,3′-diaminobenzidine tetrahydrochloride. After that, the expression of WASL in the cervical cancer tissues from the tissue microarray could be blindly quantified. Finally, the average integrated optical density (IOD) per stained area (μm^2) (IOD/area) for positive staining could be calculated by the Image-pro Plus software 6.0.

Gene Expression Profiling

The total RNA of three pairs of cervical cancer Hela cells was extracted and converted to cDNAs, which were used to synthesize double-stranded DNA (dsDNA) templates. For RNA extraction, Trizol method was used to extract total RNA from the samples. The total RNA extracted was subjected to NanoDrop 2000 ($1.7 < A260/A280 < 2.2$) and Agilent Bioanalyzer 2100 (RIN $\geq$ 7.0 and 28S/18S $>$ 0.7) for quality control, and qualified samples entered the chip experiment. Then, the dsDNA was used to synthesize biotin-labeled cRNA using GeneChip 3'IVT Express Kit (Affymetrix). Finally, the biotin-labeled cRNA was hybridized to Affymetrix Human Gene Expression Array (PrimeView) according to the manufacturer's protocol. The probe cell intensity was normalized using robust multi-array average (RMA) methods in the R/bioconductor package "affy" (Gautier et al., 2004). R package "limma" (Ritchie et al., 2015) was used to identify differentially expressed genes (DEGs) between the knockdown group and control group. Absolute fold change >1.3 and $p < 0.05$ were treated as the criteria for the screening of DEGs.

Functional Analysis

To make clear the function of the DEGs, gene ontology (GO) enrichment analysis was performed using the R/bioconductor package "clusterProfiler" (Yu et al., 2012). GO terms meeting p and false discovery rate (FDR) less than 0.05 were considered to be significant. Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), a systems biology approach, could be used to determine whether a defined gene set shows statistically significant, concordant differences between two or more biological groups. Thus, we conducted GSEA (Mootha et al., 2003; Subramanian et al., 2005) on the knockdown group and control group based on the expression profile of the three pairs of cervical cancer cells. Hallmark gene sets v.7.0 was used as a reference gene set. Other parameters were set as default. Gene sets meeting $p < 0.05$ and FDR < 0.25 were considered to be significantly enriched.

RESULTS

Relationships Between WASL Expression and the Clinical Characteristics of Patients With Cervical Cancer

Based on the optimal cutoff point (10.4851), the cervical cancer samples in the TCGA-CESC study were categorized into WASL high expression group (n = 114) and WASL low expression group (n = 176). As shown in **Table 1**, the chi-square test and multivariable logistic regression analysis suggested that lower expression of WASL was associated with a lower pathological stage (chi-square test: p = 0.022, chi-square = 9.613; logistic regression: OR = 0.869, 95% CI: 0.756–0.991, p = 0.041), while the expression of WASL was not correlated with other clinical characteristics (age, grade, ECOG score, and number of childbirths).

Protein Level of WASL Was Significantly Higher in Cervical Cancer Compared With That in Their Adjacent Normal Tissues

To identify the upregulation of WASL in cervical cancer, we detected the protein level of WASL in tissue microarray by immunohistochemistry (IHC) using WASL antibodies. The average IOD per stained area (μm^2) (IOD/area) was used to

TABLE 1 | The associations between the expression of WAS expression and clinical characteristic of patients with cervical cancer in TCGA-CESC.

	WASL expression		Chi-square test		Logistic regression			
	High	Low	*p*-Value	Chi-square	OR	LCI	UCI	*p*-Value
Age (years)								
≤50	63	86	0.345	0.893	1.012	0.984	1.04	0.41
>50	51	90						
Stage								
1	75	84	0.022	9.613	0.869	0.756	0.991	0.041
2	17	47						
3	13	28						
4	7	13						
Grade								
G1	9	9	0.08	5.049	0.621	0.349	1.086	0.098
G2	57	72						
G3	37	80						
ECOG								
0	45	67	0.541	1.228	1.374	0.831	2.294	0.214
1	29	49						
2∼4	6	5						
Number of childbirths								
≤3	72	109	1	0	1.002	0.838	1.191	0.985
>3	29	43						

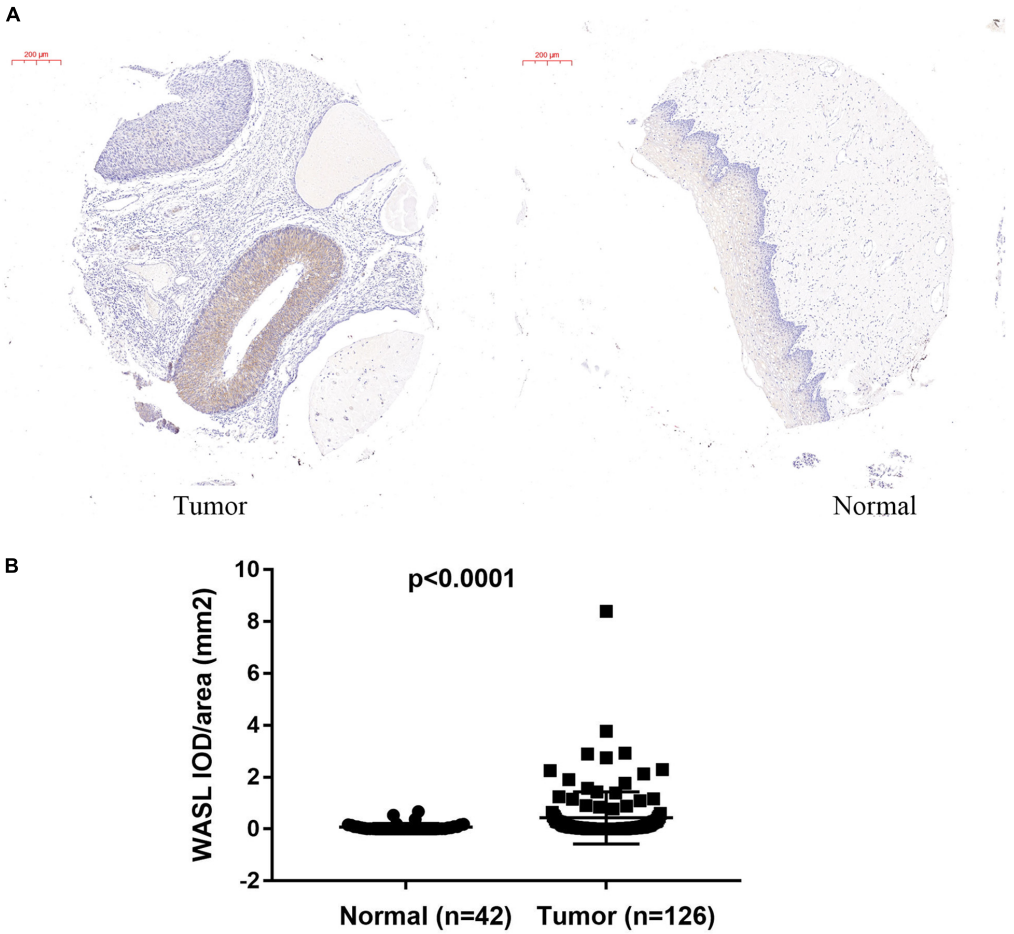


FIGURE 1 | Immunohistochemistry (IHC) staining in tissue microarray showing increased WASL protein level in cervical cancer tissues. The representative images of IHC are shown in **(A)**, while the quantification of WASL protein level is summarized in **(B)**. Scale bar in **(A)**, 200 μm. Student's *t*-test was performed and data represent means ± SEM in **(B)**. $P < 0.0001$.

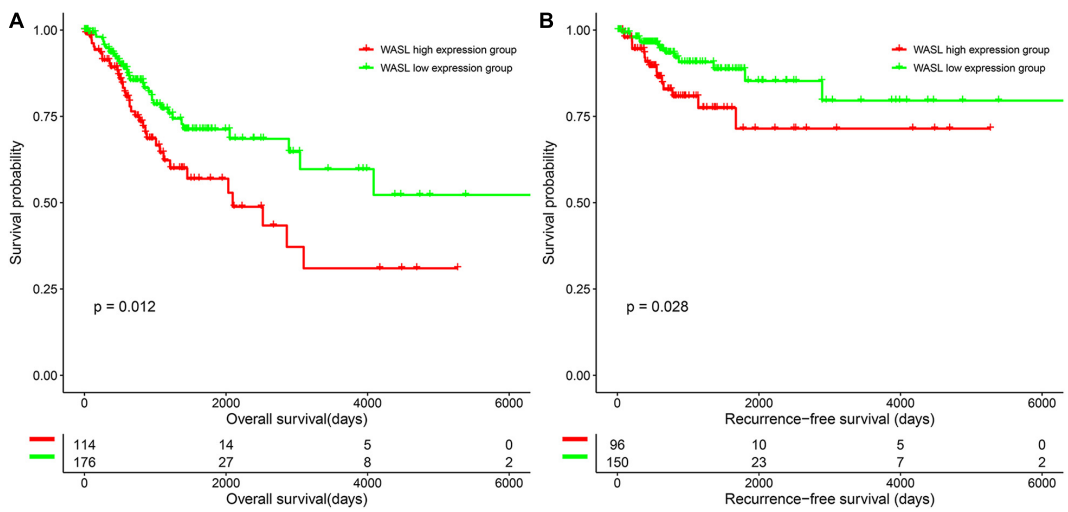


FIGURE 2 | Survival differences of patients in the WASL low expression group and WASL high expression group. **(A)** Overall survival. **(B)** Recurrence-free survival.

calculate a positive signal. As expected, the protein level of WASL was significantly higher in cervical cancer tissues compared with that in their adjacent normal tissues (**Figures 1A,B**).

Lower Expression of WASL Was Associated With Better Survival of Cervical Patients

To characterize the prognostic role of WASL, we performed survival analysis of cervical cancer patients based on the expression of WASL. As shown in **Figure 2**, patients in the WASL high expression group had worse OS (HR: 0.555, 95% CI: 0.348–0.884, log-rank $p = 0.012$, **Figure 2A** and **Table 2**) and RFS (HR = 0.449, 95% CI: 0.215–0.934, log-rank $p = 0.028$, **Figure 2B** and **Table 3**) compared with those in the WASL low expression group. In addition, the univariate and multivariable Cox proportional hazards regression model suggested that WASL expression was an independent prognostic factor for predicting OS (**Table 2**) and RFS (**Table 3**) in cervical cancer.

Differentially Expressed Genes Between WASL Low Expression Group and WASL High Expression Group

As mentioned above, the raw expression signals of the expression profile were normalized using RMA methods. Then, the DEGs between the two groups were calculated via "limma." Consequently, a total of 221 DEGs (including 83 down-regulated genes and 138 up-regulated genes) were identified according to the inclusion criteria (**Figure 3** and **Supplementary Table 1**).

Functional Enrichment Analysis

To make clear the biological meaning of the DEGs, GO enrichment analysis was performed on the DEGs. As shown in **Figure 4**, the DEGs were mostly enriched in "negative regulation of megakaryocyte differentiation," "DNA replication-dependent nucleosome assembly," "DNA replication-dependent nucleosome organization," "CENP-A containing nucleosome assembly," "CENP-A containing chromatin organization," "histone exchange," "flavonoid glucuronidation," "chromatin remodeling at centromere," "chromatin silencing at rDNA," and "protein heterotetramerization." Thus, the results of GO enrichment analysis suggested that DEGs were related with DNA replication or the proliferation of tumor cells. In addition, the results of GSEA suggested samples in the WASL knockdown group were enriched in glycolysis (normalized enrichment score (NES) = 1.429, $p < 0.0001$, FDR = 17.754%, **Figure 5A**), TNF-α signaling via NF-κB (NES = 1.475, $p < 0.0001$, FDR = 19.236%, **Figure 5B**), mTORC1 signaling (NES = 1.489, $p < 0.0001$, FDR = 23.682%, **Figure 5C**), and Wnt/β-catenin signaling (NES = 1.449, $p < 0.0001$, FDR = 18.865%, **Figure 5D**).

DISCUSSION

In the present study, we demonstrated that the expression of WASL was significantly correlated with the pathological stage of cervical cancer patients, and lower expression of WASL was associated with better survival of cervical cancer patients. Actually, WASL was reported to be associated with the biological behaviors of several human cancers. Schwickert et al. suggested that the knockdown of WASL significantly decreased

TABLE 2 | Univariate and multivariable Cox proportional hazards regression on the overall survival in TCGA-CESC.

Characteristic	Univariate analysis				Multivariable analysis			
	HR	LCI	UCI	*p*-Value	HR	LCI	UCI	*p*-Value
WASL	0.555	0.348	0.884	0.013	0.23	0.103	0.514	<0.001
Age	1.017	1	1.035	0.057	1.019	0.986	1.053	0.261
Stage	1.141	1.059	1.23	0.001	1.261	1.098	1.447	0.001
Grade	1.009	0.652	1.56	0.969	1.212	0.639	2.3	0.557
ECOG	1.769	1.165	2.688	0.007	1.722	1.067	2.779	0.026
Number of childbirths	1.042	0.935	1.162	0.458	1.013	0.845	1.214	0.891

TABLE 3 | Univariate and multivariable Cox proportional hazards regression on the recurrence-free survival in TCGA-CESC.

Characteristics	Univariate analysis				Multivariable analysis			
	HR	LCI	UCI	*p*-Value	HR	LCI	UCI	*p*-Value
WASL	0.449	0.215	0.934	0.032	0.337	0.119	0.956	0.041
Age	1.005	0.977	1.033	0.735	1.026	0.98	1.073	0.273
Stage	0.98	0.859	1.118	0.767	0.986	0.795	1.223	0.898
Grade	1.361	0.711	2.606	0.352	2.685	1.033	6.976	0.043
ECOG	2.383	1.461	3.887	0.001	2.116	1.191	3.76	0.011
Number of childbirths	0.941	0.771	1.148	0.547	0.939	0.713	1.237	0.656

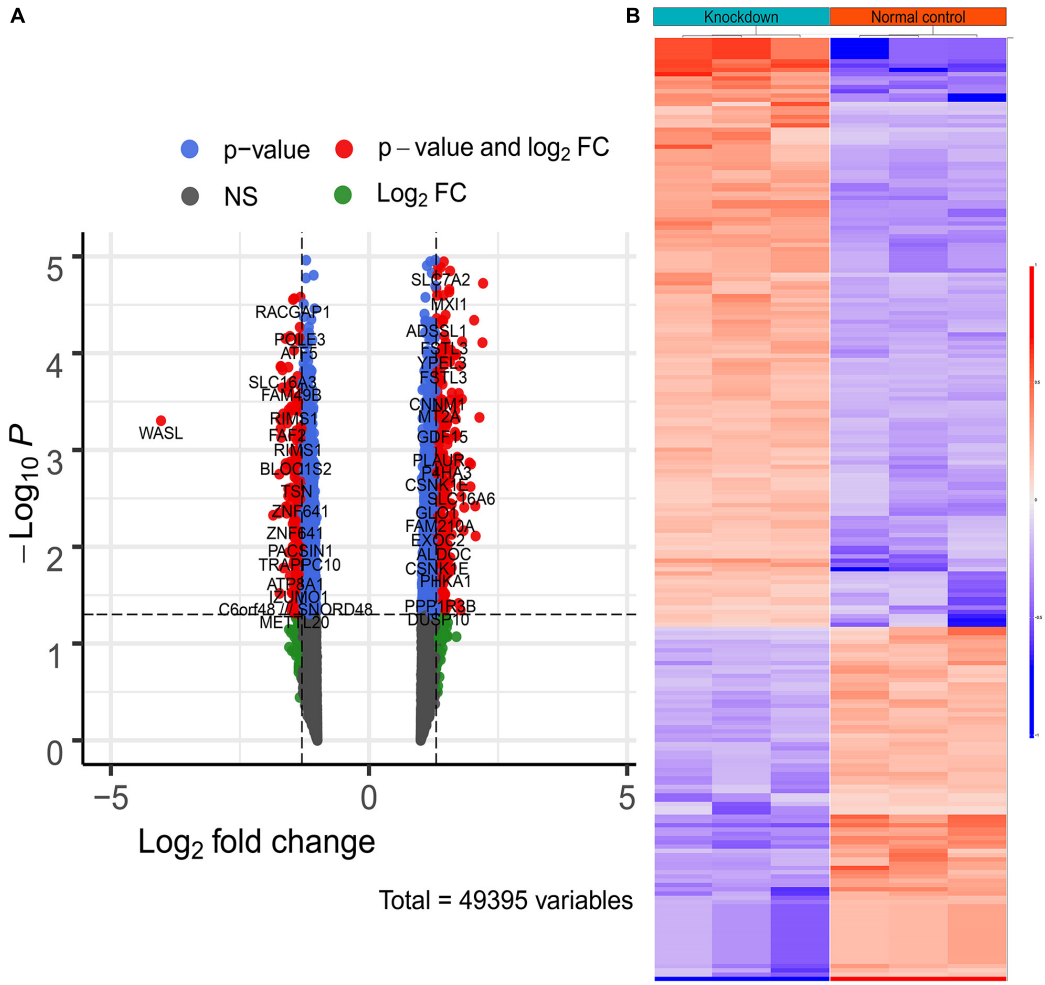


FIGURE 3 | The differentially expressed genes between WASL knockdown group and normal control group. **(A)** Volcano plot depicting the overall distribution of gene expressions between the two groups. **(B)** heatmap depicting the expression levels of differentially expressed genes.

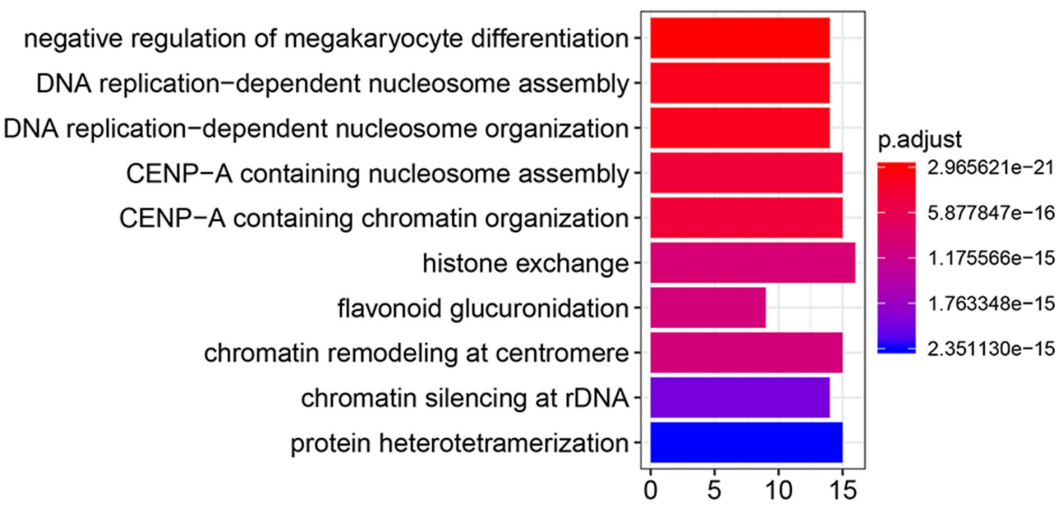


FIGURE 4 | Gene ontology enrichment analysis of differentially expressed genes.

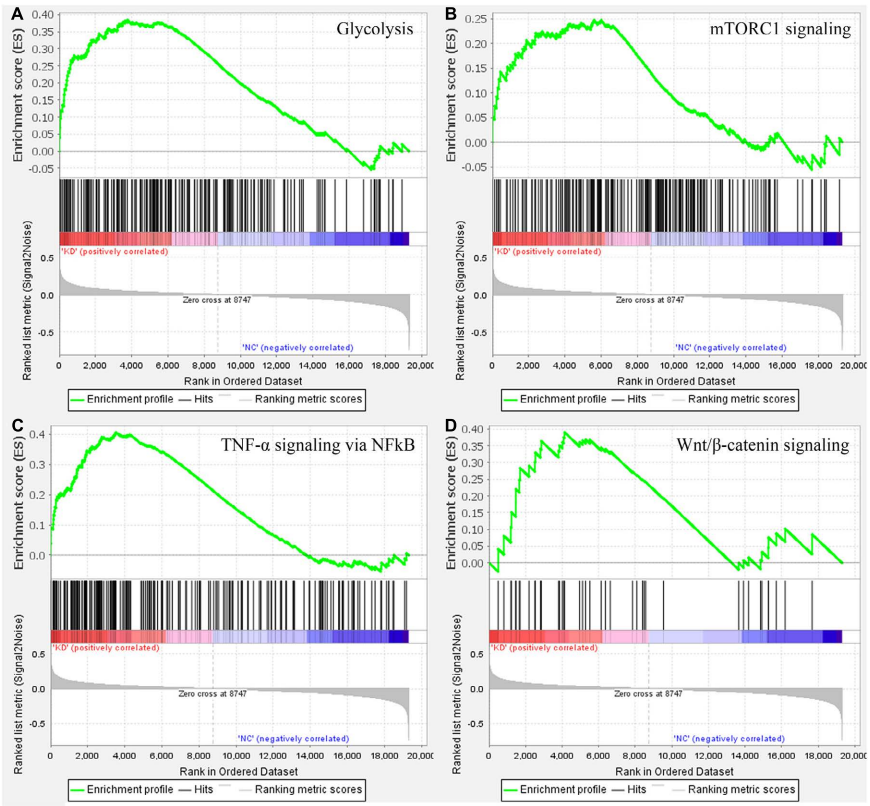


FIGURE 5 | Gene set enrichment analysis on the cervical samples with WASL knockdown. **(A)** glycolysis, **(B)** TNF-a signaling via NF-kB, **(C)** mTORC1 signaling, **(D)** Wnt/b-catenin signaling.

the invasiveness of human breast cancer cells (Schwickert et al., 2015). Peng et al. suggested down-regulation of WASL suppressed the migration of melanoma cells (Peng et al., 2019). Li et al. indicated that MicroRNA-214-5p prohibited the invasion and migration of liver cancer cells through its down-regulation of WASL (Li H. et al., 2018). These studies suggested that WASL might play a role of promoting the progression of cervical cancer cells.

Gene ontology enrichment analysis based on biological process suggested that the DEGs between WASL knockdown group and normal control suggested that the DEGs were mostly enriched in DNA replication or the proliferation of tumor cells, which might explain to some extent why WASL could be used as a prognostic marker for patients with cervical cancer. The results of GSEA suggested that samples in the WASL knockdown group were mostly enriched in glycolysis (genes encoding proteins involved in glycolysis and gluconeogenesis), TNF-α signaling via NF-êB (genes regulated by NF-êB in response to TNF), mTORC1 signaling (genes up-regulated through activation of mTORC1 complex), and Wnt/β-catenin signaling (genes up-regulated by activation of WNT signaling through accumulation of β-catenin). Glycolysis is an important way for organisms (including tumor cells) to obtain energy (Cheng et al., 2013). As a proinflammatory cytokine, TNFα had been reported to be associated with cell proliferation, differentiation, apoptosis, and immune response. Moreover, Li et al. indicated that TNF-α induced the apoptosis of cervical cells (Li J. et al., 2018). Rashmi et al. suggested that AKT inhibitors suppress the proliferation of cervical cancer cells via disruption of mTOR signaling and glucose uptake (Rashmi et al., 2014). Meanwhile, Hsu et al. (2019) reported that activation of Wnt/β-catenin signaling promoted the growth of cervical cancer. Thus, it indicated the knockdown of WASL might be correlated with biological processes such as glycolysis, TNFα signaling, mTOR signaling, and Wnt/β-catenin signaling, which, in return, might suppress the progress of cervical cancer in clinical settings.

In summary, the expression of WASL was associated with the pathological stage, OS, and RFS, and it might be an independent prognostication factor in patients with cervical cancer. DEGs between WSAL knockdown group and normal controls were associated with DNA replication- and cell proliferation-related GO terms. Knockdown of WASL might be correlated with biological process such as glycolysis, TNFα signaling, mTOR signaling, and Wnt/β-catenin signaling.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the Github repository, which could be accessed via https://github.com/xuyulab/cervical-cancer.

AUTHOR CONTRIBUTIONS

YX: conception. JH, CC, and YH: interpretation or analysis of data. JH and CC: preparation of the manuscript. JH, QG, LG, and YH: revision for important intellectual content. YX: supervision. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

This work was supported by the Doctoral Fund of Ministry of Education of China (No. 20130141120059), Zhongnan Hospital of Wuhan University Science, Technology and Innovation Cultivating Fund (znpy2018097), the Science and Technology Department of Hubei Province Key Project (2018ACA159), and Young and Middle-aged Medical Key Talents Training Project of Wuhan (WHQG201901).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.670890/full#supplementary-material

Supplementary Figure 1 | RT-PCR depicting the knockdown efficiency of WASL.

Supplementary Table 1 | Differentially expressed genes between WASL knockdown group and normal control group.

REFERENCES

Alimena, S., Yang, D. D., Melamed, A., Mahal, B. A., Worley, M. J. Jr., Feldman, S., et al. (2019). Racial disparities in brachytherapy administration and survival in women with locally advanced cervical cancer. *Gynecol. Oncol.* 154, 595–601. doi: 10.1016/j.ygyno.2019.06.022

Cancer Genome Atlas Research Network [CGARN], Albert Einstein College of Medicine [AECM], Analytical Biological Services [ABS], Barretos Cancer Hospital [BCH], Baylor College of Medicine [BCM], Beckman Research Institute of City of Hope [BRICH], et al. (2017). Integrated genomic and molecular characterization of cervical cancer. *Nature* 543, 378–384. doi: 10.1038/nature21386

Cheng, Y., Chen, G., Hong, L., Zhou, L., Hu, M., Li, B., et al. (2013). How does hypoxia inducible factor-1alpha participate in enhancing the glycolysis activity in cervical cancer? *Ann. Diagn. Pathol.* 17, 305–311. doi: 10.1016/j.anndiagpath.2012.12.002

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Hou, J., Yang, H., Huang, X., Leng, X., Zhou, F., Xie, C., et al. (2017). N-WASP promotes invasion and migration of cervical cancer cells through regulating p38 MAPKs signaling pathway. *Am. J. Transl. Res.* 9, 403–415.

Hsu, W., Liu, L., Chen, X., Zhang, Y., and Zhu, W. (2019). LncRNA CASC11 promotes the cervical cancer progression by activating Wnt/beta-catenin signaling pathway. *Biol. Res.* 52:33.

Hu, Z., and Ma, D. (2018). The precision prevention and therapy of HPV-related cervical cancer: new concepts and clinical implications. *Cancer Med.* 7, 5217–5236. doi: 10.1002/cam4.1501

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* 17:53. doi: 10.1186/s12874-017-0332-6

Li, H., Wang, H., and Ren, Z. (2018). MicroRNA-214-5p Inhibits the Invasion and Migration of Hepatocellular Carcinoma Cells by Targeting Wiskott-Aldrich Syndrome Like. *Cell Physiol. Biochem.* 46, 757–764. doi: 10.1159/000488734

Li, H., Wu, X., and Cheng, X. (2016). Advances in diagnosis and treatment of metastatic cervical cancer. *J. Gynecol. Oncol.* 27:e43.

Li, J., Zhang, Y., Chen, L., Lu, X., Li, Z., Xue, Y., et al. (2018). Cervical Cancer HeLa Cell Autocrine Apoptosis Induced by Coimmobilized IFN-gamma plus TNF-alpha Biomaterials. *ACS Appl. Mater. Interfaces* 10, 8451–8464. doi: 10.1021/acsami.7b18277

Liu, X. P., Yin, X. H., Meng, X. Y., Yan, X. H., Wang, F., and He, L. (2018). Development and Validation of a 9-Gene Prognostic Signature in Patients With Multiple Myeloma. *Front. Oncol.* 8:615. doi: 10.3389/fonc.2018.00615

Mapanga, W., Chipato, T., and Feresu, S. A. (2018). Treatment of cervical cancer in HIV-seropositive women from developing countries: a protocol for a systematic review. *Syst. Rev.* 7:22.

Mboumba Bouassa, R. S., Prazuck, T., Lethu, T., Jenabian, M. A., Meye, J. F., and Belec, L. (2017). Cervical cancer in sub-Saharan Africa: a preventable noncommunicable disease. *Expert Rev. Anti Infect Ther.* 15, 613–627. doi: 10.1080/14787210.2017.1322902

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273. doi: 10.1038/ng1180

Pedersen, K., Fogelberg, S., Thamsborg, L. H., Clements, M., Nygård, M., Kristiansen, I. S., et al. (2018). An overview of cervical cancer epidemiology and prevention in Scandinavia. *Acta Obstet. Gynecol. Scand.* 97, 795–807. doi: 10.1111/aogs.13313

Peng, D., Dong, J., Zhao, Y., Peng, X., Tang, J., Chen, X., et al. (2019). miR-142-3p suppresses uveal melanoma by targeting CDC25C, TGFbetaR1, GNAQ, WASL, and RAC1. *Cancer Manag. Res.* 11, 4729–4742. doi: 10.2147/cmar.s206461

Rashmi, R., DeSelm, C., Helms, C., Bowcock, A., Rogers, B. E., Rader, J. L., et al. (2014). AKT inhibitors promote cell death in cervical cancer through disruption of mTOR signaling and glucose uptake. *PLoS One* 9:e92948. doi: 10.1371/journal.pone.0092948

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi: 10.1093/nar/gkv007

Schwickert, A., Weghake, E., Brüggemann, K., Engbers, A., Brinkmann, B. F., Kemper, B., et al. (2015). microRNA miR-142-3p Inhibits Breast Cancer Cell Invasiveness by Synchronous Targeting of WASL, Integrin Alpha V, and Additional Cytoskeletal Elements. *PLoS One* 10:e0143993. doi: 10.1371/journal.pone.0143993

Small, W. Jr., Bacon, M. A., Bajaj, A., Chuang, L. T., Fisher, B. J., Harkenrider, M. M., et al. (2017). Cervical cancer: A global health crisis. *Cancer* 123, 2404–2412. doi: 10.1002/cncr.30667

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Yu, D. H., Chen, C., Liu, X. P., Yao, J., Li, S., and Ruan, X. L. (2021). Dysregulation of miR-138-5p/RPS6KA1-AP2M1 Is Associated With Poor Prognosis in AML. *Front. Cell Dev. Biol.* 9:641629. doi: 10.3389/fcell.2021.641629

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Hou, Chen, Hu, Gong, Gan and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 01 July 2021
doi: 10.3389/fcell.2021.689028

Uncovering the Subtype-Specific Molecular Characteristics of Breast Cancer by Multiomics Analysis of Prognosis-Associated Genes, Driver Genes, Signaling Pathways, and Immune Activity

Xinhui Li[1†], Jian Zhou[2†], Mingming Xiao[3†], Lingyu Zhao[4†], Yan Zhao[5†], Shuoshuo Wang[4], Shuangshu Gao[4], Yuan Zhuang[4], Yi Niu[5], Shijun Li[6], Xiaobo Li[4], Yuanyuan Zhu[4], Minghui Zhang[5*] and Jing Tang[4*]*

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] School of Medicine, Southern University of Science and Technology, Shenzhen, China, [3] Department of Pathology, The People's Hospital of Liaoning Province, Shenyang, China, [4] Department of Pathology, Harbin Medical University, Harbin, China, [5] Department of Oncology, Chifeng City Hospital, Chifeng, China, [6] Department of Pathology, Chifeng City Hospital, Chifeng, China

Breast cancer is a heterogeneous malignant disease with different prognoses and has been divided into four molecular subtypes. It is believed that molecular events occurring in breast stem/progenitor cells contribute to the carcinogenesis and development of different breast cancer subtypes. However, these subtype-specific molecular characteristics are largely unknown. In this study, we employed 1217 breast cancer samples from The Cancer Genome Atlas (TCGA) database for a multiomics analysis of the molecular characteristics of different breast cancer subtypes based on PAM50 algorithms. We detected the expression changes of subtype-specific genes and revealed that the expression of particular subtype-specific genes significantly affected prognosis. We also investigated the mutations and copy number variations (CNVs) of breast cancer driver genes and the representative genes of ten signaling pathways in different subtypes and revealed several subtype-specifically altered genes. Moreover, we detected the infiltration of various immune cells in different subtypes of breast cancer and showed that the infiltration levels of major immune cell types are different among these subtypes. Additionally, we investigated the factors affecting the immune infiltration level and the immune cytolytic activity in different breast cancer subtypes, namely, the mutation burden, genome instability and cancer-associated fibroblast (CAF) infiltration. This study may shed light on the molecular events contributing to carcinogenesis and development and provide potential markers and targets for the clinical diagnosis and treatment of different breast cancer subtypes.

Keywords: breast cancer, molecular subtypes, driver gene, immune infiltration, prognosis

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Yanshuo Chu,
University of Texas MD Anderson Cancer Center, United States
Ya Cui,
University of California, Irvine, United States

***Correspondence:**
Yuanyuan Zhu
yuanyuanzhu0214@163.com
Minghui Zhang
cfzhangminghui@163.com
Jing Tang
jinghmu@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *31 March 2021*
Accepted: *28 May 2021*
Published: *01 July 2021*

Citation:
Li X, Zhou J, Xiao M, Zhao L, Zhao Y, Wang S, Gao S, Zhuang Y, Niu Y, Li S, Li X, Zhu Y, Zhang M and Tang J (2021) Uncovering the Subtype-Specific Molecular Characteristics of Breast Cancer by Multiomics Analysis of Prognosis-Associated Genes, Driver Genes, Signaling Pathways, and Immune Activity. Front. Cell Dev. Biol. 9:689028. doi: 10.3389/fcell.2021.689028

INTRODUCTION

Breast cancer is a heterogeneous disease. In 2000, Perou first reported the molecular characteristics-based classification of breast cancer, namely, the luminal subtype (lumA and lumB), basal-like subtype, HER2-overexpression subtype and normal breast-like subtype (Perou et al., 2000). Sorlie et al. (2003) divided the luminal subtype into A type and B/C type. Each subtype has unique molecular signatures, prognoses, clinical behaviors and treatment responses. For example, the prognosis of patients with the lumA and lumB subtypes is relatively good, and the primary treatments are surgery, chemotherapy and endocrine therapy (Harbeck et al., 2019). The 5-year survival rate of patients with the basal-like subtype is low, and there is a lack of effective treatments. Patients with the HER2 subtype are usually treated with targeted drugs or chemoradiotherapy until the tumor has been reduced to a specific size range before undergoing surgical resection (Goldhirsch et al., 2013). Studies have shown that the clinically identified HER2 and basal-like subtypes are complex, explaining why the clinical effects of drugs such as Herceptin on patients with the HER2 subtype are general, and why triple-negative breast cancer is largely difficult to treat (Kumar and Aggarwal, 2016).

At present, numerous studies on breast cancer have focused on tumorigenesis, development, treatment, and improving prognosis (Nagini, 2017; Harbeck et al., 2019; Wang et al., 2019). An increasing number of studies have revealed that the traditional immunohistochemical classification of cancer has some limitations in understanding breast cancer heterogeneity. A more accurate and helpful subtype prediction model of breast cancer is developed through computational biology methods at various molecular levels, which makes up for the lack of immunohistochemical typing. Furthermore, through a deeper mechanism of breast cancer research, reliable and effective treatment regimens can be revealed. However, different molecular subtypes have different effects on treatment and prognosis, and the mechanism remains unclear. Wallden et al. (2015) reported that the PAM50 assay has clinical accuracy and technical precision based on several clinical validation studies. In this project, we used the current academic authoritative PAM50 subtype prediction algorithm to classify breast cancer samples in The Cancer Genome Atlas (TCGA) public database and analyzed the differences among the subtypes at the DNA and RNA levels. We analyzed the difference in driver gene mutations, copy number variation (CNV), and gene fusion in the different subtypes and assessed their prognostic impact. To assess the differences in immune cell levels among the subtypes, we also analyzed the infiltration levels of immune cells and cancer-associated fibroblasts (CAFs). We aimed to determine the potential differences among different subtypes through bioinformatics methods and to find suitable therapeutic targets for clinically conquering breast cancers of different subtypes.

MATERIALS AND METHODS

Data Source

The breast cancer data included 113 normal samples and 1104 tumor samples from TCGA database[1]. We only used samples that had available data from the UCSC Xena database[2] across the following four genomic platforms: RNA expression, gene mutation, CNV, and gene fusion. The representative genes of ten classic pan-cancer signaling pathways were obtained from the PathwayMapper database[3] (Bahceci et al., 2017). The driver genes studied in this analysis were predicted by at least seven algorithms (accounting for half of the total number of algorithms) in the DriverDBv3 online database[4] (Liu et al., 2020).

Subtype and Prognosis Analysis of Breast Cancer

Gene expression profiles were used to classify the molecular subtypes of the breast cancer samples from the TCGA. The PAM50 algorithm developed by Parker et al. was applied for this process (Parker et al., 2009). The prognostic analysis was performed on samples of the five subtypes obtained (lumA, lumB, HER2, basal-like, and normal) using Kaplan–Meier analysis.

Screening of Subtype-Specific RNA and Driver Genes in Each Subtype

The transcriptome data of the breast cancer samples from the TCGA were used to perform subtype-specific RNA analyses among the five molecular subtypes. The method used the Seurat3.0 package, and a p value less than 0.01 was considered to represent a statistically significant difference between a specific subtype and other subtypes.

We then used the 15 recognized driver gene prediction algorithms provided by the DriverDBv3 database to identify the driver genes. We defined driver genes as genes that were determined to be driver genes by more than seven algorithms.

DNA-Level Differences Between Subtypes

The DNA-level changes in the driver genes and ten oncogenic pathway representative genes were analyzed in our study, namely, CNV, gene mutation, and gene fusion. The frequency of DNA changes and the number of samples with DNA changes were analyzed in detail for each subtype.

Analysis of Tumor Mutation Burden

Tumor mutation burden (TMB) was defined as the total amount of somatic gene coding errors, base substitutions, insertions or deletions detected per million bases. TMB data were downloaded from the TCGA database through the GDC tool. We classified the samples of each subtype into low-and high-TMB groups

[1]https://portal.gdc.cancer.gov
[2]https://xenabrowser.net/datapages/
[3]http://www.pathwaymapper.org/
[4]http://ngs.ym.edu.tw/driverdb

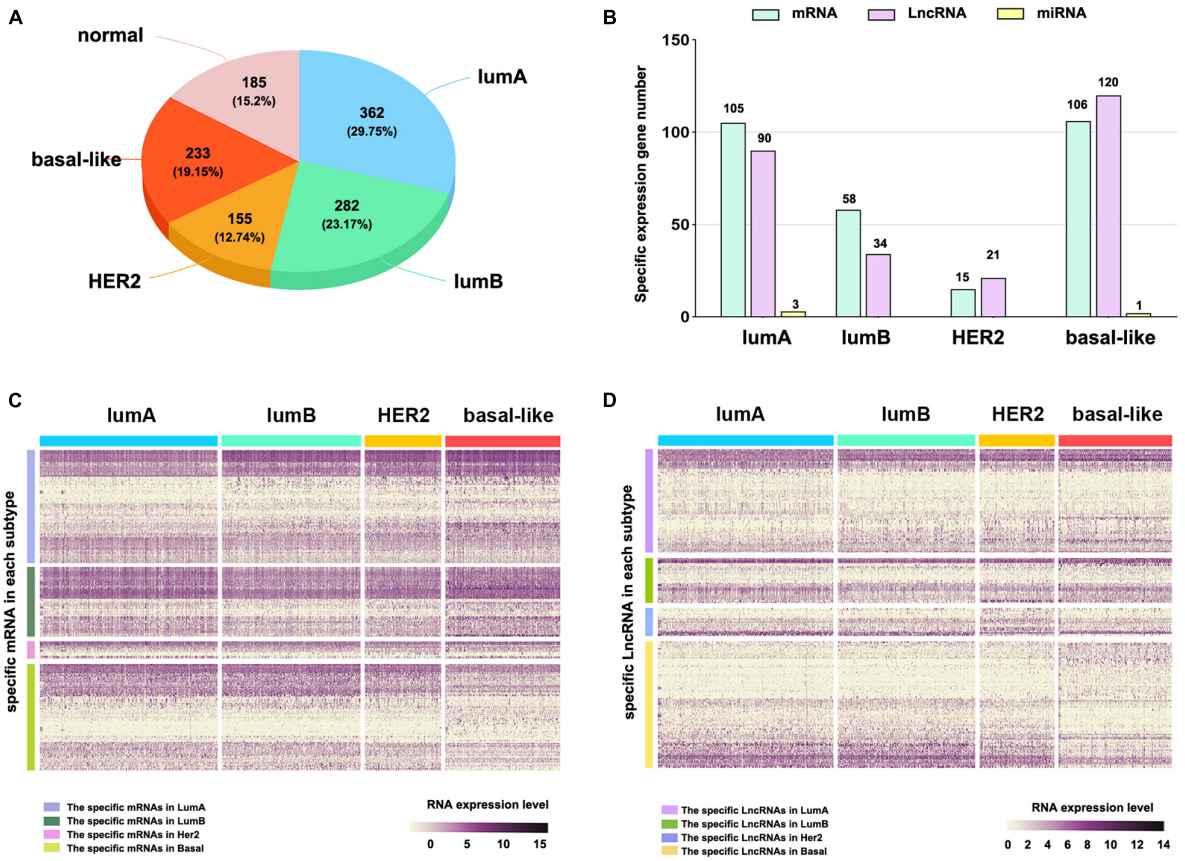


FIGURE 1 | The subtype-specific RNA expression. **(A)** The number of samples of each breast cancer subtype. **(B)** The number of subtype-specific RNAs in each subtype. **(C)** The expression of specific mRNAs in each subtype. **(D)** The expression of specific lncRNAs in each subtype.

according to the median data. Then, we merged the TMB data with corresponding survival information via the ID number of the samples. Kaplan–Meier analysis was conducted to compare the survival difference between the low-and high-TMB groups of each subtype, and the *p* value of the log-rank test was calculated.

Selection and Classification of Genes in Pathways

Genes were assigned to pathways based on a combined revision of pathway analyses from previous papers published between 2008 and 2017, a review of the scientific literature and expert curation. Several genes in the pathways, such as TGF-β, Myc, and PI3K, had been analyzed by specific working groups. These groups were led by experts on each pathway, and each published separate manuscripts (Foroutan et al., 2017; Ge et al., 2018; Knijnenburg et al., 2018; Peng et al., 2018). The pathways included in our study were (1) Hippo signaling, (2) cell cycle, (3) TGFβ signaling, (4) Notch signaling, (5) receptor tyrosine kinase (RTK)/RAS signaling, (6) β-catenin/Wnt signaling, (7) oxidative stress response/Nrf2, (8) Myc signaling, (9) P53, and (10) PI-3-kinase signaling. Gene mutation, copy number amplification and deletion were evaluated in all essential representative genes from the 10 pathways. The frequency of crucial gene changes in the pathways was also counted.

Infiltration of Immune Cells and Fibroblasts

The infiltration levels of immune cells and fibroblasts in breast cancer samples were predicted by Newman et al. (2015) and Becht et al. (2016) algorithms, respectively, using the mRNA expression profiles of the samples. The correlation between immune cells and fibroblasts was calculated by Pearson correlation. All the data processing was performed using R 3.6.1.

Immunophenoscore Analysis in Breast Cancer

An immunophenogram was used to predict anti-PD-1/PD-L1 therapy responses across cancers (Charoentong et al., 2017). The immunophenoscore (IPS) was calculated by the immunophenogram among four cancer subtypes (CTLA4_negative + PD-1_negative, CTLA4_positive + PD-1_negative, CTLA4_ negative + PD-1_ positive, CTLA4_ positive + PD-1_positive) from the TCGA-BRCA database. The IPS ranged from 0 to 10. A high PD-1_positive IPS indicated

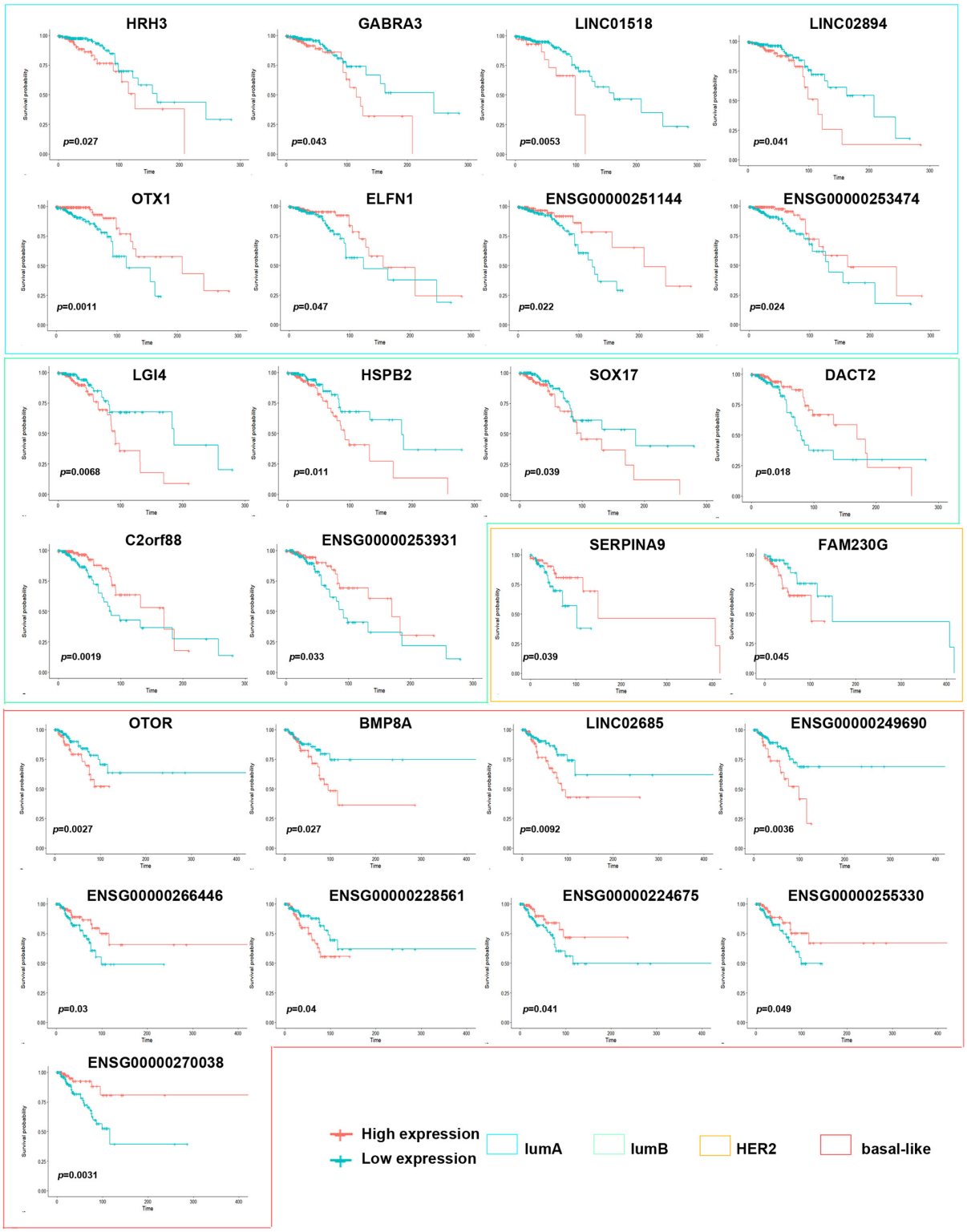


FIGURE 2 | Correlation of expression of subtype-specific RNA with overall survival in breast cancer. Kaplan–Meier survival curves were generated for subtype-specific RNA by comparing groups of high (red line) and low (blue line) gene expression. $p < 0.05$ in log-rank test.

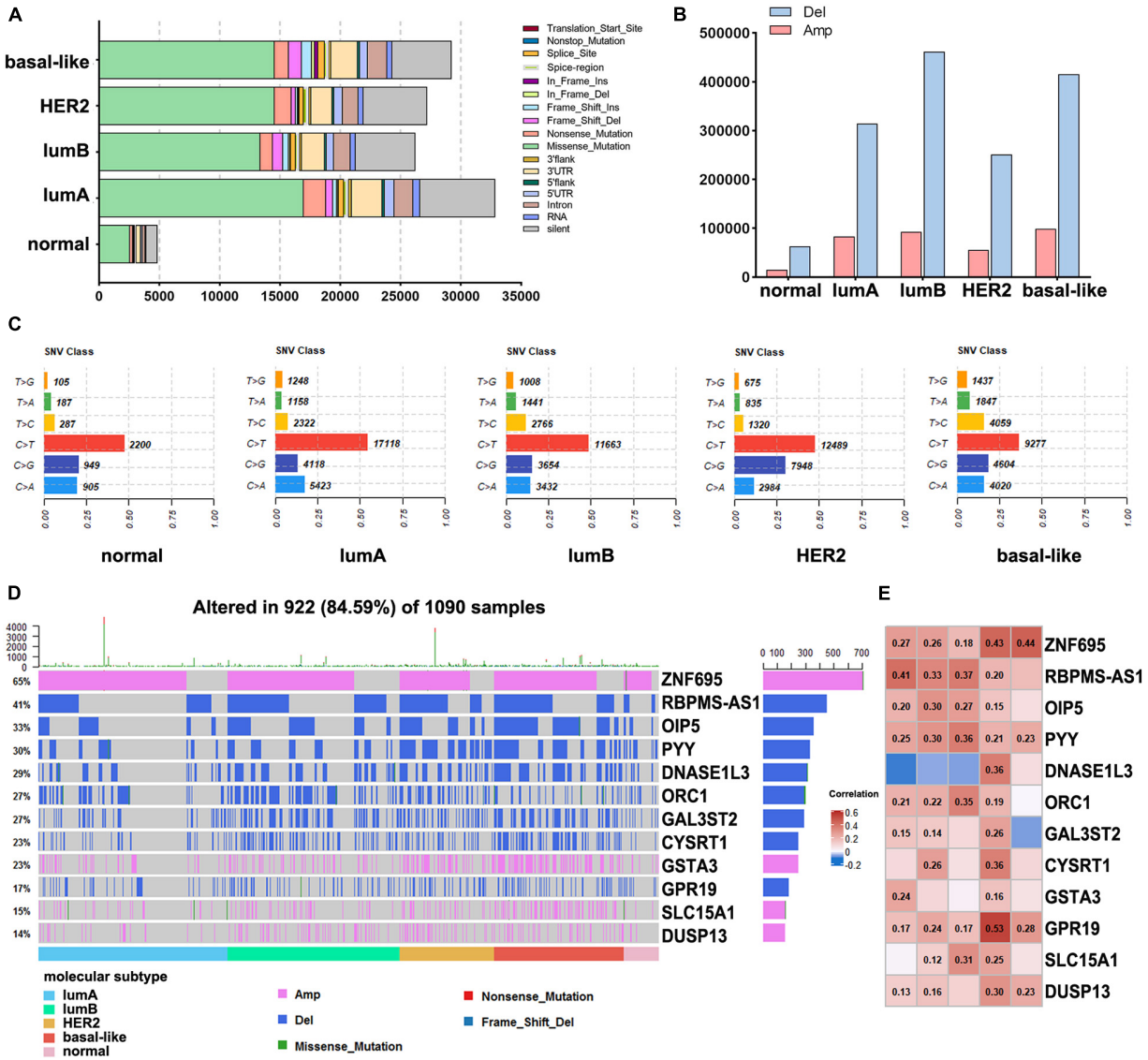


FIGURE 3 | Analysis of DNA alterations in different breast cancer subtypes. **(A)** The different classified categories of DNA mutations in each breast cancer subtype. **(B)** The number of copy number variations in each subtype. **(C)** Statistics of SNV changes in different subtypes. **(D)** Waterfall chart showing the mutations of different subtype-specific genes. **(E)** Correlation analysis of subtype-specific RNAs and their modifications in different subtypes. The numbers in the figure are correlation coefficients, a negative value represents a negative correlation, and a positive value represents a positive correlation.

a well-predicted response to anti-PD-1/PD-L1 therapy. The R code used is available at GitHub[5].

RESULTS

Identification of Molecular Subtype-Specific RNA in Breast Cancer

To explore the differences among breast cancer molecular subtypes, we first used the PAM50 method to predict the molecular subtypes of 1,217 breast cancer samples from the TCGA. The prediction results showed that the number of samples in each breast cancer subtype was lumA: 362, lumB: 282, HER2: 155, basal-like: 233, and normal: 185 (**Figure 1A**). We counted the number of RNAs specifically expressed in each subtype, as shown in **Figure 1B** (**Supplementary Table 1**). We compared differential expression levels of mRNA, lncRNA and miRNA in each subtype (**Figures 1C,D** and **Supplementary Figure 1**). We generated Kaplan–Meier survival curves to explore the potential link between subtype-specific RNA and OS. The median was used as the cutoff for high or low expression chosen for subtype-specific RNA. Among the RNAs, a total of 25 were shown to significantly predict OS (**Figure 2**, $p < 0.05$). These genes were considered to be potential subtype-related prognostic genes.

[5]https://github.com/Mayer/C-imed/Immunophenogram

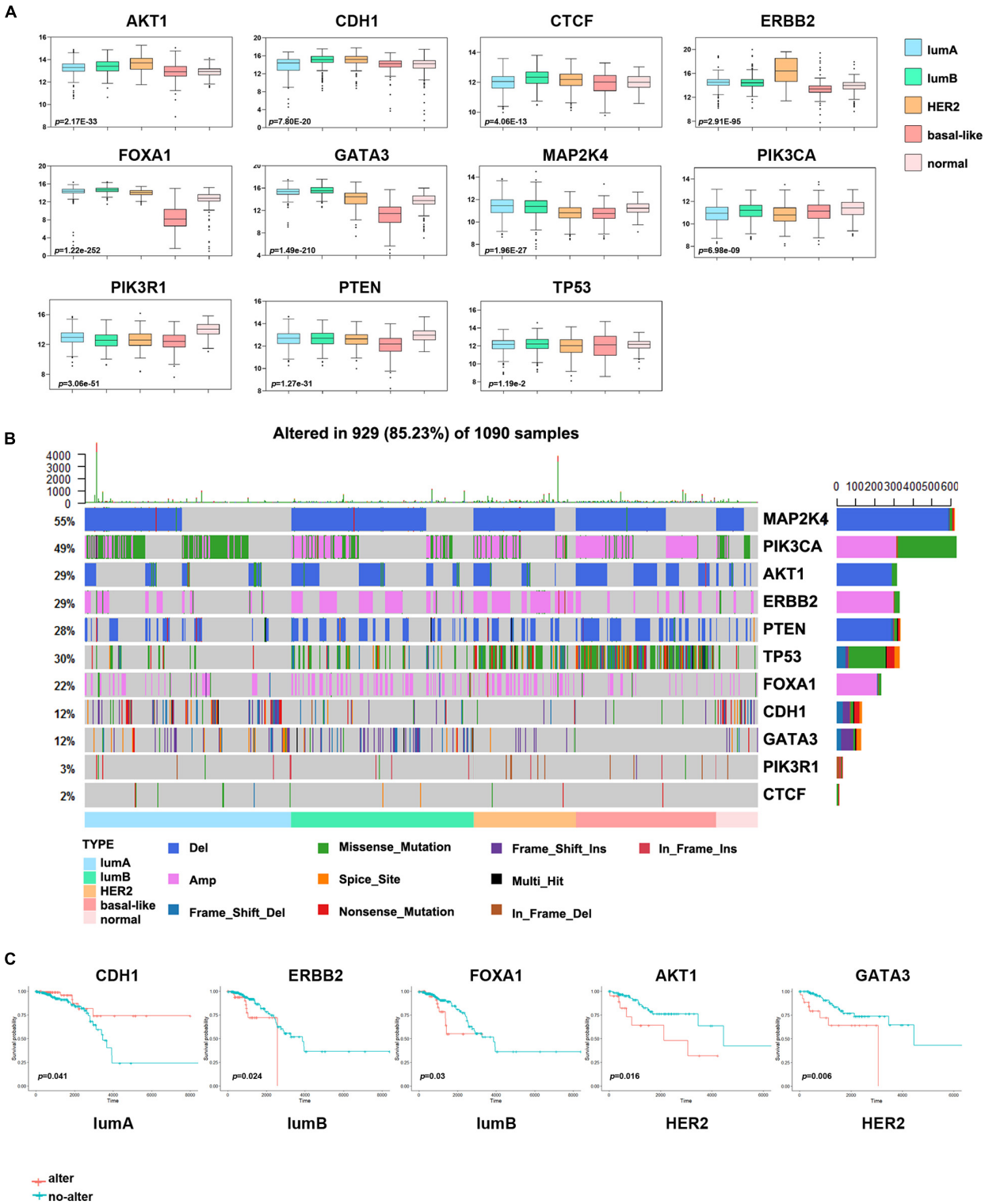


FIGURE 4 | Analysis of the driver genes in different breast cancer subtypes. **(A)** The expression levels of eleven driver genes in each subtype. **(B)** The waterfall chart shows the frequency mutations of the driver genes in each subtype. **(C)** Kaplan–Meier survival curve of the driver genes with gene mutations. Red indicates the gene alteration group, and blue indicates the non-alteration group.

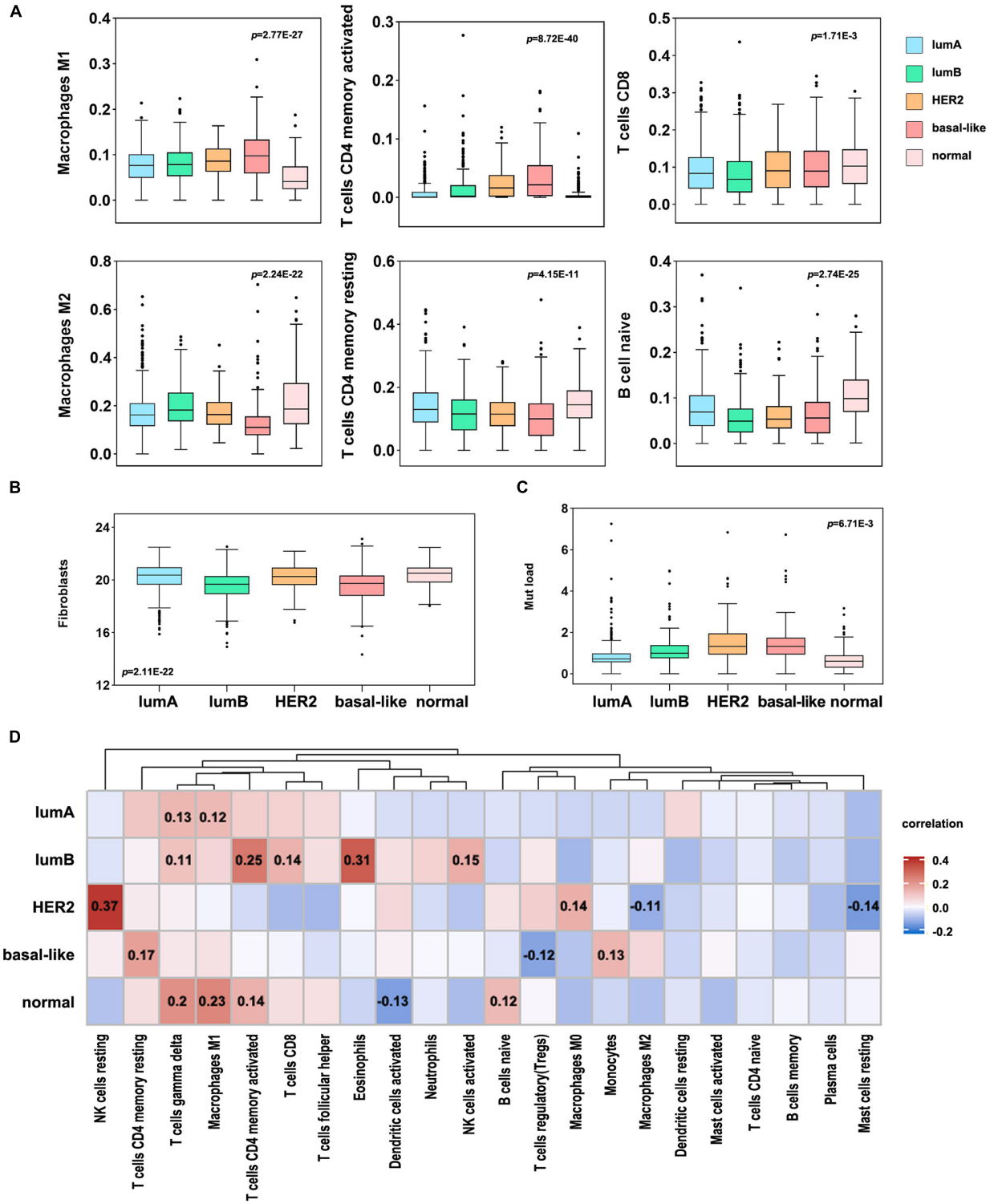


FIGURE 5 | Immune cell infiltration levels in different breast cancer subtypes. **(A)** Differential analysis of the immune cell infiltration level in different breast cancer subtypes ($p < 0.05$). **(B)** The infiltration level of fibroblasts in different breast cancer subtypes. **(C)** Analysis of TMB in different breast cancer subtypes. **(D)** Correlation analysis between TMB and 22 immune cell infiltration levels in different subtypes. Red indicates a positive correlation, and blue indicates a negative correlation. The number represents the degree of correlation, and $p < 0.05$.

Gene Mutations Among the Breast Cancer Subtypes

To assess the alteration of genes among different subtypes in breast cancer, we analyzed the mutations in each subtype, including gene mutation and CNV. In summary, these mutations were classified according to different categories, in which missense mutations accounted for the largest fraction (**Figure 3A**). The lumA subtype had the most missense mutations, and the lumB subtype had the lowest (16,967 and 13,364 mutations, respectively). The CNV analysis across the five subtypes revealed that the highest levels of amplification and deletion were detected in the basal-like and lumB subtypes (**Figure 3B**). Comparison across the five subtypes revealed that they all had increased C > T transversions (**Figure 3C**). The C > G transversions were markedly higher in the HER2 subtype than in other subtypes. The basal-like subtype had more T > C transversions than the other subtypes. We further performed mutation analysis on the subtype-specific genes in which mRNA expression was significantly different in each subtype. The results showed that the mutation of these genes, namely ZNF695, RBPMS1-AS1, OIP5, and PYY, correlated with their RNA expression levels (**Figures 3D,E**).

Identification of Driver Genes in Breast Cancer Subtypes

To further understand the molecular features of different subtypes, we used the DriverDBv3 online database to predict the driver genes of breast cancer samples in the TCGA database. To increase the accuracy of our results, we used more than seven algorithms to simultaneously predict driver genes and obtained 11 driver genes, namely ERBB2, AKT1, PIK3CA, PIK3R1, PTEN, TP53, CDH1, GATA3, MAP2K4, CTCF, and FOXA1. We analyzed the effect of driver gene expression in the different subtypes and found that the expression of PIK3CA, PIK3R1, and PTEN was significantly lower in tumor samples than in normal samples (**Figure 4A**). ERBB2, GATA3, PIK3CA, MAP2K4, and other driver genes are either overexpressed considerably or expressed at low levels in specific subtypes, which indicates that these genes have essential significance in the formation and progression of tumors in these subtypes and may be used as subtype-determining markers and breast cancer treatment targets.

Other studies have also shown that breast cancer cells often have many driver gene mutations (Kaur et al., 2018). To more intuitively show the impact of driver genes on the prognosis of patients with breast cancer, we analyzed the driver gene mutations in each subtype and the abnormal alterations of frequently altered genes and essential cancer genes in different signaling pathways. A large proportion of samples from patients with various subtypes of breast cancer had abnormal changes in the MAP2K4, TP53 and PIK3CA genes (**Figure 4B**). The change frequency of TP53 in basal-like samples was more significant than that in other subtypes, and missense mutation of PIK3CA was more extensive in the lumA subtype. In addition, many patients with the HER2 subtype had ERBB2 gene alterations, and patients with the lumB subtype had a deletion of MAP2K4. And we discovered that some key oncogenes and tumor suppressor genes in 10 classical signaling pathways, namely MDM4, MTOR, MYC, CCND1, and RB1, had a higher proportion of mutations. Consistent with other studies, the average mutation rate of oncogene PIK3CA and tumor suppressor gene TP53 was as high as 10% (**Supplementary Figure 2**).

Next, we evaluated the impact of driver gene alterations on the prognosis of patients with breast cancer. We divided the samples of each subtype into an altered group and a non-altered group according to whether the driver gene had a change in mutation, CNV, or gene fusion. This analysis found that the survival effect of the abnormally altered ERBB2 and FOXA1 groups in the lumB subtype group was significantly lower than that of the non-altered group, and the survival effect of the abnormally altered AKT1 and GATA3 groups in the HER2 subtype group was substantially lower than that of the non-altered group (**Figure 4C**). The survival effect of the abnormally altered CDH1 group in the lumA subtype group was significantly higher than that of the non-altered group.

Differences in Immune Cell Infiltration of Different Subtypes

A high number of mutations in breast cancer samples indicates inferior genome stability, and many mutations in tumor tissues can induce the production of new antigens. Simultaneously, patients with specific gene mutations are suitable candidates for immunotherapy, such as patients with BRCA1/2 gene mutations (Mateo et al., 2019). Consequently, to study the difference in immune cell infiltration among breast cancer subtypes and identify personalized immunotherapy for patients. Here, we used CIBERSORT to analyze the differences in the infiltration of 22 immune cell types in each subtype of tumor tissues. In the more malignant tumor tissues of the basal-like and HER2 subtypes, the infiltration level of M1 macrophages, activated memory CD4 T cells, and CD8 T cells was significantly higher. In contrast, the infiltration levels of M2 macrophages, naive B cells, and resting memory CD4 T cells were substantially lower (**Figure 5A**). Then, we used an MCP counter to analyze the fibroblast infiltration levels in each subtype (**Figure 5B**), among which the infiltration levels of fibroblasts in the lumB and basal-like subtypes were low. Correlation analysis between immune cells and fibroblasts showed that T cells and fibroblast levels were negatively correlated (**Supplementary Figure 3**).

Furthermore, we analyzed the correlation between TMB and immune cell infiltration in each subtype and found that HER2 mutation was significantly correlated with NK cell infiltration, and the other subtypes were significantly correlated with T cell infiltration (**Figures 5C,D**). The basal-like subtype was considerably correlated with CD4 T cell infiltration, and the lumB subtype was significantly correlated with CD4 and CD8 infiltration. This suggests that basal-like and lumB subtype tumors with higher TMB have more T cell infiltration and lower fibroblast infiltration. We speculate that the cellular components involved in positive and negative immune responses are complex, leading to poor immunotherapy effects. The proportions of these immune responses could potentially be changed. The

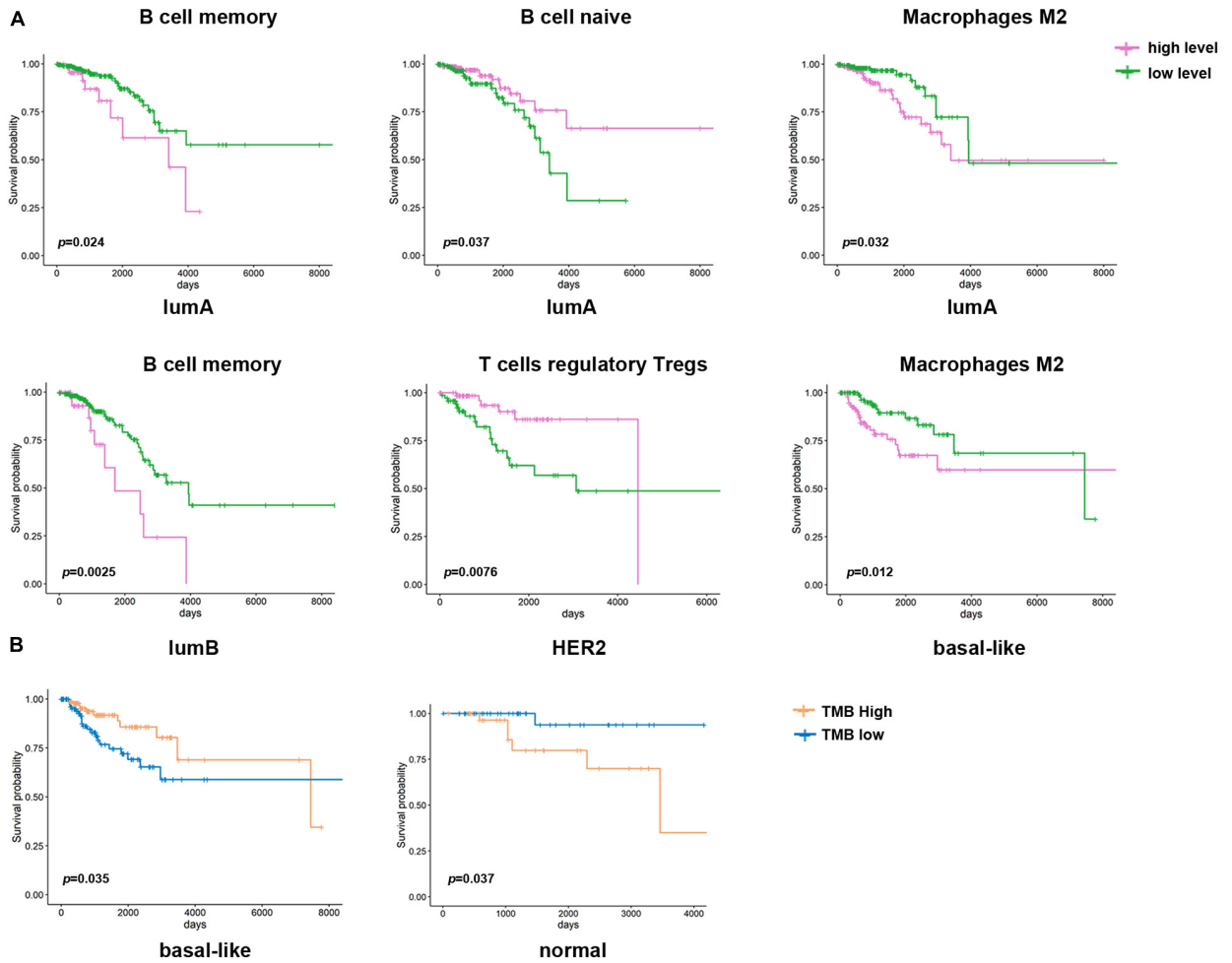


FIGURE 6 | Kaplan–Meier survival curve of immune cells and TMB in different breast cancer subtypes. **(A)** Kaplan–Meier survival curve of the immune cell infiltration level in different breast cancer subtypes ($p < 0.05$). Purple represents a high level of infiltration, and green represents a low level of infiltration. **(B)** Kaplan–Meier survival curve of TMB in different breast cancer subtypes. Blue represents the high TMB group, and orange represents the low TMB group.

components of the positive immune response can work to inhibit negative immune cells, thereby improving the efficacy of immunotherapy for this type of tumor.

Then, we assessed the correlation between immune cell infiltration and prognosis among the different subtypes (**Figure 6A**). It was shown that a low degree of memory B cell infiltration indicated a better prognosis in the lumA and lumB subtype groups, while naive B cell infiltration indicated a poorer prognosis in the lumA subtype group. The group with high Treg cell infiltration had a more extended survival period in the HER2 subtype group; high M2 macrophage infiltration in the basal-like subtype group had a worse prognosis. The above results suggest that the immune cell infiltration in different subtypes is related to patient prognosis.

Different Responses to Immunotherapy Among the Different Molecular Subtypes

To evaluate the different responses to immunotherapy among the subtypes, we analyzed the mRNA expression of breast cancer immune checkpoint genes, namely PD-1, PD-L1, PD-L2, LAG3, VTCN1, IDO1, and TIM3 (**Figure 7A**). The expression of PD-L1 and PD-L2 was significantly higher in the basal-like subtype than in other subtypes (**Figure 7B**). The expression of other immune checkpoint genes (CD40, CD80, CD86, IDO1, and LAG3) was also significantly elevated in the basal-like subtype (**Figure 7A** and **Supplementary Figure 4**). VTCN1 and TIM3 expression was higher in the lumA and HER2 subtypes, respectively ($p < 0.0001$). Then, we used immunophenogram analysis to predict the response to anti-PD-1/PD-L1 therapy among the subtypes. We found that in the CTLA4_negative + PD-1_negative subtype, the lumB and HER2 subtypes exhibited a lower IPS than the other subtypes (**Figure 7C**). In the CTLA4_negative + PD-1_positive and CTLA4_positive + PD-1_positive subtypes, the IPS of the basal-like subtype was significantly higher (**Figure 7C**). We assessed the correlation between immune checkpoint gene expression and prognosis among the different subtypes. The results showed that patients with the basal-like subtype with higher levels of PD-L2 expression had a better OS rate

(**Figure 7D**). These results indicated that patients with the basal-like subtypes were likely to have a higher positive response to anti-PD-1/PD-L1 therapy or a combination of anti-PD-1/PD-L1 and anti-CTLA4 treatments.

DISCUSSION

Breast cancer is a heterogeneous disease with different molecular characteristics and various clinical treatment responses and prognoses. This study found that the level of subtype-specific RNA expression can assess patient prognosis within different subtypes. For example, HRH3 and GABRA3 expression was lower in the lumA subtype than in the other subtypes, and the better prognosis of patients with the lumA subtype with low HRH3 and GABRA3 expression. Similarly, Kiranmai Gumireddy et al. reported that high expression of GABRA3 was inversely correlated with breast cancer survival. GABRA3 can promote breast cancer cell migration, invasion and metastasis by activating the AKT pathway (Gumireddy et al., 2016). However, an A-to-I RNA-edited form of GABRA3 showed that edited GABRA3 suppresses breast cancer cell invasion and metastasis. This is the first report in which an edited RNA was found to play a crucial role in the progression, invasion and metastasis of breast cancer, and it may be a potential therapeutic target. Therefore, it is suggested that subtype-specific RNA expression in breast cancer may be targeted therapeutically, along with providing information regarding the prognosis of patients with different breast cancer subtypes.

Driver genes play an essential role in tumor progression. There is apparent heterogeneity in the frequency and diversity of tumor driver gene mutations. Stephens et al. (2012) reported that few driver genes have high-frequency repeated mutations in breast cancer, except PIK3CA and TP53, which is mutated in approximately 30% of tumors, and ERBB2, FGFR1, and CCND1, which are amplified in approximately 15% of tumors. Significantly, even small numbers of driver gene mutations might correlate with the clinical response of breast cancer. In this study, we found that MAP2K4 (55%), PIK3CA (49%), and TP53 (30%) were mutated in a large number of cancer samples. Among them, the level of TP53 mutation was significantly higher in the basal-like subtype than in the other subtypes, which may be the reason for the higher malignancy of TNBC. In addition, the copy number amplification of ERBB2 in the HER2 subtype was significantly higher than that in the other subtypes, which may promote the development of the HER2 subtype. The driver genes abnormally altered in different subtypes may help explain the differences in response to clinical treatment. Our results were consistent with Shah et al. (2012), who reported that in primary TNBCs, TP53 mutation was the most frequent clonal event (53.8%), followed by PIK3CA mutations (10.7%). The mutation frequencies of TP53 are 12–29 and 72–80% in the lumB, HER2, and basal-like subtypes, respectively. This indicates that TP53 mutation plays a critical carcinogenic effect in most breast cancers (Cancer Genome Atlas Network (CGAN)., 2012). Other studies also reported that the frequency of PIK3CA mutations in ER + and HER2 + tumors is 29–45 and 22.7–39%, respectively (Yang et al., 2016). Although the mutation frequency is significantly different, highly active PI3K pathway expression suggests that PI3K inhibitors may be an effective targeted treatment for breast cancer. Some studies have shown that many genes, such as CDK4, MDM2, and CDH1, have significant CNV. Therefore, further research is needed to understand the driver gene mutations in breast cancer, which may have substantial value for targeted therapy (Luen et al., 2016).

In the past, breast cancer was considered a tumor type with poor immunogenicity. Compared with other cancers with a high mutation load that respond well to immunotherapy (such as non-small cell lung cancer and melanoma), breast cancer has a lower mutation load (Smith et al., 2019). However, recent studies have revealed that some molecular subtypes of breast cancer are infiltrated by immune cells, suggesting that immunotherapy may improve the prognosis of these patients (Dieci et al., 2018). Therefore, it is essential to predict whether patients will respond to immunotherapy. In this study, we found that different infiltration levels of immune cells are related to the prognosis of breast cancer. The high infiltration level of M2 macrophages indicated a worse prognosis in the lumA and basal-like subtypes (**Figure 6A**). The high level of Treg cell infiltration in the HER2 subtype indicates a longer survival time. However, it is generally recognized that Treg cells have a high potential to suppress the immune system. They promote tumor development by inhibiting the effective antitumor immune mechanism of malignant tumors. Therefore, most studies have demonstrated that Treg cell accumulation in breast cancer tumors is related to shorter overall survival times. In addition, we found that the CAF and TMB were also different among the breast cancer subtypes, and both were correlated with immune cell infiltration. In all subtypes of breast cancer, CAFs are negatively correlated with T cell infiltration. In the basal-like subtype, TMB and CD4 + T cell infiltration were higher, CAFs were lower, and high TMB indicated a better prognosis (**Figure 6B**).

Recent years, PD1 and PDL1 are hot spots in the immune regulation. Several studies reported that they governed pathways acting as feedback to prevent excessive T cell response (Yang et al., 2019). In this study, the basal-like subtype samples showed higher expression of immune checkpoint genes (PD-L1, PD-L2, CD40, CD80, CD86, IDO1, and LAG3) than other subtypes. Immunophenogram analysis of the different subtypes also showed that the basal-like subtype responds well to anti-PD-1 therapy but not to anti-CTLA4 therapy. These features suggest that patients with the basal-like subtype may respond better to immunotherapy. Other studies on TMB and immune cell infiltration also showed that higher TMB tends to promote T cell and NK cell infiltration. Patients with bladder cancer with higher TMB levels have a better prognosis (Fan et al., 2020; Mi et al., 2020; Wu et al., 2020). However, in head and neck squamous cell carcinoma and melanoma, high TMB can lead to a shorter survival period, which may be related to infiltrating immune cells (Jiang et al., 2020, 2021). The above results indicate that the cellular components of the tumor involved in the immune response are complex, which leads to different immunotherapy effects in patients. In the future, it may be possible to change

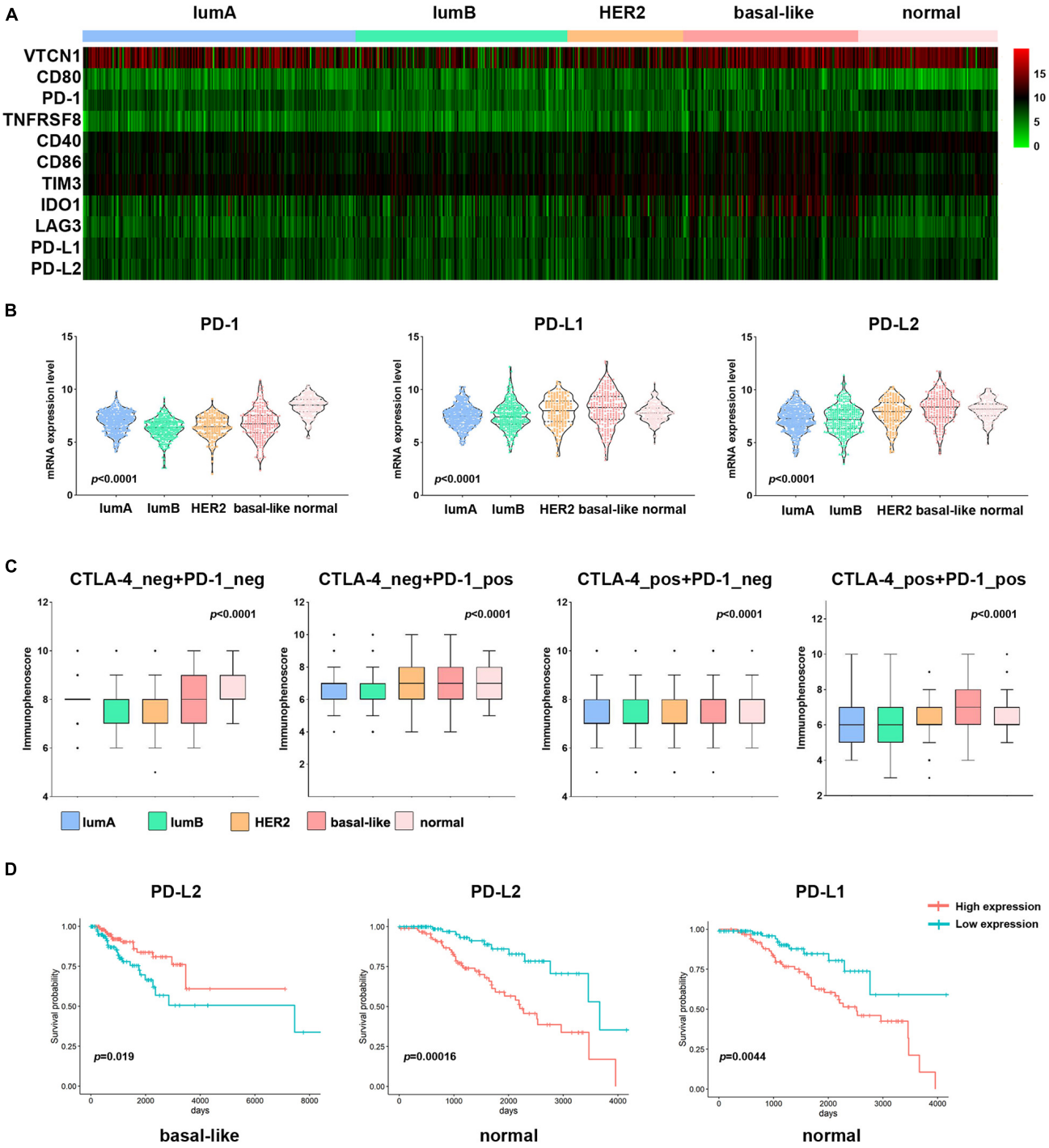


FIGURE 7 | Association of PD-L1 expression and IPS among subtypes in patients with breast cancer. **(A)** Heatmap representation of differences in mRNA expression levels of immune inhibitory checkpoint-related genes. **(B)** Comparison of PD-1, PD-L1, and PD-L2 expression levels between each subtype in breast cancer. **(C)** IPS comparison between each subtype in breast cancer patients in the CTLA4-negative/positive or PD-1-negative/positive groups. CTLA4-positive or PD1-positive represented anti-CTLA4 or anti-PD-1/PD-L1 therapy, respectively. **(D)** Kaplan–Meier survival curves of PD-L1 and PD-L2 in different breast cancer subtypes. Red represents the high expression group, and blue represents the low expression group.

the ratio of immune cells to increase the positive immune cell component and suppress negative immune cells, thereby improving the effect of tumor immunotherapy.

Several recent retrospective and prospective studies have shown that the classification of molecular subtypes and the mechanisms of interaction between tumors and immune cells

of different subtypes are significant for predicting therapeutic response and prognosis and developing individualized treatment plans (Waks and Winer, 2019). Therefore, we analyzed the specific molecular characteristics of different subtypes of breast cancer from a multiomics perspective, providing a theoretical basis for selecting patients most likely to benefit from immunotherapy and providing potential biomarkers for future treatments.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

XbL, YyZ, MZ, and JT designed and supervised this study. XhL, JZ, and JT drafted the manuscript. SW, LZ, and SG illustrated the figures for the manuscript. YZha, YN, and MX searched the database. XhL, SW, and YZhu analyzed the data. All authors approved the final manuscript.

FUNDING

This study was supported by the Natural Science Foundation of Inner Mongolia (No. 2020MS08084 to MZ) and the Inner Mongolia Science & Technology Plan Project (Nos. 201802133 and 2020GG0297 to MZ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.689028/full#supplementary-material

Supplementary Figure 1 | Comparison of specific miRNA expression levels among the breast cancer subtypes.

Supplementary Figure 2 | Mutation frequency of key genes in ten classic signaling pathways. The three black dots on the left side of the gene name represent gene mutation, copy number amplification and deletion. The colors represent oncogenes and tumor suppressor genes. The intensity of the color represents the frequency of mutations.

Supplementary Figure 3 | Correlation analysis of fibroblast infiltration levels and twenty-two immune cell infiltration levels in different subtypes. Red indicates a positive correlation, and blue indicates a negative correlation. The number represents the degree of correlation, and $p < 0.05$.

Supplementary Figure 4 | Comparison of immune checkpoint gene mRNA expression levels among the breast cancer subtypes.

REFERENCES

Bahceci, I., Dogrusoz, U., La, K. C., Babur, O., Gao, J., and Schultz, N. (2017). PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data. *Bioinformatics* 33, 2238–2240. doi: 10.1093/bioinformatics/btx149

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* 17:218. doi: 10.1186/s13059-016-1070-5

Cancer Genome Atlas Network (CGAN). (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* 18, 248–262. doi: 10.1016/j.celrep.2016.12.019

Dieci, M. V., Radosevic-Robin, N., Fineberg, S., van den Eynden, G., Ternes, N., Penault-Llorca, F., et al. (2018). Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: a report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. *Semin. Cancer Biol.* 52, 16–25. doi: 10.1016/j.semcancer.2017.10.003

Fan, S., Gao, X., Qin, Q., Li, H., Yuan, Z., and Zhao, S. (2020). Association between tumor mutation burden and immune infiltration in ovarian cancer. *Int. Immunopharmacol.* 89:107126. doi: 10.1016/j.intimp.2020.107126

Foroutan, M., Cursons, J., Hediyeh-Zadeh, S., Thompson, E. W., and Davis, M. J. (2017). A Transcriptional Program for Detecting TGFbeta-Induced EMT in Cancer. *Mol. Cancer Res.* 15, 619–631. doi: 10.1158/1541-7786.MCR-16-0313

Ge, Z., Leighton, J. S., Wang, Y., Peng, X., Chen, Z., Chen, H., et al. (2018). Integrated Genomic Analysis of the Ubiquitin Pathway across Cancer Types. *Cell Rep.* 23, 213–226.e3. doi: 10.1016/j.celrep.2018.03.047

Goldhirsch, A., Winer, E. P., Coates, A. S., Gelber, R. D., Piccart-Gebhart, M., Thurlimann, B., et al. (2013). Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Ann. Oncol.* 24, 2206–2223. doi: 10.1093/annonc/mdt303

Gumireddy, K., Li, A., Kossenkov, A. V., Sakurai, M., Yan, J., Li, Y., et al. (2016). The mRNA-edited form of GABRA3 suppresses GABRA3-mediated Akt activation and breast cancer metastasis. *Nat. Commun.* 7:10715. doi: 10.1038/ncomms10715

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., et al. (2019). Breast cancer. *Nat. Rev. Dis. Primers* 5:66. doi: 10.1038/s41572-019-0111-2

Jiang, A. M., Ren, M. D., Liu, N., Gao, H., Wang, J. J., Zheng, X. Q., et al. (2021). Tumor Mutation Burden, Immune Cell Infiltration, and Construction of Immune-Related Genes Prognostic Model in Head and Neck Cancer. *Int. J. Med. Sci.* 18, 226–238. doi: 10.7150/ijms.51064

Jiang, F., Wu, C., Wang, M., Wei, K., Zhou, G., and Wang, J. (2020). Multi-omics analysis of tumor mutation burden combined with immune infiltrates in melanoma. *Clin. Chim. Acta* 511, 306–318. doi: 10.1016/j.cca.2020.10.030

Kaur, R. P., Vasudeva, K., Kumar, R., and Munshi, A. (2018). Role of p53 Gene in Breast Cancer: focus on Mutation Spectrum and Therapeutic Strategies. *Curr. Pharm. Des.* 24, 3566–3575. doi: 10.2174/1381612824666180926095709

Knijnenburg, T. A., Wang, L., Zimmermann, M. T., Chambwe, N., Gao, G. F., Cherniack, A. D., et al. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* 23, 239–254.e6. doi: 10.1016/j.celrep.2018.03.076

Kumar, P., and Aggarwal, R. (2016). An overview of triple-negative breast cancer. *Arch. Gynecol. Obstet.* 293, 247–269. doi: 10.1007/s00404-015-3859-y

Liu, S. H., Shen, P. C., Chen, C. Y., Hsu, A. N., Cho, Y. C., Lai, Y. L., et al. (2020). DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res.* 48, D863–D870. doi: 10.1093/nar/gkz964

Luen, S., Virassamy, B., Savas, P., Salgado, R., and Loi, S. (2016). The genomic landscape of breast cancer and its interaction with host immunity. *Breast* 29, 241–250. doi: 10.1016/j.breast.2016.07.015

Mateo, J., Lord, C. J., Serra, V., Tutt, A., Balmana, J., Castroviejo-Bermejo, M., et al. (2019). A decade of clinical development of PARP inhibitors in perspective. *Ann. Oncol.* 30, 1437–1447. doi: 10.1093/annonc/mdz192

Mi, J. L., Xu, M., Liu, C., and Wang, R. S. (2020). Interactions between tumor mutation burden and immune infiltration in ovarian cancer. *Int. J. Clin. Exp. Pathol.* 13, 2513–2523.

Nagini, S. (2017). Breast Cancer: current Molecular Therapeutic Targets and New Players. *Anticancer Agents Med. Chem.* 17, 152–163. doi: 10.2174/1871520616666160502122724

Newman, A. M., Liu, C. L., Green M. R., Gentles, A. J., Feng, W., Xu, Y. et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167. doi: 10.1200/JCO.2008.18.1370

Peng, X., Chen, Z., Farshidfar, F., Xu, X., Lorenzi, P. L., Wang, Y., et al. (2018). Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers. *Cell Rep.* 23, 255–269.e4. doi: 10.1016/j.celrep.2018.03.077

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752. doi: 10.1038/35021093

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 486, 395–399. doi: 10.1038/nature10933

Smith, K. N., Llosa, N. J., Cottrell, T. R., Siegel, N., Fan, H., Suri, P., et al. (2019). Persistent mutant oncogene specific T cells in two patients benefitting from anti-PD-1. *J. Immunother. Cancer* 7:40. doi: 10.1186/s40425-018-0492-x

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A.* 100, 8418–8423. doi: 10.1073/pnas.0932692100

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature* 486, 400–404. doi: 10.1038/nature11017

Waks, A. G., and Winer, E. P. (2019). Breast Cancer Treatment: a Review. *JAMA* 321, 288–300. doi: 10.1001/jama.2018.19323

Wallden, B., Storhoff, J., Nielsen, T., Dowidar, N., Schaper, C., Ferree, S., et al. (2015). Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* 8:54. doi: 10.1186/s12920-015-0129-6

Wang, Q., Yang, M., Zhang, Y., Zhong, L., and Zheng, X. (2019). Novel Combination Oncolytic Adenoviral Gene Therapy Armed with Dm-dNK and CD40L for Breast Cancer. *Curr. Gene Ther.* 19, 54–65. doi: 10.2174/1566523219666190307094713

Wu, Z., Wang, M., Liu, Q., Liu, Y., Zhu, K., Chen, L., et al. (2020). Identification of gene expression profiles and immune cell infiltration signatures between low and high tumor mutation burden groups in bladder cancer. *Int. J. Med. Sci.* 17, 89–96. doi: 10.7150/ijms.39056

Yang, S. X., Polley, E., and Lipkowitz, S. (2016). New insights on PI3K/AKT pathway alterations and clinical outcomes in breast cancer. *Cancer Treat. Rev.* 45, 87–96. doi: 10.1016/j.ctrv.2016.03.004

Yang, T., Li, J., Li, R., Yang, C., Zhang, W., Qiu, Y., et al. (2019). Correlation between MDSC and Immune Tolerance in Transplantation: cytokines, Pathways and Cell-cell Interaction. *Curr. Gene Ther.* 19, 81–92. doi: 10.2174/1566523219666190618093707

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Li, Zhou, Xiao, Zhao, Zhao, Wang, Gao, Zhuang, Niu, Li, Li, Zhu, Zhang and Tang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 05 July 2021
doi: 10.3389/fcell.2021.676142

Comprehensive Analysis of Ferroptosis Regulators With Regard to PD-L1 and Immune Infiltration in Clear Cell Renal Cell Carcinoma

Song Wang†, Shiming Chen†, Yufan Ying†, Xueyou Ma, Haixiang Shen, Jiangfeng Li, Xiao Wang, Yiwei Lin, Ben Liu, Xiangyi Zheng and Liping Xie**

Department of Urology, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China

OPEN ACCESS

Edited by:
Yan Zhang,
Harbin Institute of Technology, China

Reviewed by:
Xueli Zhang,
Shanghai Cancer Institute, China
Jun Zhao,
Yale University, United States

***Correspondence:**
Xiangyi Zheng
zheng_xy@zju.edu.cn
Liping Xie
xielp@zju.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *04 March 2021*
Accepted: *03 May 2021*
Published: *05 July 2021*

Citation:
Wang S, Chen S, Ying Y, Ma X, Shen H, Li J, Wang X, Lin Y, Liu B, Zheng X and Xie L (2021) Comprehensive Analysis of Ferroptosis Regulators With Regard to PD-L1 and Immune Infiltration in Clear Cell Renal Cell Carcinoma. Front. Cell Dev. Biol. 9:676142. doi: 10.3389/fcell.2021.676142

Clear cell renal cell carcinoma (ccRCC) is one of the tumor types with sensitivity to ferroptosis, and immunotherapy has emerged as a standard pillar for metastatic ccRCC treatment, while it remains largely obscure whether ferroptosis influences the tumor immune microenvironment in ccRCC. Based on available data in The Cancer Genome Atlas, divergent expression profiles of ferroptosis regulators were noted in ccRCC and normal tissues, and we also found that the ferroptosis regulators correlated with the PD-L1 expression. Two independent subtypes were determined by consensus clustering analysis according to the expression level of ferroptosis regulators in ccRCC. Cluster 1 showed lower histological tumor stage and grade, more favorable prognosis, and higher PD-L1 expression compared to cluster 2. CIBERSORT analysis revealed that cluster 2 harbored higher infiltrated levels of CD8+ T cell, Tregs, and T follicular helper cell, while cluster 1 more correlated with the monocyte, M1 macrophage, and M2 macrophage. Gene set enrichment analysis indicated that the ERBB signaling and JAK_STAT signaling pathways were significantly enriched in cluster 1. We subsequently identified CARS as the potentially key immune infiltration-related ferroptosis regulator, whose high expression showed dismal prognosis and was positively correlated with PD-L1 expression in ccRCC. We also verified the upregulation of CARS in ccRCC tissues and cell lines *via* qRT-PCR method. Additionally, a pan-cancer analysis demonstrated that CARS closely related to the expression of immune checkpoint-related genes (especially PD-L1) and an unfavorable prognosis in diverse cancer types. In summary, our study suggested the crucial role of ferroptosis in immune infiltration of ccRCC, and CARS is a potentially novel prognostic biomarker and potential target for cancer immunotherapy.

Keywords: ferroptosis, clear cell renal cell carcinoma, PD-L1, immune infiltration, cars, therapeutic target

INTRODUCTION

Kidney cancer is a prevalent human malignancy and among the top 10 causes of cancer-related deaths globally, with estimated 430,000 new cases and 180,000 deaths annually (Sung et al., 2021). Clear cell renal cell carcinoma (ccRCC, also named KIRC) is the major subtype, accounting for nearly 70% of kidney cancer cases (Capitanio et al., 2019). Although the improved understanding

and advances in cancer treatment are acknowledged, the clinical outcomes of patients with ccRCC remain unsatisfactory, specifically those diagnosed with advanced ccRCC. In recent years, immunotherapy has demonstrated unprecedented antitumor activity in patients with advanced malignancies—emerging as a standard therapeutic modality for recurrent or metastasized ccRCC (Barata and Rini, 2017). Majority of patients with ccRCC are still unable to acquire durable clinical benefits from mainstream immunotherapies, such as the combined therapies of anti-PD-1 and/or anti-CTLA4 antibodies, despite having historically high response rates to these immunotherapies (Motzer et al., 2015; Choueiri et al., 2018). An in-depth understanding of the heterogeneity of ccRCC tumor immune microenvironment (TIME) might help in providing personalized immunotherapy management to improve patient survival.

Ferroptosis is an iron-dependent programmed cell death process, characterized by the production of free radicals and excessive accumulation of lipid peroxides (Dixon et al., 2012). Notably, ferroptosis has been implicated in ischemic organ injuries (Tuo et al., 2017), neurodegeneration (Chen et al., 2015), autoimmune diseases (Hu et al., 2019), and cancer development (Viswanathan et al., 2017). Specifically, significant evidence suggests that ferroptosis plays a vital role in cancer inhibition, and its activation enhances the therapeutic efficacy of anticancer agents (Eling et al., 2015; Yu et al., 2017; Zhang et al., 2021). Nonetheless, most studies have focused on elucidating the intrinsic carcinogenic mechanisms of tumors; whether ferroptosis influences the TIME remains largely elusive. Emerging evidence indicates that tumor-killing T cell and anti-PD-L1 antibodies induce ferroptosis in tumor cells, while liproxstatin-1 (a type of highly efficient ferroptosis depressor) inhibits the anticancer efficacy of these agents. Additionally, a synergistic effect has been reported between anti-PD-L1 antibodies and ferroptosis activators (such as erastin), suppressing tumor growth both *in vitro* and *in vivo* (Wang et al., 2019). However, there is still a lack of integrated understanding of ferroptosis in ccRCC, including the crosstalk between ferroptosis regulators and tumor immune microenvironment.

Here, a systematical analysis was performed, involving the expression profiles and the correlations with prognosis, PD-L1, and roles in TIME of ferroptosis regulators in ccRCC. Besides this, we clustered subtypes based on the expression level of ferroptosis regulators. Consequently, apparent tumor heterogeneity, distinct PD-L1 expression, and tumor immune microenvironment were noted between the two subtypes, promoting risk stratification and precision therapy for patients with ccRCC. Subsequently, CARS was identified as a potential immune infiltration-related ferroptosis regulator, whose high expression demonstrated dismal prognosis and positively correlated with PD-L1 expression in ccRCC. Furthermore, this work also comprehensively explored the pan-cancer expression profiles, prognostic values, and correlations with immune checkpoints of CARS. Our findings provide novel insights into the regulatory mechanisms associated with TIME and the approaches for ccRCC immunotherapy.

MATERIALS AND METHODS

Data Acquisition

RNA-seq data and clinical information of the ccRCC cohort were downloaded from the Genomic Data Commons (GDC) data portal of The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/) in January 2020. Data included 72 normal kidney and 530 ccRCC tissues. In the part of pan-cancer analysis of CARS, RNA-seq data and corresponding clinical information of 33 cancer types cohort were obtained from the GDC portal of TCGA. Meanwhile, normal tissue samples were obtained from GTEx V8 release version (https://gtexportal.org/home/datasets). GSE53757 and GSE15641 datasets from Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) database were downloaded and used to further validate the expression level of CARS. The abbreviations of multiple cancer types, ferroptosis regulators, and immune-related checkpoints in this study are listed in **Supplementary Table 1**.

Bioinformatics Analysis

The R software "Consensus Cluster Plus" (v1.54.0) package was adopted for consistency analysis (the maximum number of clusters is six, and 80% of the total sample drawn 100 times) according to the expression level of selected ferroptosis regulators in ccRCC patients; the "ggplot2" package was used for principal component analysis (PCA). The "Cluster Profiler" package was used to perform the gene set enrichment analysis (GSEA), Gene Ontology (GO) analysis, and Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway analysis in R software. The permutation tests were performed 1,000 times, and *p*-values were adjusted for multiple testing by performing the Benjamini–Hochberg procedure, Notably, $p < 0.05$ or FDR < 0.05 was a significantly enriched pathway.

To obtain a reliable evaluation of immune infiltration in ccRCC, CIBERSORT was used to detect the relative proportion of infiltrating immune cells in each ccRCC cancer sample. The algorithm of 1,000 permutations was adopted to obtain meaningful results, and the results were implemented by R, version 4.0.3, and software packages "ggplot2" and "pheatmap." The protein expression level of CARS in ccRCC compared to normal tissue was obtained from the Human Protein Atlas database (https://www.proteinatlas.org/). The CARS survival was externally validated by using the Kaplan–Meier plotter (Nagy et al., 2018) (http://kmplot.com/analysis/). Univariate and multivariate Cox regression analyses were performed to identify the prognostic value of CARS in ccRCC. Forest was used to show the *P* value, hazard ratio (HR), and 95% CI of each variable *via* the "forestplot" R package.

Human Clinical Samples, Cell Lines, and qRT-PCR Analysis

In total, 22 pairs of ccRCC and adjacent normal tissues were obtained from patients subjected to radical nephrectomy between January and October 2013 at the First Affiliated Hospital of Zhejiang University. Written informed consent was acquired from each patient, and the Institutional Ethics Committee in the hospital approved this study. The human ccRCC cell lines 786-O,

ACHN, and caki-1 and the normal kidney cell line HK-2 were purchased from Shanghai Institute of Cell Biology (Shanghai, China). Total RNA was extracted with TRIzol reagent (Takara) and then reverse-transcribed into cDNA by using PrimeScript RT Reagent Kit (Takara). The CARS relative expression was detected by RT-qPCR, through using the SYBR Premix Ex Taq (Takara) and ABI 7500 fast real-time PCR System (Applied Biosystems). GAPDH was used as an endogenous normalization reference. All patients' clinical information and primers used are shown in **Supplementary Tables 2, 3**.

Statistical Analysis

Statistical tests were conducted using R, version 4.0.3, and GraphPad Prism 8.0. Student's *t*-test was used to analyze CARS expression in 22 pairs of ccRCC and adjacent normal tissues; one-way ANOVA was utilized to compare the expression of CARS in ccRCC cell lines to that in normal kidney cell line. Wilcox test and Kruskal–Wallis test were used to perform the group comparisons of two groups and more than two groups, respectively. Spearman correlation analysis was performed to evaluate the correlation between ferroptosis regulators' expression levels with checkpoint -elated genes. For the Kaplan–Meier curves, P-values, and HR with 95% confidence interval (CI) were generated by log-rank tests and univariate Cox proportional hazards regression. Univariate and multivariate analyses were conducted using Cox regression models to establish the independent prognostic value of the CARS integrated other clinical features. The predictive efficiency of genes for 1-, 3-, and 5-year overall survival (OS) was estimated using the receiver operating characteristic (ROC) curves by timeROC package. $P < 0.05$ indicated statistical significance.

RESULTS

Expression Divergence of Ferroptosis Regulators Between ccRCC and Adjacent Normal Tissues

A total of 24 ferroptosis regulators derived from a previous study were defined as key ferroptosis regulator genes because of their pivotal roles in regulating ferroptosis (Liu et al., 2020). To evaluate the biological functions of ferroptosis regulators in ccRCC, the expression patterns of selected ferroptosis regulators between ccRCC and adjacent normal pairs were systematically explored based upon TCGA dataset. We downloaded and analyzed the expression profiles of 530 cases of ccRCC individuals and 72 normal patients. Consequently, distinct expression levels of ferroptosis regulators were found in ccRCC and normal tissues (**Figures 1A,B**), including upregulated ferroptosis regulators CDKN1A, HSPA5, EMC2, SLC7A11, MT1G, HSPB1, FANCD2, SLC1A5, RPL8, LPCAT3, DPP4, and CARS ($p < 0.05$) and downregulated regulators NFE2L2, GPX4, CISD1, FDFT1, NCOA4, GLS2, CS, ATP5MC3, and ACSL4 ($p < 0.05$). However, no significant differences were noted between the ccRCC and normal tissues regarding the expression of SAT1, TFRC, and ALOX15 ($p > 0.05$). Furthermore, correlation combined with prognosis analysis demonstrated that the expression of most ferroptosis regulatory genes was positively correlated and played a vital prognostic role in ccRCC (**Figure 1C**). These findings suggest that ferroptosis regulators might play a crucial role in regulating tumorigenesis and ccRCC development.

Consensus Clustering Analysis of Ferroptosis Regulators Revealed Significant Differences in Baseline Characteristics and Survival Between Two Patient Clusters

The Consensus Cluster Plus package of R software was used for consistency analysis. According to the expression level of selected ferroptosis regulators and the proportion of ambiguous clustering measure, $k = 2$ was identified as optimal clustering stability from $k = 2$ to 6 (**Figures 2A,B**; **Supplementary Figures 1A–E**). Therefore, the 530 KIRC patients were divided into two subtypes, namely, cluster 1 ($n = 444$) and cluster 2 ($n = 86$). In contrast to cluster 2, most of the ferroptosis regulators (20/24) were highly expressed in cluster 1, and two regulators (HSPB1 and GPX4) were lowly expressed in cluster 1, while there was no difference in the expression of two regulators (MT1G and SLC1A5) between two clusters (**Figure 2C**). Subsequently, differences were noted in the clinicopathological characters and prognosis between the two subtypes, as shown in **Table 1**. Cluster 1 was preferentially correlated with lower tumor stage as well as cancer grade ($P < 0.01$), whereas no statistical differences in age, sex, and race were found between the two clusters ($P > 0.05$). Additionally, cluster 1 patients possessed favorable OS ($P < 0.001$) and progression-free survival ($P < 0.001$) than cluster 2 (**Figures 2D,E**). The results indicated that there is significant heterogeneity between the two subtypes of KIRC patients. To further confirm the clustering results defined by the expression of ferroptosis regulators, we subsequently applied the PCA method to analyze the gene expression profiles between the two subtypes (**Supplementary Figure 1F**). As a result, the gene expression profiles between the two subtypes revealed perfect differences.

Association of Ferroptosis Regulators With PD-L1 Expression Level and Immune Cell Infiltration in ccRCC

To investigate the relationship between PD-L1 and ferroptosis in ccRCC, this work evaluated the divergent expression in two clusters and the association of PD-L1 with the ferroptosis regulators in ccRCC. Unlike normal adjacent tissues and cluster 2 patients, the expression levels of PD-L1 in ccRCC tissues and cluster 1 patients were distinctly overexpressed, respectively ($P < 0.01$; **Figures 3A,B**). An analysis involving 530 ccRCC individuals revealed that PD-L1 showed a positive association with the expression levels of most ferroptosis regulators (such as ACSL4, CARS, NCOA4, etc.) but a negative association between HSPB1, MT1G, RPL8, GPX4, and PD-L1 (**Figure 3C**). We further explore the impact of ferroptosis regulators on the tumor immune microenvironment of ccRCC. The two subtypes classified according to the

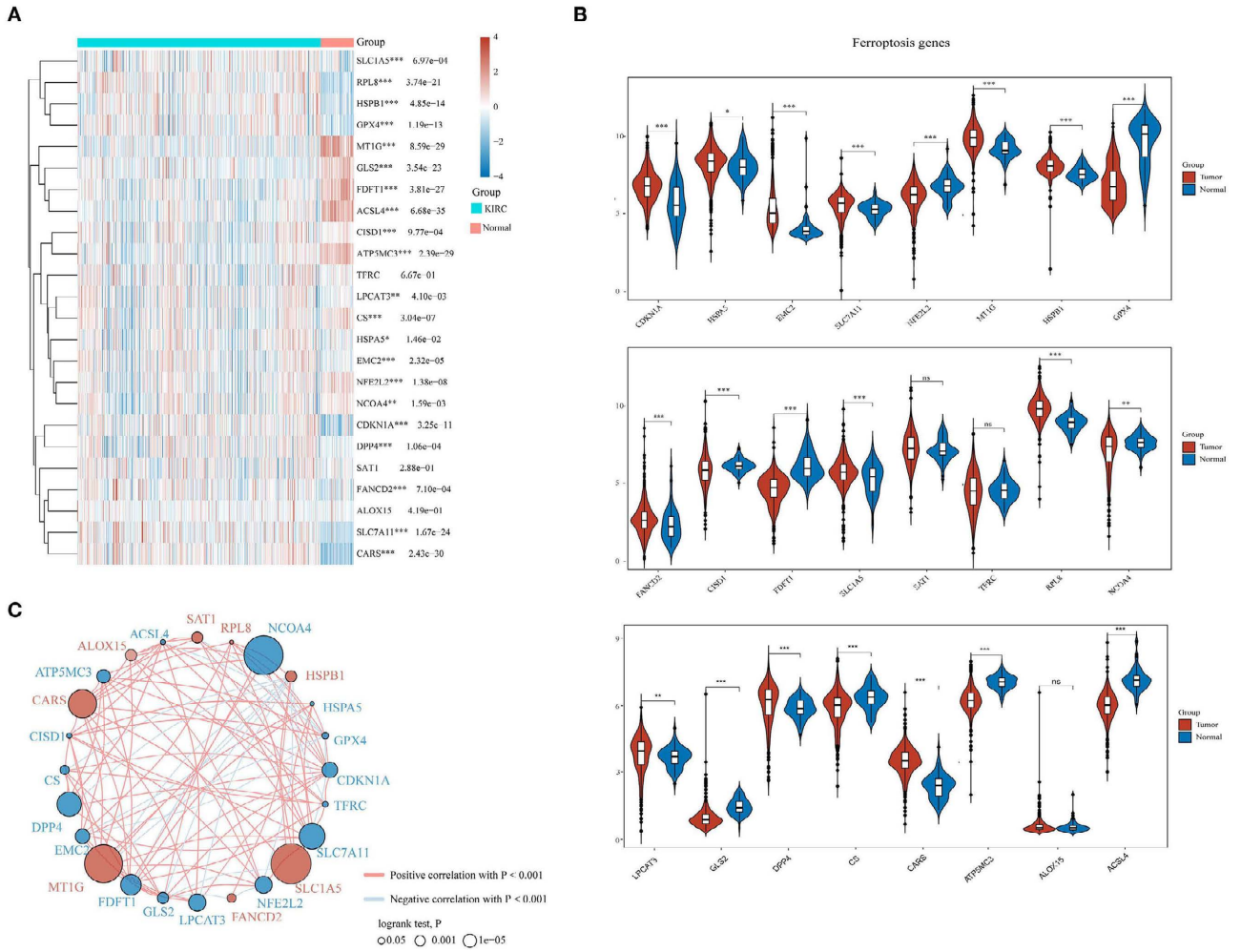


FIGURE 1 | The expression distribution, correlation, and prognostic values of ferroptosis regulators in clear cell renal cell carcinoma (ccRCC) patients. Heat map **(A)** and violin plots **(B)** of ferroptosis regulators in ccRCC compared to normal tissues. **(C)** Spearman correlation and prognostic values of ferroptosis regulators in ccRCC. The red and blue line respectively represents the positive and negative correlation. The red and blue dot represents bad and good prognosis, respectively. The larger the circle, the smaller the prognosis log rank *p*. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

expression levels of selected ferroptosis regulators revealed significant differences in immune cell infiltration (**Figure 3D**). Cluster 2 possessed higher infiltrated levels of activated NK cell, memory B cell, CD8+ T cell, Tregs, and T follicular helper cell (**Figures 3D**, **4A–C**), while cluster 1 was more correlated with the CD4+ memory resting T cell, naive B cell, neutrophil, monocyte, M1 macrophage, and M2 macrophage (**Figures 3D**, **4D–F**). The GSEA method was adopted to clarify the underlying regulatory mechanisms causing the TIME difference between the two clusters. Consequently, ERBB signaling (normalized enrichment score, NES = 1.379, p and $q < 0.01$) and JAK_STAT signaling (NES = 1.245, p and $q < 0.05$) were significantly associated with cluster 1 (**Figure 4G**). Therefore, the ERBB and JAK_STAT signaling pathways are potentially implicated in the divergent TIME of two subtypes in ccRCC.

Key Ferroptosis Regulator CARS Is Upregulated in ccRCC

As described previously, the two ccRCC subtypes defined according to the expression levels of ferroptosis regulators have apparent tumor heterogeneity and distinct PD-L1 expression and tumor immune microenvironment. These results suggested that ferroptosis regulators play an important role in tumor development and tumor immune infiltration. An intersection of ferroptosis regulators highly expressed in ccRCC, linked to poor prognosis, and positively correlated with the expression of PD-L1 was used to determine key regulators linked to poor prognosis and PD-L1 expression among the 24 selected ferroptotic regulators. As shown in **Figure 5A**, CARS was the potentially key poor prognostic ferroptosis regulator that is highly expressed in ccRCC and positively correlated with PD-L1 expression. Additionally, CARS was the only ferroptosis

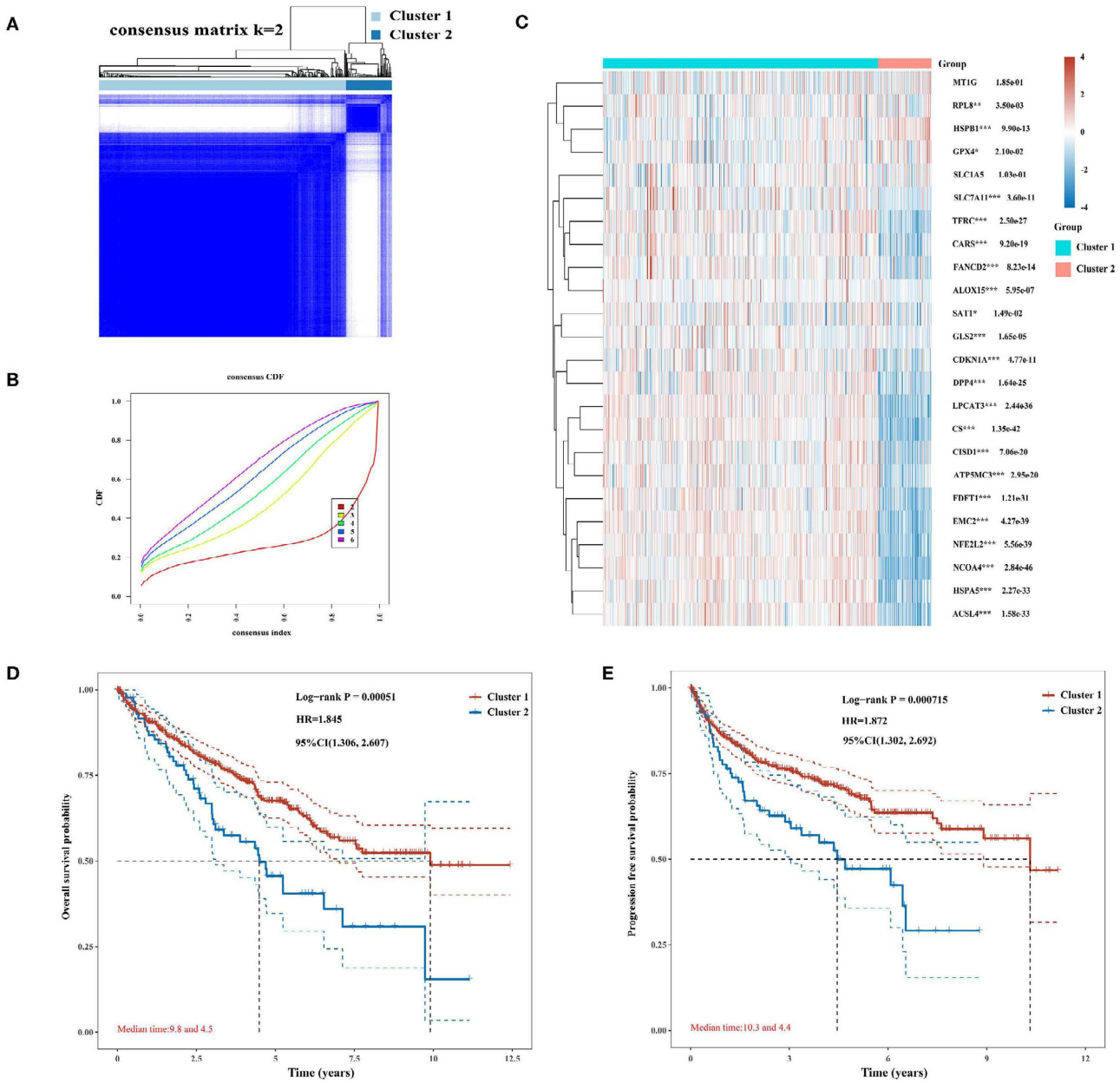


FIGURE 2 | Differential expression pattern of ferroptosis regulators and survival in two clear cell renal cell carcinoma (ccRCC) subtypes. **(A)** Consensus clustering matrix for $k = 2$. **(B)** Cumulative distribution function curves for $k = 2$–6. **(C)** Heat map visualized the expression patterns of ferroptosis regulators in two ccRCC subtypes or clusters. **(D,E)** The Kaplan–Meier curves show the overall survival **(D)** and the disease-free survival **(E)** for two clusters of ccRCC patients. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

regulator highly expressed in ccRCC and cluster 1, positively corelated with PD-L1 expression and worse OS (**Figure 5B**). To further illustrate the expression of CARS in ccRCC, we analyzed the expression data from the TCGA and GEO database. The results showed that, in the TCGA cohort and GSE53757 and GSE15641 cohorts, CARS was significantly more highly expressed in ccRCC than in normal tissues ($P < 0.05$, **Figures 5C–E**). Moreover, the expression level of CARS was significantly higher in advanced cancer stage compared to early-cancer-stage patients ($P < 0.05$, **Figure 5C**). Subsequently, the relative expression of CARS was verified in 22 pairs of ccRCC and normal tissues in ccRCC cell lines and normal kidney cell line *via* qRT-PCR method ($P < 0.05$, **Figures 5F,G**). Correspondingly, the protein expression of CARS was also upregulated in ccRCC tissue compared to normal tissue (**Figure 5H**). These results demonstrated that CARS is highly expressed in ccRCC at both transcriptional and translational levels.

TABLE 1 | Clinical characteristics of two clusters of ccRCC patients.

	Feature	Cluster 1	Cluster 2	*P*-value
Age	Mean (SD)	60.4 (12.1)	61.4 (12.6)	
	Median (min, max)	61 (26, 90)	60.5 (33, 85)	0.483
Gender	Female	159	27	
	Male	285	59	0.508
Race	Black or Asian	52	12	
	White	385	74	0.763
pTNM_stage	I	236	29	
	II	48	9	
	III	91	32	
	IV	66	16	0.002
Grade	G1	13	1	
	G2	201	26	
	G3	167	39	
	G4	55	20	0.009

High CARS Expression Was a Significantly Independent Prognostic Biomarker in ccRCC

Furthermore, the prognostic value of highly expressed CARS was analyzed in ccRCC, where the upregulated CARS group show worse OS compared to the lowly expressed group ($P < 0.05$, **Figure 6A**). A similar result was also confirmed in an independent ccRCC cohort from the Kaplan–Meier plotter ($P < 0.05$; **Figure 6B**). A stacked bar chart and a violin plot (shown in **Figure 6C**) demonstrated that high CARS ccRCC patients possessed a lower proportion of live/death and survival time compared to the low CARS ccRCC patients ($P < 0.05$). Besides this, CARS also showed a powerful prognostic ability since the ROC curve showed that the AUC values of CARS expression for predicting 1-, 3-, and 5-year survival were 0.651, 0.579, and 0.562, while the AUC values of β-actin and GAPDH expression for predicting 1-, 3-, and 5-year survival were 0.598, 0.544, and 0.513 and 0.580, 0.565, 0.551, respectively (**Figure 6D**). The Cox analysis method was also adopted to explore the relationship between CARS expression and OS in ccRCC; univariate analysis revealed that age (HR = 1.0289, p = 1e-05), TNM stage (HR = 1.866, $p < 0.0001$), tumor grade (HR = 2.2907, $p < 0.0001$), and CARS expression (HR = 1.9503, $p = 0.00095$) are significantly correlated with OS (**Figure 6E**). Moreover, a multivariate analysis, exhibited as a forest plot in **Figure 6F**, revealed that CARS expression ($p = 0.047$) is an independent factor for prognosis in ccRCC patients.

Correlation Analysis of CARS Expression With PD-L1 and Infiltrating Immune Cells

To further investigate the underlying functions of CARS in ccRCC, we utilized the Cluster Profiler package to analyze GO term and KEGG pathway in samples with high expression levels of CARS (as shown in **Figure 7A**). The biological processes of GO annotation suggested five categories that were positively linked to highly expressed CARS: cell growth, positive regulation of cytokine production, T cell activation, positive regulation of proteolysis, and cell cycle G1/S phase transition. KEGG pathway analysis revealed that five pathways positively associated with the expression of CARS: PI3K-AKT signaling pathway, MAPK signaling pathway, proteoglycans in cancer, chemokine signaling pathway, and cell cycle (**Figure 7B**). These results indicated that the CARS may be involved in multiple signaling pathways and tumor immune microenvironment in ccRCC. A further correlation analysis showed that CARS was significantly positively correlated with the expression of PD-L1 in ccRCC ($P < 0.001$, Spearman = 0.32) (**Figure 7C**). To identify a discrepancy in the TIME of ccRCC patients with high and low expression levels of CARS, the CIBERSORT algorithm was carried out to explore the difference of numerous immune cell subtypes between the CARS high expression group and the low expression group in ccRCC. As shown in **Figure 7D**, monocytes, M1 macrophages, and neutrophil increased in the high CARS expression group, whereas CD8+ T cell, activated NK cell, and gamma delta T cell decreased ($P < 0.05$), and the heat map visualized the percentage of abundance of tumor-infiltrating immune cells in each sample of low and high CARS group (**Figure 7E**).

Comprehensive Analysis of Key Ferroptosis Regulator CARS in Pan-Cancer

Considering that ferroptosis and immune infiltration are widely implicated in the occurrence and development of multiple cancer types, we desired to determine whether CARS, the key ferroptosis regulator of our study, also predicts poor prognosis and affects the tumor immune microenvironment in other cancer types. Therefore, we firstly analyzed the relative expression of CARS in more than 10,000 tumor and normal tissue samples in the TCGA and GTEx database. The results demonstrated that, compared to normal tissues, CARS was significantly upregulated in 12 (CHOL, COAD, DLBC, GBM, HNSC, KICH, KIRC, KIRP, LGG, LIHC, PAAD, and PCPG) cancer types and downregulated in 13 (ACC, BRCA, ESCA, LUAD, LUSC, OV, PRAD, SKCM, STAD, TGCT, THCA, UCEC, and UCS) cancer types ($P < 0.05$; **Figure 8A**). Meanwhile, the prognostic values of CARS were also investigated across multiple cancer types. The univariate Cox regression analyses in pan-cancer suggested that the upregulation of CARS was correlated with poor OS in 13 (ACC, BRCA, COAD, KICH, KIRC, KIRP, LAML, LGG, LIHC, MESO, PAAD, PCPG, and PRAD) cancer types ($P < 0.05$, **Figure 8B**). In addition, a further expression correlation analysis between CARS and eight most common immune checkpoint-related genes (including PD-L1, CTLA4, HAVCR2, LAG3, PDCD1, PDCD1LG2, SIGLEC15, and TIGIT) was performed. The results revealed that CARS closely related to the expression of checkpoint-related genes in most cancer types, especially the expression of PD-L1 (**Figure 8C**). In summary, a high expression of CARS was correlated with immune checkpoint-related gene expression and dismal prognosis in diverse prevailing cancers.

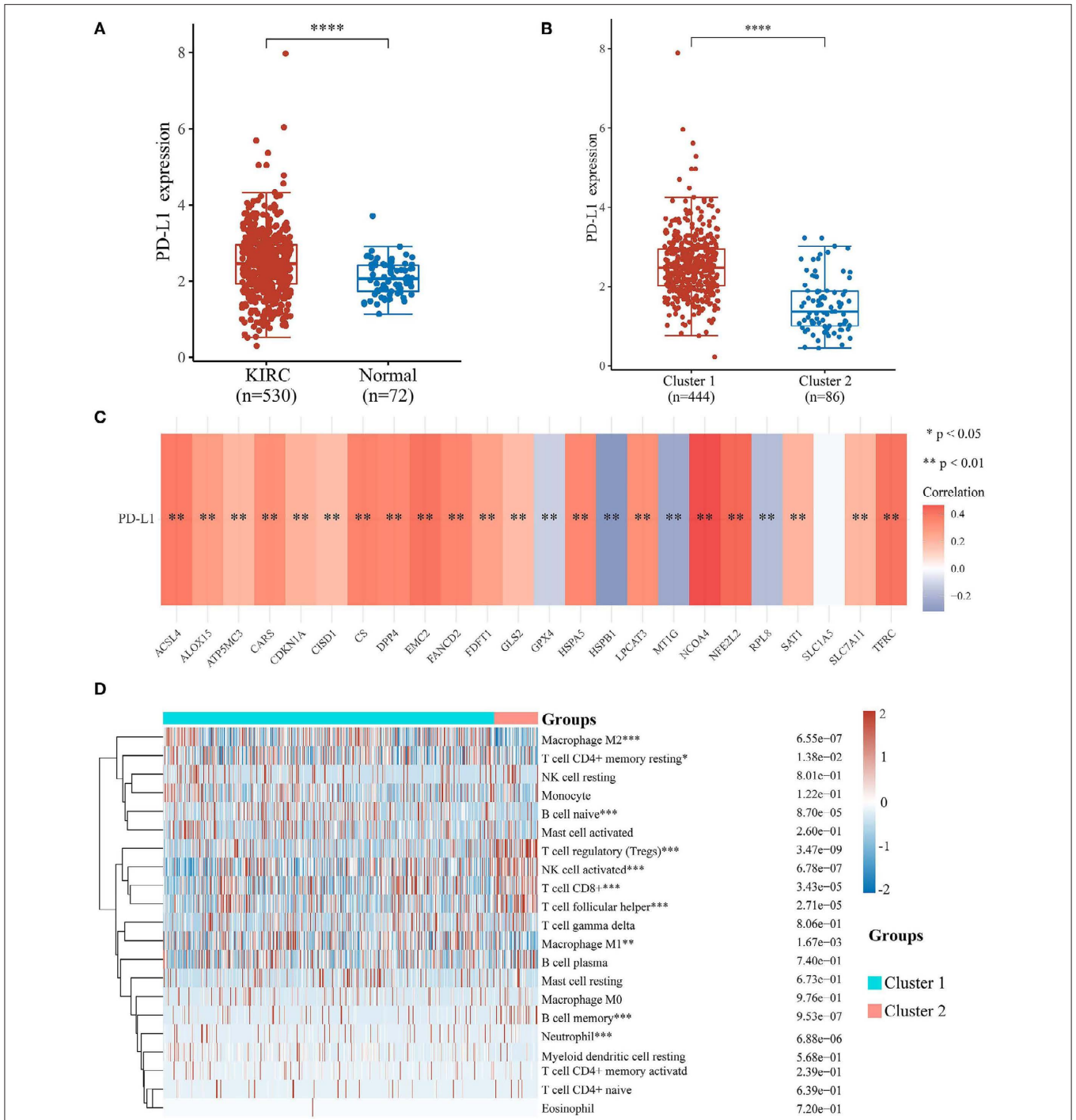


FIGURE 3 | Association of PD-L1 with ferroptosis regulators and the differential infiltration level of tumor immune cells in two clear cell renal cell carcinoma (ccRCC) subtypes. **(A,B)** The expression level of PD-L1 in ccRCC/normal group **(A)** and cluster 1/2 subtype **(B)** in ccRCC. **(C)** The correlation of PD-L1 with ferroptosis regulators in the TCGA-KIRC cohort. **(D)** The infiltrating levels of various immune cell types in two subtypes in the TCGA-KIRC cohort. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, and ****$p < 0.0001$.

DISCUSSION

Ferroptosis is a novel iron-dependent programmed cell death process, which is characteristically distinct from autophagy, apoptosis, and other forms of necrosis (Lei et al., 2020). Recently, emerging evidence reported the vital roles of ferroptosis in the tumorigenesis of human malignancies (Ding et al., 2021), especially the ccRCC which is intrinsically vulnerable to

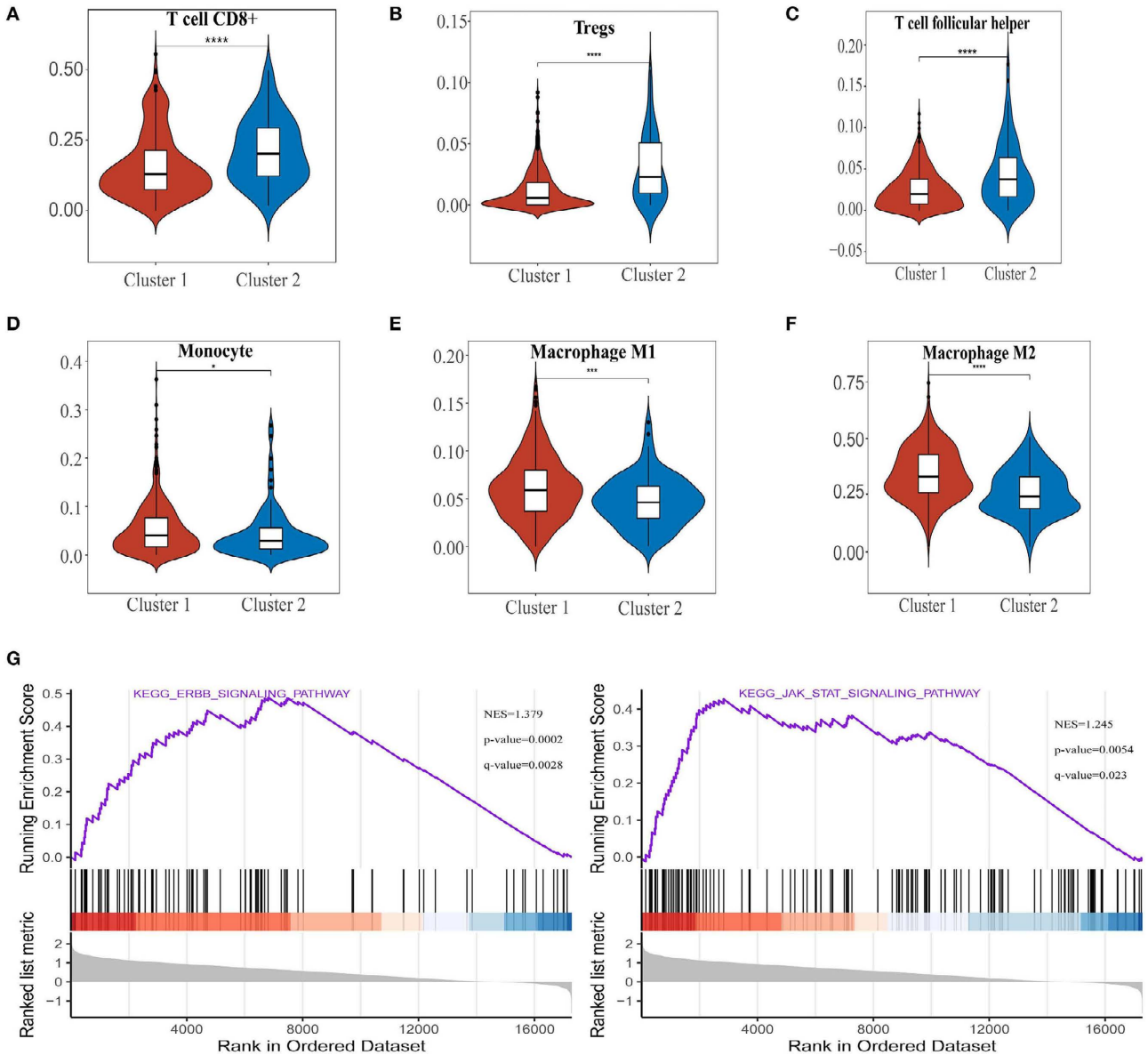


FIGURE 4 | The differences of immune cell infiltration levels in two clear cell renal cell carcinoma (ccRCC) subtypes. **(A–F)** The infiltration level of the CD8+ T cell **(A)**, Tregs **(B)**, and T follicular helper cell **(C)** were downregulated in cluster 1 ccRCC patients. The infiltration levels of the monocyte **(D)**, M1 macrophage cell **(E)**, and M2 macrophage cell **(F)** were upregulated in cluster 1 ccRCC patients. **(G)** Gene set enrichment analysis indicated that ERBB signaling and the JAK-STAT signaling pathways are significantly enriched in cluster 1. NES, normalized enrichment score; *q*-value, false discovery rate. $^{*}p < 0.05$, $^{***}p < 0.001$, and $^{****}p < 0.0001$.

ferroptosis (Zou et al., 2019). However, the potential role of ferroptosis in the ccRCC immune microenvironment remains elusive. Here we distinguished two independent subtypes with differential clinical features, prognoses, PD-L1 expression, and TIME *via* consensus clustering for selected ferroptosis regulators. Among these, CARS was identified as the potential immune infiltration-related ferroptosis regulator. Furthermore, pan-cancer analysis revealed a close correlation of CARS with the expression of PD-L1 and poor overall survival in multiple cancer types.

In recent years, precision medicine has revolutionized the treatment of human cancers (Dupont et al., 2021), which emphasizes that, by integrating multidimensional biological data and clinical characteristics, highly heterogeneous cancers can be subdivided into more accurate subtypes for individualized treatment (Zhang et al., 2020). Indeed accumulating evidence established subgroups of tumor patients according to their molecular profiles, representing distinct phenotypes, prognoses, and therapy responses (Yan et al., 2018; Tan et al., 2019). For instance, Zhou et al. classified colorectal cancer patients into

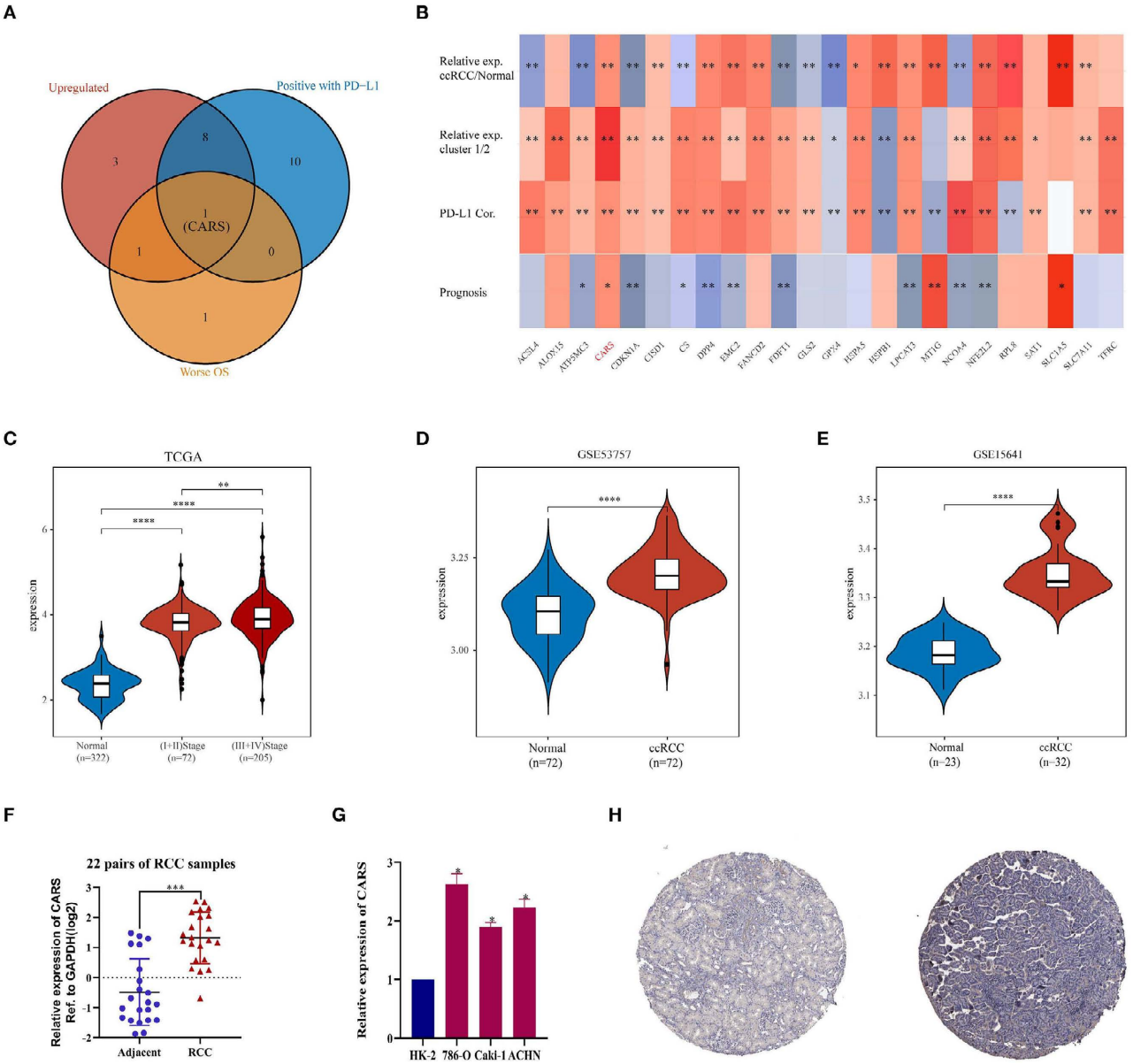


FIGURE 5 | Analysis of the CARS gene expression in ccRCC. **(A)** The Venn diagram suggested that the upregulation of CARS was an unfavorable prognostic factor and positively correlated with the PD-L1 expression in clear cell renal cell carcinoma (ccRCC). **(B)** The heat map visualized the relative expression of 24 ferroptosis regulators in ccRCC/normal (row 1) and cluster 1/2 (row 2) and the correlations of 24 regulators with the PD-L1 expression (row 3) and prognosis (row 4). The red color indicates a high expression in ccRCC or cluster 1, positively correlated with PD-L1 or bad prognosis. $^{*}p < 0.05$ and $^{**}p < 0.01$. **(C)** CARS was upregulated in ccRCC tissues compared to normal tissues, and the expression significantly correlated with AJCC stages. **(D,E)** Validation sets showed that CARS was highly expressed in ccRCC tissues compared to adjacent normal tissues in GSE53757 **(D)** and GSE156417 **(E)**. **(F)** Relative expression of CARS was detected by qRT-PCR in 22 pairs of ccRCC and normal tissues. **(G)** Compared with HK-2 cell line (normal), CARS was highly expressed in 786-O, ACHN, and caki-1 cell lines (ccRCC). **(H)** Immunohistochemistry staining indicated that, compared with normal kidney tissue (left), CARS was significantly elevated in ccRCC tissue (right) in the human protein atlas (antibody HPA002384, ×10). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, and $^{****}p < 0.0001$.

high- and low-risk groups according to their autophagy features, and consequently, more aggressive and targeted therapies were required in the high-risk group (Zhou et al., 2019). Based on immune-related gene expression signatures, lung adenocarcinoma patients can be divided into two prognostically and clinically relevant subtypes. Patients from the high-risk immune subtype were more responsive to immune checkpoint blockade treatment (Wang et al., 2020). In this study, we applied consensus clustering analysis to classify ccRCC patients into two subgroups or clusters on the basis of ferroptosis regulator expression signatures. Low tumor grade, favorable prognosis, and upregulated PD-L1 expression level were observed in cluster

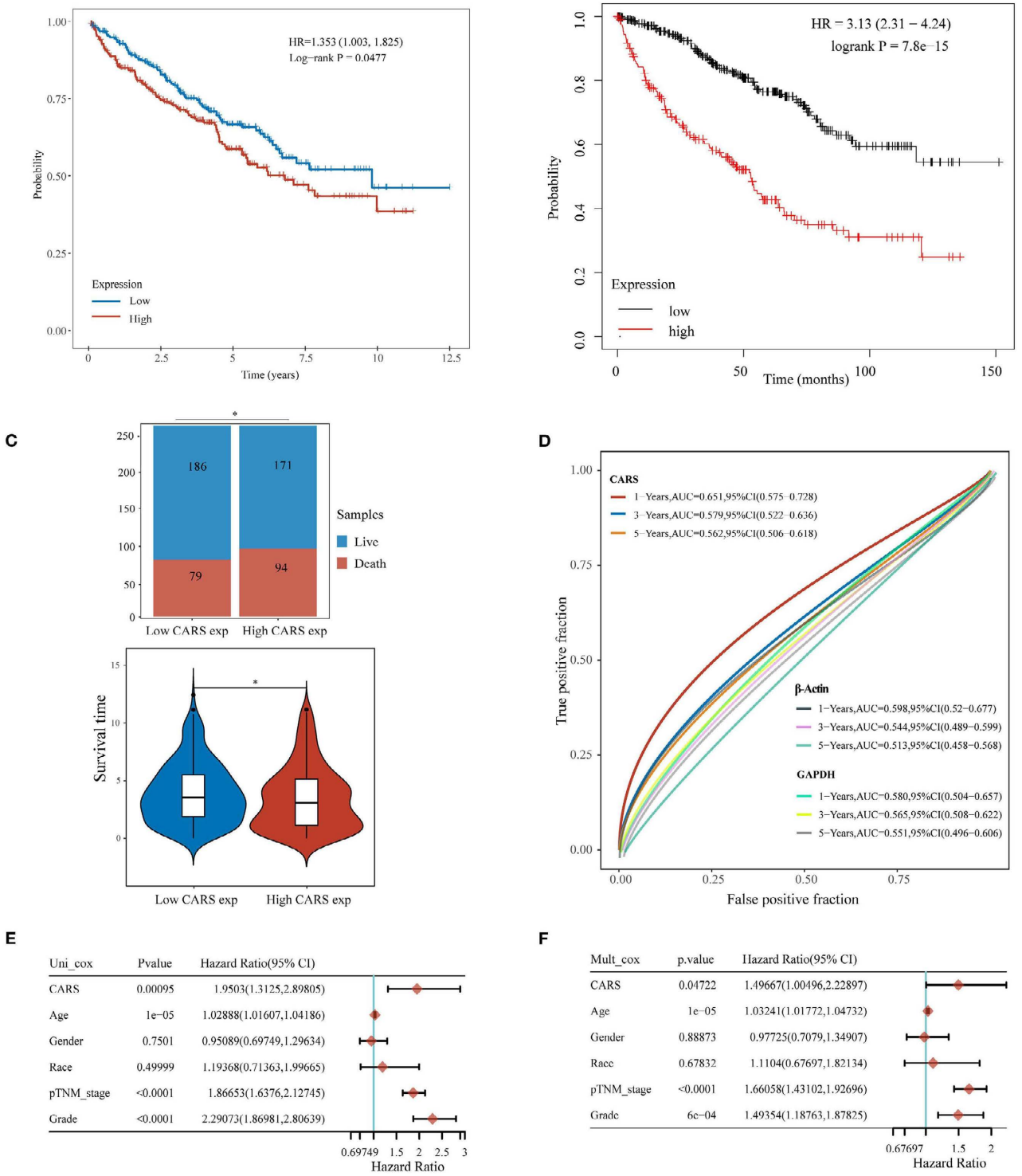


FIGURE 6 | Upregulated CARS expression is associated with poor outcomes of clear cell renal cell carcinoma (ccRCC). **(A,B)** The Kaplan–Meier analysis of ccRCC patients with high and low CARS expression level in the TCGA cohort **(A)** and Kaplan–Meier plotter database **(B)**. **(C)** The stacked bar chart and violin plot visualized the proportion of live/death and survival time in high/low CARS groups in ccRCC patients. **(D)** Time-dependent receiver operating characteristics analysis of CARS and the two most common housekeeping genes, GAPDH and β-Actin. **(E,F)** Forest plots based upon the outcomes of univariate **(E)** and multivariate Cox regression **(F)** of CARS expression and other clinicopathological factors. $^*p < 0.05$.

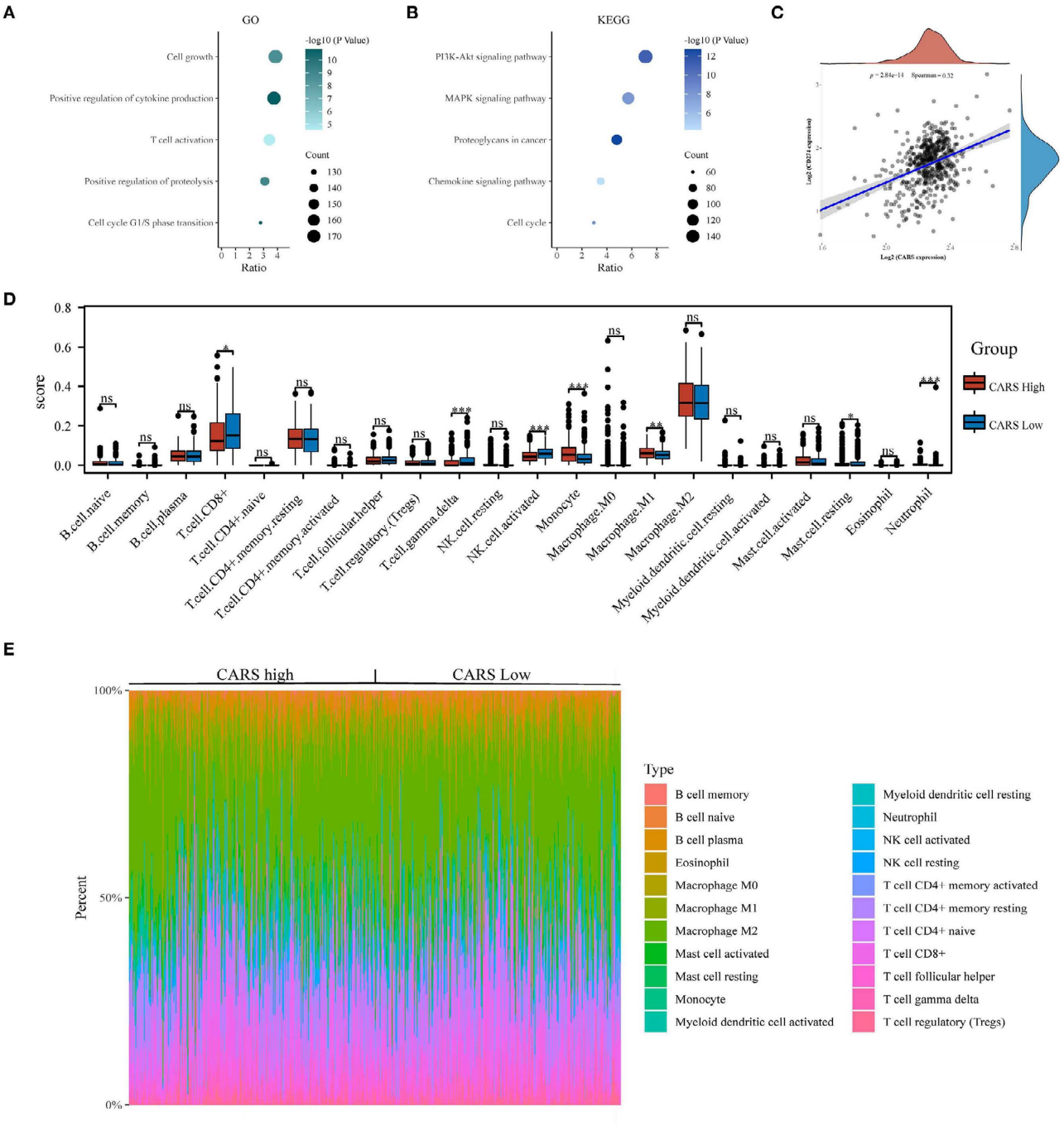


FIGURE 7 | Correlation analysis of CARS expression with PD-L1 and infiltrating immune cells in clear cell renal cell carcinoma (ccRCC). **(A,B)** The enriched Gene Ontology **(A)** and Kyoto Encyclopedia of Genes and Genomes signaling pathways **(B)** analysis of highly expressed CARS in ccRCC. **(C)** Spearman correlation analysis of CARS expression and PD-L1 expression in ccRCC. **(D)** The infiltrating levels of immune cells in high and low CARS expression groups in ccRCC patients. **(E)** The heat map visualized the percentage abundance of tumor-infiltrating immune cells in each sample. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

1 patients. This also meant that cluster 1 patients benefit more from anti-PD-L1 antibodies even with the occurrence of metastasis. Meanwhile, further analysis revealed that cluster 1 harbored significantly higher infiltrated levels of monocyte, M1 macrophage cell, and M2 macrophage cell. On the other hand, cluster 2, with unfavorable survival, was more correlated with CD8+ T cell, Tregs, and T follicular helper cells. The role of Tregs and T follicular helper cells, which were upregulated in cluster 2, in tumor immune escape was well-established (Serrels et al., 2015). Intriguingly, CD8+ T cells are a type of well-known

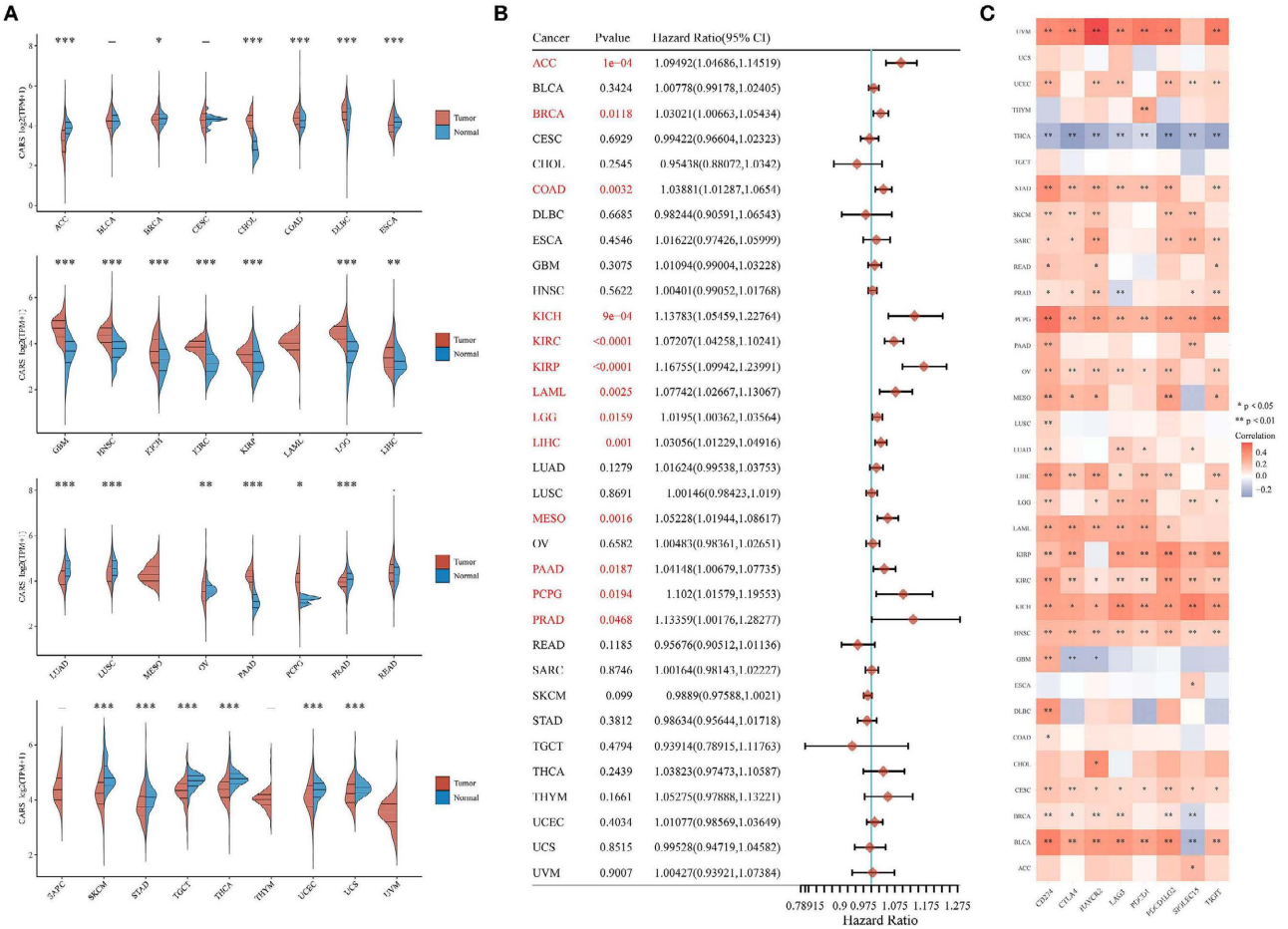


FIGURE 8 | Comprehensive analysis of key ferroptosis regulator CARS in pan-cancers. **(A)** CARS was differentially expressed in most cancer types. **(B)** Highly expressed CARS showed poor overall survival in various cancers. **(C)** The association of CARS expression with the expression of eight common immune checkpoints was observed in most cancer types, especially the expression level of PD-L1. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

effector cells. Unlike the vast majority of cancers, high CD8+ T cell infiltration level was recognized as an adverse prognostic factor in ccRCC (Braun et al., 2020; Bruni et al., 2020; Dai et al., 2021). Moreover, Clark et al. classified ccRCC into four subtypes. Among these, CD8+-inflamed ccRCC with a highly infiltrated level of CD8+ T cell is tightly linked to an increased proportion of higher-grade tumors and worse prognosis, while CD8- inflamed tumors with low infiltration of CD8+ T cell were characterized by an innate immune signature, which showed an increased level of macrophages (Clark et al., 2019). Additionally, several recent studies suggested that tumor-killing T cell and anti-PD-L1 antibodies can induce ferroptosis in tumor cells, while liproxstatin-1 (ferroptosis depressor) can reduce the anticancer efficacy of these agents (Lang et al., 2019; Wang et al., 2019). These findings suggest an overwhelming complexity in the role of ferroptosis in the TIME. To further explore the possible mechanisms of the distinct TIME between two clusters, we carried out the GSEA analysis. The results revealed that the ERBB signaling and the JAK_STAT signaling were significantly enriched in cluster 1. Notably, ERBB family members activated downstream of signaling pathways are widespread, including the PI3K–AKT–mTOR and MAPK pathways, which not only participate in the endogenous carcinogenic pathway but also affect the tumor immune microenvironment (Yarden and Pines, 2012). Yi et al. disclosed that activation of PI3K–AKT–mTOR signaling suppresses ferroptosis in cancer (Yi et al., 2020). Liu et al. found elsewhere that the interaction between GPX4 and mTOR signaling modulates ferroptotic cancer cell death (Liu et al., 2021). Ferroptosis regulator DPP4 was reported to activate the MAPK pathway in papillary thyroid carcinoma (Hu et al., 2021). These studies have indicated that the interaction between ferroptosis and ERBB signaling pathway might be implicated in regulating the TIME of ccRCC.

Given the vital roles of ferroptosis in ccRCC TIME, we identified adverse prognostic regulators among the selected 24 ferroptosis regulators, elevated in tumors and positively correlated with PD-L1 expression. Interestingly, CARS was the only key ferroptosis regulator corresponding to these three conditions. CARS (also CARS1) promote cysteine metabolism (Akaike et al., 2017), hence regulating ferroptosis. Hayano

et al. demonstrated that the knockdown of CARS can induce the transsulfuration pathway, causing cystine deprivation and thereby inhibiting ferroptosis in osteosarcoma, Ewing's sarcoma, and pancreatic carcinoma (Hayano et al., 2016). A recent study found that CARS1 has a unique domain (UNE-C1) that influences the immune response (Cho et al., 2020). Nevertheless, the potential role of CARS in ccRCC TIME remains puzzling. Herein we demonstrated the overexpression and unfavorable prognostic value of CARS in multiple ways. Furthermore, CIBERSORT analysis revealed that the levels of monocyte and M1 macrophages were increased in the high CARS expression group, whereas the levels of CD8+ T cell and NK-activated cells were reduced. This implied a possible regulation role of CARS in the polarization of TAM. Notably, significantly high infiltration levels of NK-activated cells were observed in the low CARS expression group. NK-activated cells are a type of cytotoxic cell; they are directly cytotoxic to tumor cells but also produce cytokines (TNF-α, IFN-γ, etc.), inhibiting cancer angiogenesis and proliferation (Whiteside, 2020). Therefore, CARS might play a pivotal role in the regulation of the tumor microenvironment in ccRCC. Given that ferroptosis is widely involved in various biological processes, we sought to establish whether CARS, the key ferroptosis regulator of our study, also plays a vital role in other cancer types. Subsequently, a pan-cancer analysis was performed, and the results demonstrated that CARS was closely related to the expression of immune checkpoint genes (specifically the expression of PD-L1) in most cancer types. Also, CARS overexpression showed poor OS in 13 cancer types but showed no protective effect in any cancer types. These results implied that CARS is a potential target for cancer immunotherapy and an effective prognostic biomarker not limited to ccRCC.

Nevertheless, several potential limitations in this study deserve to be mentioned. First, we selected 24 key ferroptosis regulators for analysis in this study, according to a comprehensive previous study. Thus, it is probable that ferroptosis regulators were left out before and after our analysis. Furthermore, the clustering subtypes as well as the interactions between the TIME and ferroptosis regulators were analyzed only based on available data from the TCGA due to lacking sufficient data in our own cohort. As such, further reliable validation analysis in larger cohorts needs to be carried out in the future. Besides this, additional experiments *in vitro* and *in vivo* are warranted to be performed to elucidate the regulatory mechanism between ferroptosis regulators and TIME.

In summary, this research systematically analyzed the expression profiles and the correlations with prognosis, PD-L1, and role in TIME of ferroptosis regulators in ccRCC. The ferroptosis regulators' expressions were significantly associated with PD-L1. Two independent subtypes were also established by consensus clustering for ferroptosis regulators. Obvious tumor heterogeneity, distinct PD-L1 expression, and TIME were observed between the two subtypes, which will contribute to the risk stratification and precision therapy for patients with ccRCC, and we also found that the interaction between ferroptosis and the ERBB signaling pathway might be implicated in regulating the TIME of ccRCC. Among the selected ferroptosis regulators, CARS was subsequently identified as an independent prognostic indicator of ccRCC patients and positively correlated with PD-L1 expression. Further pan-cancer analysis demonstrated that CARS closely related to the expression of immune checkpoint genes (especially the expression of PD-L1) and poor prognosis in diverse cancer types. Certainly, a series of clinical trials and basic researches might help to determine targets for enhancing the efficacy of cancer immunotherapy.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by First Affiliated Hospital, School of Medicine, Zhejiang University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

SW drafted and wrote the manuscript. SW, XM, YY, and HS collected related data. SW and SC conducted the bioinformatics analysis with R software. JL, YL, and BL searched the existing related literatures. XW, XZ, and LX revised the manuscript. All the authors read and approved the final manuscript.

FUNDING

This study was supported by the Natural Science Foundation of Zhejiang Province, Grant/award No. LY20H160022, and the National Natural Science Foundation of China, Grant/award Nos. 81802564, 81402096, and 81972374.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.676142/full#supplementary-material

Supplementary Figure 1 | Consensus clustering for ferroptosis regulators in clear cell renal cell carcinoma (ccRCC). **(A–D)** Four heat maps exhibit the clustering matrix for ferroptosis regulators in ccRCC patients for $k = 3$, 4, 5, and 6. The tighter and clearer the clusters are, the more optimal the cluster. **(E)** Delta area curve of consensus clustering for $k = 2$–6. **(F)** Principal component analysis of ccRCC patients' ferroptosis regulator expression profiles demonstrates two patient clusters.

Supplementary Table 1 | The abbreviations of multiple cancer types, ferroptosis regulators and immune-related checkpoints in this study.

Supplementary Table 2 | The patients' clinical information ($n = 22$) in this study.

Supplementary Table 3 | The oligonucleotides used in this study. *F, forward primer; R, reverse primer.

REFERENCES

Akaike, T., Ida, T., Wei, F., Nishida, M., Kumagai, Y., Alam, M., et al. (2017). Cysteinyl-tRNA synthetase governs cysteine polysulfidation and mitochondrial bioenergetics. *Nat. Commun.* 8:1177. doi: 10.1038/s41467-017-01311-y

Barata, P., and Rini, B. (2017). Treatment of renal cell carcinoma: current status and future directions. *CA Cancer J. Clin.* 67, 507–524. doi: 10.3322/caac.21411

Braun, D., Hou, Y., Bakouny, Z., Ficial, M., Sant' Angelo, M., Forman, J., et al. (2020). Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. *Nat. Med.* 26, 909–918. doi: 10.1038/s41591-020-0839-y

Bruni, D., Angell, H., and Galon, J. (2020). The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat. Rev. Cancer* 20, 662–680. doi: 10.1038/s41568-020-0285-7

Capitanio, U., Bensalah, K., Bex, A., Boorjian, S., Bray, F., Coleman, J., et al. (2019). Epidemiology of renal cell carcinoma. *Euro. Urol.* 5, 74–84. doi: 10.1016/j.eururo.2018.08.036

Chen, L., Hambright, W., Na, R., and Ran, Q. (2015). Ablation of the ferroptosis inhibitor glutathione peroxidase 4 in neurons results in rapid motor neuron degeneration and paralysis. *J. Biol. Chem.* 290, 28097–28106. doi: 10.1074/jbc.M115.680090

Cho, S., Kim, S., Lee, Y., Song, E., Kim, U., Kim, H., et al. (2020). Endogenous TLR2 ligand embedded in the catalytic region of human cysteinyl-tRNA synthetase 1. *J. Immunother. Cancer* 8:e000277. doi: 10.1136/jitc-2019-000277

Choueiri, T., Larkin, J., Oya, M., Thistlethwaite, F., Martignoni, M., Nathan, P., et al. (2018). Preliminary results for avelumab plus axitinib as first-line therapy in patients with advanced clear-cell renal-cell carcinoma (JAVELIN Renal 100): an open-label, dose-finding and dose-expansion, phase 1b trial. *Lancet Oncol.* 19, 451–460. doi: 10.1016/S1470-2045(18)30107-4

Clark, D., Dhanasekaran, S., Petralia, F., Pan, J., Song, X., Hu, Y., et al. (2019). Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell* 179, 964–983.e931. doi: 10.1016/j.cell.2019.10.007

Dai, S., Zeng, H., Liu, Z., Jin, K., Jiang, W., Wang, Z., et al. (2021). Intratumoral CXCL13CD8T cell infiltration determines poor clinical outcomes and immunoevasive contexture in patients with clear cell renal cell carcinoma. *J. Immunother. Cancer* 9:e001823. doi: 10.1136/jitc-2020-001823

Ding, Y., Chen, X., Liu, C., Ge, W., Wang, Q., Hao, X., et al. (2021). Identification of a small molecule as inducer of ferroptosis and apoptosis through ubiquitination of GPX4 in triple negative breast cancer cells. *J. Hematol. Oncol.* 14:19. doi: 10.1186/s13045-020-01016-8

Dixon, S., Lemberg, K., Lamprecht, M., Skouta, R., Zaitsev, E., Gleason, C., et al. (2012). Ferroptosis: an iron-dependent form of nonapoptotic cell death. *Cell* 149, 1060–1072. doi: 10.1016/j.cell.2012.03.042

Dupont, C., Riegel, K., Pompaiah, M., Juhl, H., and Rajalingam, K. (2021). Druggable genome and precision medicine in cancer: current challenges. *FEBS J.* 288, 2311–2331. doi: 10.1111/febs.15788

Eling, N., Reuter, L., Hazin, J., Hamacher-Brady, A., and Brady, N. (2015). Identification of artesunate as a specific activator of ferroptosis in pancreatic cancer cells. *Oncoscience* 2, 517–532. doi: 10.18632/oncoscience.160

Hayano, M., Yang, W., Corn, C., Pagano, N., and Stockwell, B. (2016). Loss of cysteinyl-tRNA synthetase (CARS) induces the transsulfuration pathway and inhibits ferroptosis induced by cystine deprivation. *Cell Death Differ.* 23, 270–278. doi: 10.1038/cdd.2015.93

Hu, C., Nydes, M., Shanley, K., Morales Pantoja, I., Howard, T., and Bizzozero, O. (2019). Reduced expression of the ferroptosis inhibitor glutathione peroxidase-4 in multiple sclerosis and experimental autoimmune encephalomyelitis. *J. Neurochem.* 148, 426–439. doi: 10.1111/jnc.14604

Hu, X., Chen, S., Xie, C., Li, Z., Wu, Z., and You, Z. (2021). DPP4 gene silencing inhibits proliferation and epithelial-mesenchymal transition of papillary thyroid carcinoma cells through suppression of the MAPK pathway. *J. Endocrinol. Invest*. doi: 10.1007/s40618-020-01455-7. [Epub ahead of print].

Lang, X., Green, M., Wang, W., Yu, J., Choi, J., Jiang, L., et al. (2019). Radiotherapy and immunotherapy promote tumoral lipid oxidation and ferroptosis via synergistic repression of SLC7A11. *Cancer Discov.* 9, 1673–1685. doi: 10.1158/2159-8290.CD-19-0338

Lei, G., Zhang, Y., Koppula, P., Liu, X., Zhang, J., Lin, S., et al. (2020). The role of ferroptosis in ionizing radiation-induced cell death and tumor suppression. *Cell Res.* 30, 146–162. doi: 10.1038/s41422-019-0263-3

Liu, Y., Wang, Y., Liu, J., Kang, R., and Tang, D. (2021). Interplay between MTOR and GPX4 signaling modulates autophagy-dependent ferroptotic cancer cell death. *Cancer Gene Ther.* 28, 55–63. doi: 10.1038/s41417-020-0182-y

Liu, Z., Zhao, Q., Zuo, Z., Yuan, S., Yu, K., Zhang, Q., et al. (2020). Systematic analysis of the aberrances and functional implications of ferroptosis in cancer. *iScience* 23:101302. doi: 10.1016/j.isci.2020.101302

Motzer, R., Escudier, B., McDermott, D., George, S., Hammers, H., Srinivas, S., et al. (2015). Nivolumab versus everolimus in advanced renal-cell carcinoma. *N. Engl. J. Med.* 373, 1803–1813. doi: 10.1056/NEJMoa1510665

Nagy, Á., Lánczky, A., Menyhárt, O., and Gyorffy, B. (2018). Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Sci. Rep.* 8:9227. doi: 10.1038/s41598-018-27521-y

Serrels, A., Lund, T., Serrels, B., Byron, A., McPherson, R., von Kriegsheim, A., et al. (2015). Nuclear FAK controls chemokine transcription, Tregs, and evasion of anti-tumor immunity. *Cell* 163, 160–173. doi: 10.1016/j.cell.2015.09.001

Sung, H., Ferlay, J., Siegel, R., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 72, 209–249. doi: 10.3322/caac.21660

Tan, T., Rouanne, M., Tan, K., Huang, R., and Thiery, J. (2019). Molecular subtypes of urothelial bladder cancer: results from a meta-cohort analysis of 2411 tumors. *Euro. Urol.* 75, 423–432. doi: 10.1016/j.eururo.2018.08.027

Tuo, Q., Lei, P., Jackman, K., Li, X., Xiong, H., Li, X., et al. (2017). Tau-mediated iron export prevents ferroptotic damage after ischemic stroke. *Mol. Psychiatry* 22, 1520–1530. doi: 10.1038/mp.2017.171

Viswanathan, V., Ryan, M., Dhruv, H., Gill, S., Eichhoff, O., Seashore-Ludlow, B., et al. (2017). Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. *Nature* 547, 453–457. doi: 10.1038/nature23007

Wang, Q., Li, M., Yang, M., Yang, Y., Song, F., Zhang, W., et al. (2020). Analysis of immune-related signatures of lung adenocarcinoma identified two distinct subtypes: implications for immune checkpoint blockade therapy. *Aging* 12, 3312–3339. doi: 10.18632/aging.102814

Wang, W., Green, M., Choi, J., Gijón, M., Kennedy, P., Johnson, J., et al. (2019). CD8 T cells regulate tumour ferroptosis during cancer immunotherapy. *Nature* 569, 270–274. doi: 10.1038/s41586-019-1170-y

Whiteside, T. (2020). NK cells in the tumor microenvironment and thioredoxin activity. *J. Clin. Invest.* 130, 5115–5117. doi: 10.1172/JCI141460

Yan, H., Siu, H., Law, S., Ho, S., Yue, S., Tsui, W., et al. (2018). A Comprehensive Human gastric cancer organoid biobank captures tumor subtype heterogeneity and enables therapeutic screening. *Cell Stem Cell* 23, 882–897.e811. doi: 10.1016/j.stem.2018.09.016

Yarden, Y., and Pines, G. (2012). The ERBB network: at last, cancer therapy meets systems biology. *Nat. Rev. Cancer* 12, 553–563. doi: 10.1038/nrc3309

Yi, J., Zhu, J., Wu, J., Thompson, C., and Jiang, X. (2020). Oncogenic activation of PI3K-AKT-mTOR signaling suppresses ferroptosis via SREBP-mediated lipogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 117, 31189–31197. doi: 10.1073/pnas.2017152117

Yu, H., Guo, P., Xie, X., Wang, Y., and Chen, G. (2017). Ferroptosis, a new form of cell death, and its relationships with tumourous diseases. *J. Cell. Mol. Med.* 21, 648–657. doi: 10.1111/jcmm.13008

Zhang, H., Ahearn, T., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., et al. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* 52, 572–581. doi: 10.1038/s41588-020-0609-2

Zhang, Z., Lu, M., Chen, C., Tong, X., Li, Y., Yang, K., et al. (2021). Holo-lactoferrin: the link between ferroptosis and radiotherapy in triple-negative breast cancer. *Theranostics* 11, 3167–3182. doi: 10.7150/thno.52028

Zhou, Z., Mo, S., Dai, W., Ying, Z., Zhang, L., Xiang, W., et al. (2019). Development and validation of an autophagy score signature for the prediction of post-operative survival in colorectal cancer. *Front. Oncol.* 9:878. doi: 10.3389/fonc.2019.00878

Zou, Y., Palte, M., Deik, A., Li, H., Eaton, J., Wang, W., et al. (2019). A GPX4-dependent cancer cell state underlies the clear-cell morphology and confers sensitivity to ferroptosis. *Nat. Commun.* 10:1617. doi: 10.1038/s41467-019-09277-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Wang, Chen, Ying, Ma, Shen, Li, Wang, Lin, Liu, Zheng and Xie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 06 July 2021
doi: 10.3389/fcell.2021.702112

Characterizing the Metabolic and Immune Landscape of Non-small Cell Lung Cancer Reveals Prognostic Biomarkers Through Omics Data Integration

Fengjiao Wang[1†], Yuanfu Zhang[2†], Yangyang Hao[2†], Xuexin Li[3†], Yue Qi[2], Mengyu Xin[2], Qifan Xiao[1] and Peng Wang[2]*

[1] Department of Thoracic Surgery, Harbin Medical University Cancer Hospital, Harbin, China, [2] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [3] Department of Urinary Surgery, Harbin Medical University Cancer Hospital, Harbin, China

Non-small cell lung cancer (NSCLC) is one of the most common malignancies worldwide. The development of high-throughput single-cell RNA-sequencing (RNA-seq) technology and the advent of multi-omics have provided a solid basis for a systematic understanding of the heterogeneity in cancers. Although numerous studies have revealed the molecular features of NSCLC, it is important to identify and validate the molecular biomarkers related to specific NSCLC phenotypes at single-cell resolution. In this study, we analyzed and validated single-cell RNA-seq data by integrating multi-level omics data to identify key metabolic features and prognostic biomarkers in NSCLC. High-throughput single-cell RNA-seq data, including 4887 cellular gene expression profiles from NSCLC tissues, were analyzed. After pre-processing, the cells were clustered into 12 clusters using the t-SNE clustering algorithm, and the cell types were defined according to the marker genes. Malignant epithelial cells exhibit individual differences in molecular features and intra-tissue metabolic heterogeneity. We found that oxidative phosphorylation (OXPHOS) and glycolytic pathway activity are major contributors to intra-tissue metabolic heterogeneity of malignant epithelial cells and T cells. Furthermore, we constructed T-cell differentiation trajectories and identified several key genes that regulate the cellular phenotype. By screening for genes associated with T-cell differentiation using the Lasso algorithm and Cox risk regression, we identified four prognostic marker genes for NSCLC. In summary, our study revealed metabolic features and prognostic markers of NSCLC at single-cell resolution, which provides novel findings on molecular biomarkers and signatures of cancers.

Keywords: omics data integration, single cell sequencing, prognostic biomarkers, NSCLC, cellular phenotypes

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Shengli Li,
Shanghai Jiao Tong University, China
Pengping Li,
Nanjing Medical University, China

****Correspondence:***
Peng Wang
wpgqy@163.com

[†]These authors share first authorship

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *29 April 2021*
Accepted: *16 June 2021*
Published: *06 July 2021*

Citation:
Wang F, Zhang Y, Hao Y, Li X, Qi Y, Xin M, Xiao Q and Wang P (2021) Characterizing the Metabolic and Immune Landscape of Non-small Cell Lung Cancer Reveals Prognostic Biomarkers Through Omics Data Integration. Front. Cell Dev. Biol. 9:702112. doi: 10.3389/fcell.2021.702112

INTRODUCTION

Lung cancer is the leading cause of cancer-related deaths worldwide (Torre et al., 2015). Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) are the two classic histological subtypes of lung cancer. SCLC generally occurs in people of advanced age with a history of heavy smoking and accounts for approximately 15% of lung cancer cases (van Meerbeeck et al., 2011).

NSCLC represents the remaining 85% of lung cancers and contains two main histological subtypes: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) (Goldstraw et al., 2011). Since NSCLC has a wide range of genomic variation (Cancer Genome Atlas Research Network, 2014), it can respond better to immune checkpoint blockade (Rizvi et al., 2015), although there are exceptions (Topalian et al., 2016). With the development of single-cell sequencing technology, there is an increasing number of studies on identifying molecular features in the tumor microenvironment (TME) of NSCLC (Guo et al., 2018; Kim et al., 2020). However, there has been no significant improvement in the treatment of NSCLC. Therefore, it is necessary to explore the molecular features and prognostic markers of malignant cells at single-cell resolution.

During tumor tissue deterioration, cancer cells are reprogrammed by physiological mechanisms, including metabolic reprogramming, to support the demand for energy, biomass, and cellular communication (DeBerardinis and Chandel, 2016). Cell metabolism is influenced by genetic and environmental factors, including mutations that determine the direction of cell evolution, nutrients, and tissue origin (Boroughs and DeBerardinis, 2015; Pavlova and Thompson, 2016). The metabolic activity of cells is determined by the concentration of metabolically relevant molecules and biomolecule conversion rate, but these indicators are difficult to measure. Therefore, it is necessary to evaluate the expression of metabolic genes to indirectly determine the metabolic activity of cells (Puram et al., 2017).

The tumor tissue microenvironment (TME) is composed of malignant cells, fibroblasts, immune cells, and many other stromal cells (Bian et al., 2019). Each cell type plays an active role in tumor proliferation and metastasis. For example, cancer-associated fibroblasts (CAFs) assist in the invasion of tumor cells (Belle and DeNardo, 2019). Due to the different functions of each cell type, they all have specific metabolic requirements. Not only does each cell type have specific metabolic activities, but at the cellular level, each cell also has specific metabolic activities depending on its environment and evolutionary direction (Sottoriva et al., 2015; Reina-Campos et al., 2017; Reznik et al., 2018). Most conclusions about the metabolic features of the TME are derived from *in vitro* experiments (Carmona-Fontaine et al., 2017; Liu et al., 2018) or univariate measurements of metabolic enzyme expression (Miller et al., 2017), but these studies modify the TME to some extent.

The lymphocyte lineage is an important component of the TME, which has become a popular area in cancer immunotherapy. Cancer immunotherapy shows individual differences in the treatment of NSCLC (Topalian et al., 2012; Hellmann et al., 2019), which depends, in part, on the amount and properties of tumor-infiltrating lymphocytes (Rizvi et al., 2015; Huang et al., 2017). CD8+ and CD4+ T cells, as two subtypes of T cells, play an important role in the anti-tumor process. Although studies have pointed out a positive correlation between elevated CD8+ T cells and a good prognosis of cancer (Maimela et al., 2019), the mechanism that drives T-cell differentiation is unclear.

In this study, we explored the expression profiles of 4887 cells from the tumor tissues of four patients with NSCLC to identify the metabolic features of malignant cells and mechanisms that drive T-cell differentiation at the single-cell level, as well as to discover new therapeutic targets and prognostic markers.

MATERIALS AND METHODS

Data Collection and Pre-processing

Single-cell RNA-seq profiles for NSCLC were collected from the Gene Expression Omnibus [GEO (Barrett et al., 2013); [1]] under accession number GSE117570. The profiles were derived from four tumor and four paracancerous tissue samples from four patients with NSCLC. To reveal the metabolic and immune features of NSCLC and identify therapeutic targets, four tumor samples containing 4,887 cells were used in this study. Fifteen low-quality cells were filtered out through quality control, leaving 4,872 cells. A gene was selected if it was expressed in at least 3 cells. The ineligible genes were filtered out, and 10,050 genes were retained for analysis. Considering the sequencing depth of these tumor samples, the Scran algorithm (Lun et al., 2016) was used to standardize the count data. The gene sets of the metabolic pathway were downloaded from the Molecular Signatures Database [MSigDB (Liberzon et al., 2011), [2])]. Bulk RNA-seq profiles and clinical data of LUAD and LUSC belonging to NSCLC were collected from The Cancer Genome Atlas (Angelin et al., 2017) (TCGA, [3]). Moreover, the microarray sequencing data and survival data of the GSE3141 and GSE42127 datasets were obtained from the GEO database for the verification of prognostic markers. Information on all the samples used in this study is given in **Supplementary Table 1**.

Dimensionality Reduction and Clustering of Cells

The preprocessed gene expression matrix and cell annotation information were encapsulated using the R package Seurat (version 3.2.2) (Satija et al., 2015). The top 3000 variant genes calculated by the standard deviation (SD) algorithm were used for principal component analysis (PCA). We adopt the strategy of using the least number of principal components to explain more data information (Butler et al., 2018). Thus, the top 16 principal components were manually selected for cell clustering analysis using the t-SNE algorithm. Marker genes of specific cell types in NSCLC tissues collected from published literature (Song et al., 2019) and CellMarker (Zhang et al., 2019b)[4] database were used to define cell clusters.

Metabolic Reprogramming Analysis of Malignant Cells

Due to the different origins and environments of malignant cells, specific malignant cell clusters may have unique metabolic

[1]https://www.ncbi.nlm.nih.gov/gds

[2]http://www.broadinstitute.org/msigdb

[3]https://portal.gdc.cancer.gov/

[4]http://biocc.hrbmu.edu.cn/CellMarker/

mechanisms. The weighted relative pathway activity algorithm (Xiao et al., 2019) was used to evaluate specific metabolic features among specific malignant cell clusters. In this algorithm, the relative expression level of metabolic genes in each cell cluster was defined as the ratio of the mean expression value of cells in a specific cluster to the mean expression value of all cells. Furthermore, the activity score of a pathway of specific cell types was the weighted average of the relative expression of all genes in this pathway. Most importantly, weighting factors, the reciprocal of the number of pathways that include a certain gene, were used to eliminate commonalities between various metabolic pathways. Genes with low expression levels or high deletion rates in the pathway were also deleted to avoid the pathway activity score being affected by these factors. Cell type labels were randomly swapped 5000 times to construct a null distribution of pathway activity scores, which were used to examine the statistical significance of metabolic pathway activity scores in each cell cluster. For each pathway score, a p-value was calculated to assess whether the activity of the pathway was significantly higher or lower than the average.

Evaluation of Intra-Tissue Metabolism Heterogeneity

The SD of each metabolic gene expression in malignant cell expression profiles, which reflects the variability of each metabolic gene in malignant cells, was calculated. Furthermore, genes were sorted in descending order according to their SD, and gene set enrichment analysis (GSEA) (Subramanian et al., 2005) was used to identify metabolic pathways enriched in metabolic genes with the highest variability using the R package fgsea (version 1.14.0).

Evaluation of Developmental Trajectory of T Cells

Based on the definition of various T-cell characteristics in previous studies (O'Shea and Paul, 2010; Michalek et al., 2011), CD4+ (Th/Treg) T cells and CD8+ T cells were distinguished from the original T-cell clusters. For these T-cell types, pseudo-time developmental trajectories were constructed by monocle (Qiu et al., 2017) (version: 2.16.0), which is an algorithm that describes multiple fate decisions in a completely unsupervised manner using the reverse graph embedding method. Multiple branches and nodes were observed throughout the developmental trajectory, and cells on the same branch were considered to have the same state.

Construction of Transcriptional Regulatory Network

The count of biomolecules varies during T-cell development. Pearson correlation analysis was used to explore genes related to the development of T cells. The correlation coefficients of genes with the developmental processes of CD4+ and CD8+ T cells was evaluated. Genes with Pearson's correlation coefficient $|R| > 0.2$ (Shin et al., 2015) and p-value < 0.05, were defined as genes associated with T-cell development, including those associated with the development of CD4+ and CD8+ T cells. To explore the effects of these genes on the development of T cells and their transcriptional regulatory relationships, functional enrichment analysis was conducted, and a transcriptional regulatory network was constructed. The R package clusterProfiler (version 3.16.1) (Yu et al., 2012) was used to examine the enrichment of genes positive related to T-cell development in Gene Ontology (GO) (The Gene Ontology Consortium, 2019) function nodes and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017) pathways. Human transcription factor data were collected from the AnimalTFDB (Hu et al., 2019) 3.0 database[5]. Data for the relationship between transcription factors and their target gene regulation were collected from TRRUST (Han et al., 2018)[6] and ORTI (Vafaee et al., 2016)[7] databases.

Construction of Survival Prediction Model Based on Critical Factors

The lasso algorithm was used to screen for critical genes from the identified T-cell development-related genes, which are associated with the overall survival (OS) of patients with LUAD. Based on these critical factors, a multivariate Cox regression model was constructed, and the significance indicator for each gene was calculated. Furthermore, to predict the OS of patients with LUSC, the critical genes with a p-value < 0.05 were retained to establish a risk prediction model and nomogram for survival analysis. The reliability of this risk prediction model was depicted by the receiver operating characteristic (ROC) curve, and the area under the curve (AUC) was also calculated. Furthermore, the prognostic markers and regression coefficients obtained from the multivariate Cox regression model were used to construct a risk score model. The samples were divided into high-risk and low-risk groups based on the median risk score calculated by the risk score model, and Kaplan-Meier survival analysis was performed to study the difference in OS between these two groups using the bilateral logarithmic rank test. These prognostic marker genes used in the risk score model of LUAD were also connected to other NSCLC survival data.

RESULTS

Individual Differences in Malignant Transformation of Tumor

We used a calculational pipeline to analyze the gene expression profile of NSCLC at the single-cell level (**Figure 1A**). After quality control and normalization, gene expression profiles of 4872 cells from tumors of 4 patients with NSCLC were used for subsequent analysis. All cells were divided into 12 clusters according to the t-SNE clustering algorithm (**Figure 1B**). We annotated the cell types of the 12 cell clusters spanning malignant epithelial cells (0, 7, 8, and 9 clusters marked by EPCAM and KRT18/19), macrophages (1, 3, and 4 clusters marked by MSR1, CD68, and MARCO), T cells (2 clusters marked by CD7, BATF, PPP1R2, and PPP2R5C), stromal cells (5 clusters marked by

[5]http://bioinfo.life.hust.edu.cn/AnimalTFDB/

[6]https://www.grnpedia.org/trrust/

[7]http://orti.sydney.edu.au/about.html

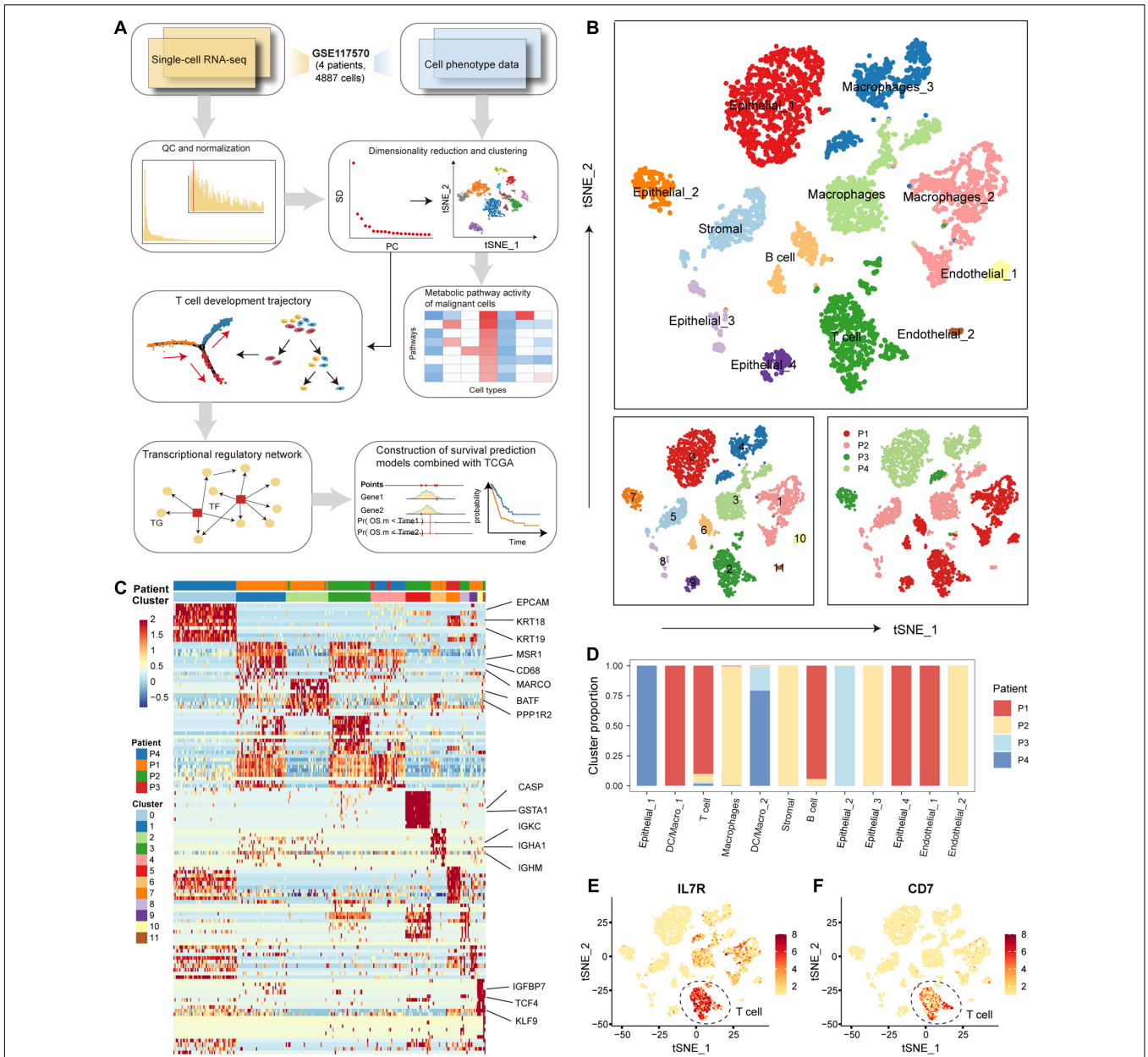


FIGURE 1 | The cellular landscape of non-small cell lung cancer (NSCLC). **(A)** Workflow diagram representing the pipeline for scRNA-seq data analysis. **(B)** Tumor cells were clustered by the t-SNE clustering algorithm into 12 clusters with a specific color marker. **(C)** Heatmap of differentially expressed genes in each cluster. Expression levels of the top 30 genes (rows) that are differentially expressed in each cluster (column). Marker genes in each cluster were annotated. **(D)** Bar plot of the origin of the cells in each cell type. The horizontal axis is the 12 cell types, and the vertical axis is the proportion of cells. **(E)** Distribution of *IL7* gene expression in all tumor cells. The higher the gene expression, the darker the color. **(F)** Same as in **(E)** but for the distribution of gene *CD7*.

CASP and GSTA1), B cells (6 clusters marked by IGKC, IGHA1, IGHG1, and IGHM), endothelial cells (10 and 11 clusters marked by IGFBP7, TCF4, KLF9, and ITGB1) (**Figures 1B,C**). We found 4 clusters of malignant epithelial cells corresponding to their tumors of origin (i.e., from which tumor the cell was derived) and 8 clusters of other stromal cells also corresponding to their tumors of origin, suggesting that NSCLC tissues have obvious individual differences (**Figures 1B,D**). The first patient (patient_1) had a significantly higher proportion of T and B cells compared to the other three patients (**Figure 1D**), suggesting that patient_1 may be more suitable for immune-targeted therapy. The results of the following studies support our inference. We found significantly high expression of *IL7R* and *CD7* genes in T cells (**Figures 1E,F**). Signaling pathways mediated by IL7R are critical for T-cell development and homeostasis *in vivo*, and aberrant IL7R activation was strongly associated with the development of human T-cell leukemogenesis (Zenatti et al., 2011; Yasunaga, 2020). Blocking the expression of

CD7, an immune transmembrane protein, is beneficial in the treatment of T-cell malignancies (Gomes-Silva et al., 2017; Png et al., 2017). These findings suggest that individual differences in NSCLC uncovered at single-cell resolution are critical for the development of precision medicine.

Intra-Tissue Metabolic Heterogeneity of Malignant Cells

During the malignant transformation of the tumor, the metabolic processes of each cell are influenced by the microenvironment, including nutrient concentrations and interactions with other cells in the same space. Therefore, it is intriguing to investigate the metabolic features of malignant cell clusters and the microenvironmental factors that contribute to the intra-tissue metabolic differences. For malignant epithelial cell clusters from four patients, we re-clustered the cells according to their metabolic gene expression profiles. The original 4 clusters of malignant cells were re-clustered into 9 clusters (**Supplementary Figure 1**) 2 clusters from patient 1, 3 clusters from patient 2, 1 cluster from patient 3, and 3 clusters from patient 4. We also separately re-clustered epithelial cells based on the origin of tumor tissue; the cells were clustered into 9 clusters that were similar to the above clustering results (**Figure 2A**), indicating the individual differences of malignant epithelial cells and the stability of the subtype. Furthermore, the weighted pathway activity score algorithm was used to measure the relative activity of the metabolic pathways of these 9 clusters. Among the 85 metabolic pathways, 58 pathways containing at least 5 genes had significantly upregulated activity scores (pathway activity score >1 and permutation test p-value < 0.01) in at least one cell cluster compared to other cell clusters (**Figure 2B**). Malignant cell clusters in patient 3 (p3_0) had the highest number of significantly upregulated metabolic pathways (36 pathways upregulated in p3_0 compared to 30 in p4_2, 22 in p4_0, 21 in p1_0, and <10 in other clusters; **Supplementary Table 2**), which included many different parts of cellular metabolism, such as glycolysis, oxidative phosphorylation (OXPHOS), and the pentose phosphate pathway (**Figure 2B**). We found significant differences between the metabolic pathway activity scores of malignant cell clusters in patients 1–3 (**Figures 2C–E**), suggesting that the activity of metabolic pathways is determined by tumor origin. Although patient 3, cluster 2 of patient 4, and cluster 0 of patient 1 had more upregulated metabolic pathways compared with other clusters, their activity scores demonstrated a poor correlation (**Figure 2F**), indicating specific metabolic reprogramming between tumors.

Next, we identified the microenvironmental factors that contribute to the intra-tissue metabolic heterogeneity. We performed GSEA to identify metabolic pathways enriched in genes explaining most of the variation among the malignant cells of each tumor cluster. We found that OXPHOS was the top-scoring pathway in most tumor clusters (**Figure 2G**). Similarly, glycolysis also made a major contribution to the metabolic heterogeneity of several tumor clusters, indicating that energy metabolic factors (mitochondrial activity) are important contributors to intra-tissue metabolic heterogeneity.

Metabolic Features During T-Cell Differentiation

In the TME, immune cells that differentiate into subtypes with distinct roles constitute an important component. Next, we used single-cell RNA-seq profile to characterize the developmental trajectory and metabolic features of T cells, which constitute the major immune cell population. According to previous studies on T cells, T cells were first separated into CD4+ and CD8+ subtypes based on the expression of the cell-surface markers CD4 and CD8A (O'Shea and Paul, 2010) (**Figure 3A**). CD4+ T cells were further differentiated into regulatory T cells (Tregs) and T helper cells (Ths) based on the expression of FOXP3 and CD25 (Michalek et al., 2011) (**Figure 3A**). We defined T cells with CD8A expression less than 1 and CD4 expression more than 1 as CD4+ T cells (45 CD4+ T cells), and cells with the opposite expression status were defined as CD8+ T cells (60 CD8+ T cells; **Figure 3B**; **Supplementary Figure 2**). Furthermore, CD4+ T cells with a sum of FOXP3 and CD25 expression higher than 2 were identified as Tregs (7 Tregs), and those cells that did not express FOXP3 and CD25 were defined as Ths (26 Ths; **Figure 3B**; **Supplementary Figure 2**). The R package monocle was then used to emulate the pseudo-developmental trajectory of the T-cell subpopulations. Three branches (defined as "B1," "B2," and "B3") were found in the developmental trajectory of T cells, which were divided into three states (**Figure 3C**). Based on the visualization results, CD4+ T cells, including Tregs and Ths, were mainly concentrated on state 2 and CD8+ T cells were mainly concentrated on state 3 (**Figure 3D**). Based on the above information, we concluded that the cell development from B1 through branch point 1 to state 3 correspond to the conversion of T cells to CD4+ T cells, and the cell development from B1 through branch point 1 to state 2 correspond to the conversion of T cells to CD8+ T cells. We identified multiple branch-dependent genes for branching point 1 ($p < 0.01$), which were closely associated with T-cell differentiation (**Figure 3E**).

T-cell differentiation was accompanied by reprogramming of metabolic pathways to satisfy the physiological demands of the new cellular state (DeBerardinis and Chandel, 2016; Xiao et al., 2019). We then performed GSEA analysis to identify metabolic pathways enriched in highly variable genes of each subtype. We found that OXPHOS and glycolysis had the top normalized functional enrichment scores (NESs) in CD4+ T and CD8+ T cells, suggesting that mitochondrial activity is also a major contributor to metabolic heterogeneity among T cells (**Figure 3F**). Glutathione metabolism and purine metabolism were found to be important metabolic pathways that distinguished the T-cell subtypes (**Figure 3F**). Interestingly, OXPHOS and glycolysis also had the highest NES in Tregs and Ths cells (**Figure 3G**). We found that the rates of glycolysis were upregulated in Tregs compared to Ths cells (**Supplementary Figure 3**), which seems to contradict previous studies showing that Ths are more susceptible to glycolysis than Tregs derived from healthy mice without tumors (Michalek et al., 2011). In contrast, the preference for OXPHOS in Tregs was consistent with previous studies (Ma et al., 2016; Angelin et al., 2017; Buck et al., 2017), highlighting that enhanced mitochondrial oxidative

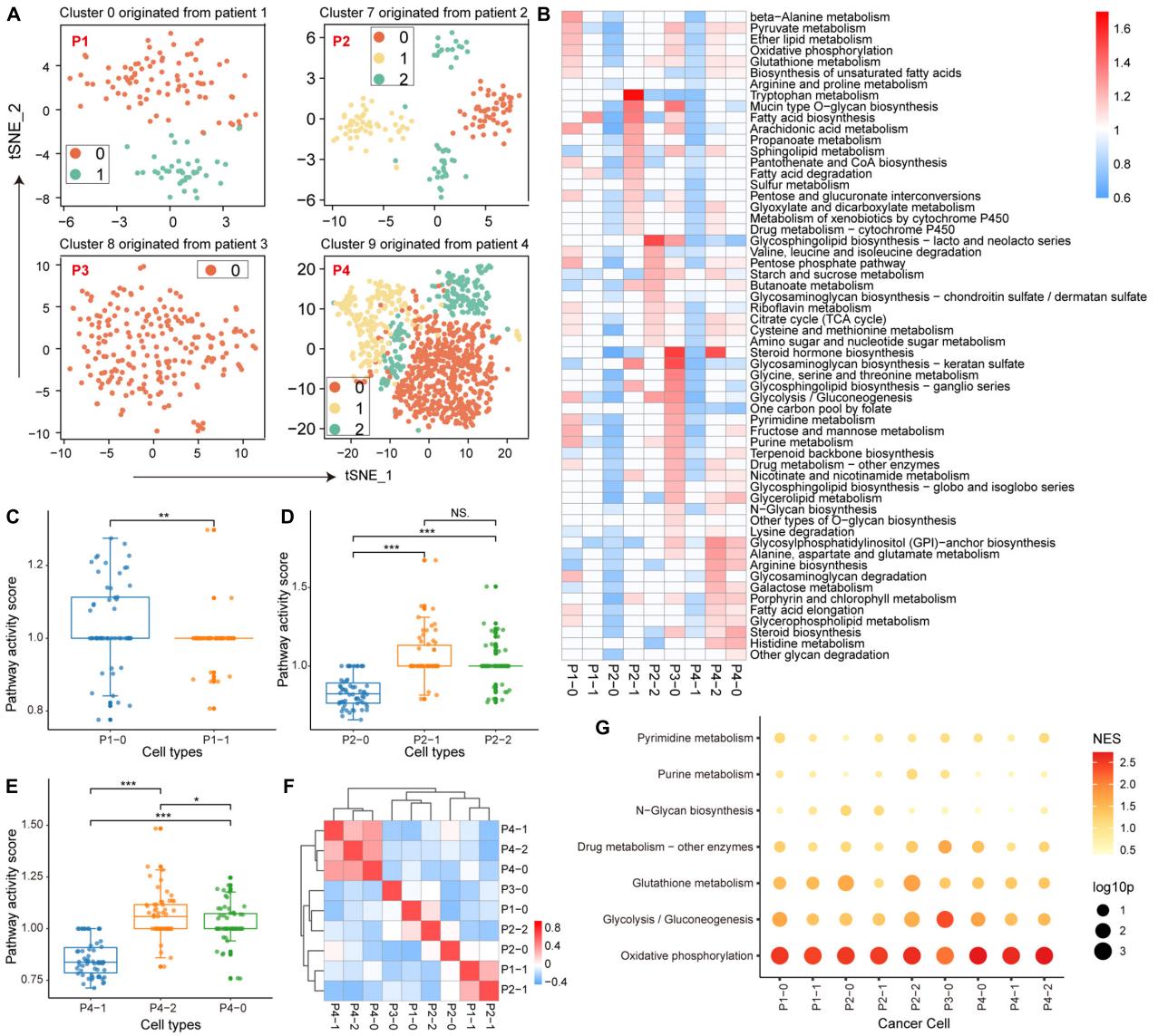


FIGURE 2 | Metabolic features of malignant epithelial cells. **(A)** Four clusters of malignant epithelial cells originating from four patients with non-small cell lung cancer (NSCLC) were individually re-clustered using t-SNE. **(B)** Metabolic pathway activity in each cell cluster. Values with statistically non-significant pathway activity (random permutation test, $p > 0.05$) were shown as blank. **(C–E)** Boxplot of pathway activity scores of malignant cell clusters originating from patients 1, 2, and 4. **(F)** Correlation of each malignant cell cluster in metabolic pathway activity scores. Additionally, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. **(G)** Enrichment results of metabolic pathways in variant genes using gene set enrichment analysis (GSEA). The size of the bubbles represents statistical significance and the shade of the color represents the normalized enrichment score (NES).

metabolism is a characteristic of Tregs. These results suggest that the metabolic features of T cells in the TME differ from those of normal tissues.

Key Factors in T-Cell Differentiation Correlate With Tumor Metastasis and Patient Prognosis

T cells in tumor tissues have specific metabolic and physiological mechanisms that are influenced by the TME. To explore the effects of genes associated with T-cell differentiation on cellular physiological functions, Pearson correlation analysis was used to identify genes associated with the T-cell differentiation process. We identified 308 genes (216 positive and 92 negative related genes) and 284 genes (187 positive and 92 negative related genes) that were associated with T/CD4+ differentiation and T/CD8+ differentiation, respectively. The R package clusterprofiler was used for GO functional enrichment and KEGG pathway analysis of genes related to T-cell differentiation. We found that genes upregulated in T/CD4+ differentiation were significantly enriched in functional pathways associated with antigen processing (**Figure 4A** and **Supplementary Figure 4A**) and

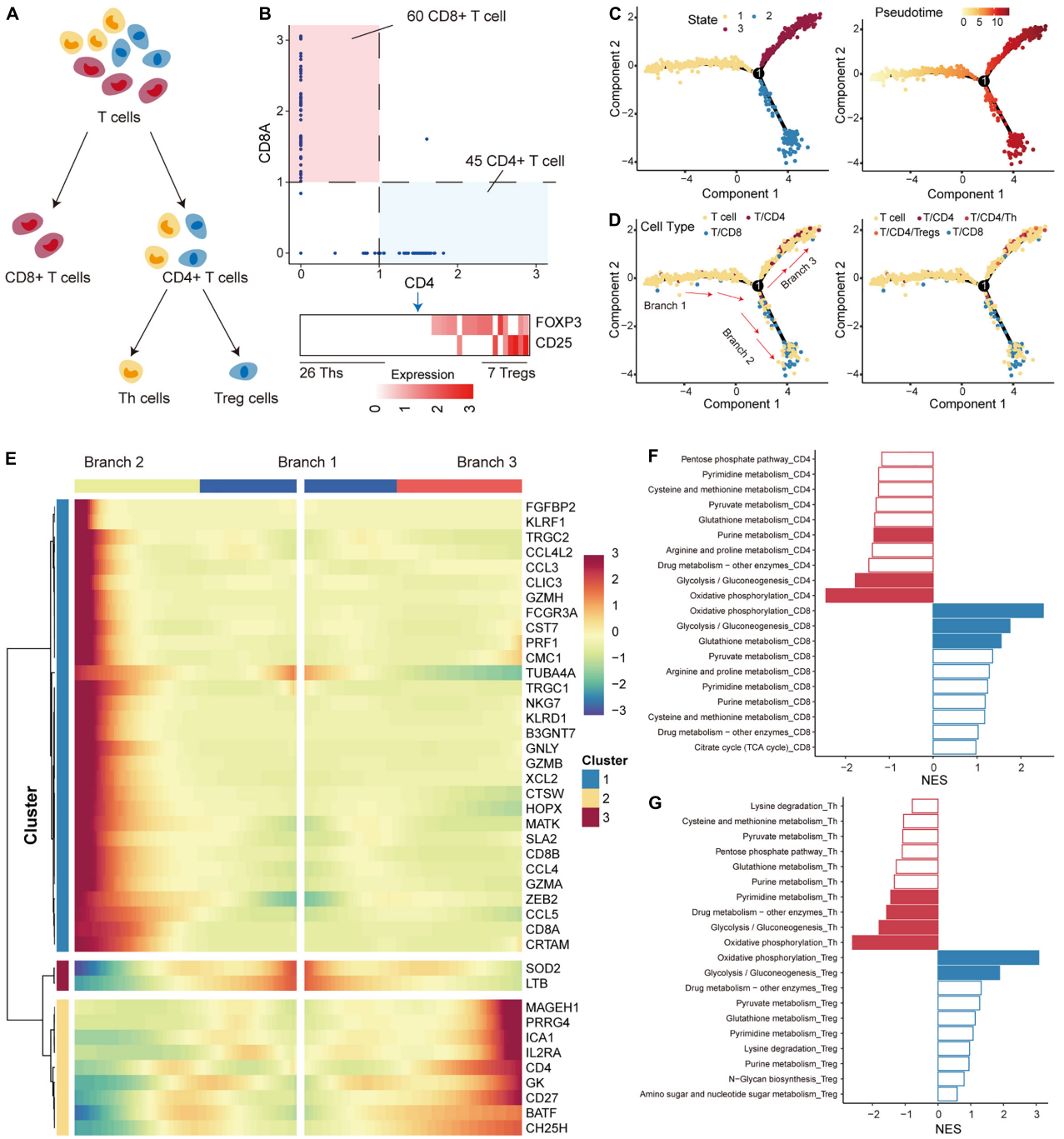


FIGURE 3 | Differentiation trajectory and metabolic features of T cells. **(A)** The panel shows the differentiation process of T cells; T cells were divided into CD4+ T, CD8+ T, Th, and Treg cells. **(B)** Expression levels of marker genes (including CD4, CD8A, FOXP3, and CD25) were used to separate T-cell subtypes in non-small cell lung cancer (NSCLC). **(C)** The pseudo-time trajectory of T-cell differentiation. Each point represents a cell and is marked with the cell state (left) and pseudo-time (right). **(D)** Same as in **(C)** but for cells marked with T-cell subtypes. **(E)** The heatmap shows the branch-dependent genes at branch point 1. The center of the heatmap is branch B1, the left is B2, and the right is B3. **(F)** Top 10 metabolic pathways enriched in CD4+ or CD8+ T cells. Significantly enriched pathways with gene set enrichment analysis (GSEA) p-values < 0.05 are highlighted in red (CD4) or blue (CD8). **(G)** Top 10 metabolic pathways enriched in Th or Treg cells. Significantly enriched pathways with GSEA p-values < 0.05 are highlighted in red (Th) or blue (Treg).

presentation, and genes upregulated in T/CD8+ differentiation were significantly enriched in the regulation of cell killing and the immune functions in which neutrophils are involved (**Figure 4B** and **Supplementary Figure 4B**). These findings are consistent with previous studies revealing the function of CD4+ and CD8+ T cells (Imanishi and Saito, 2020), which supports the reliability

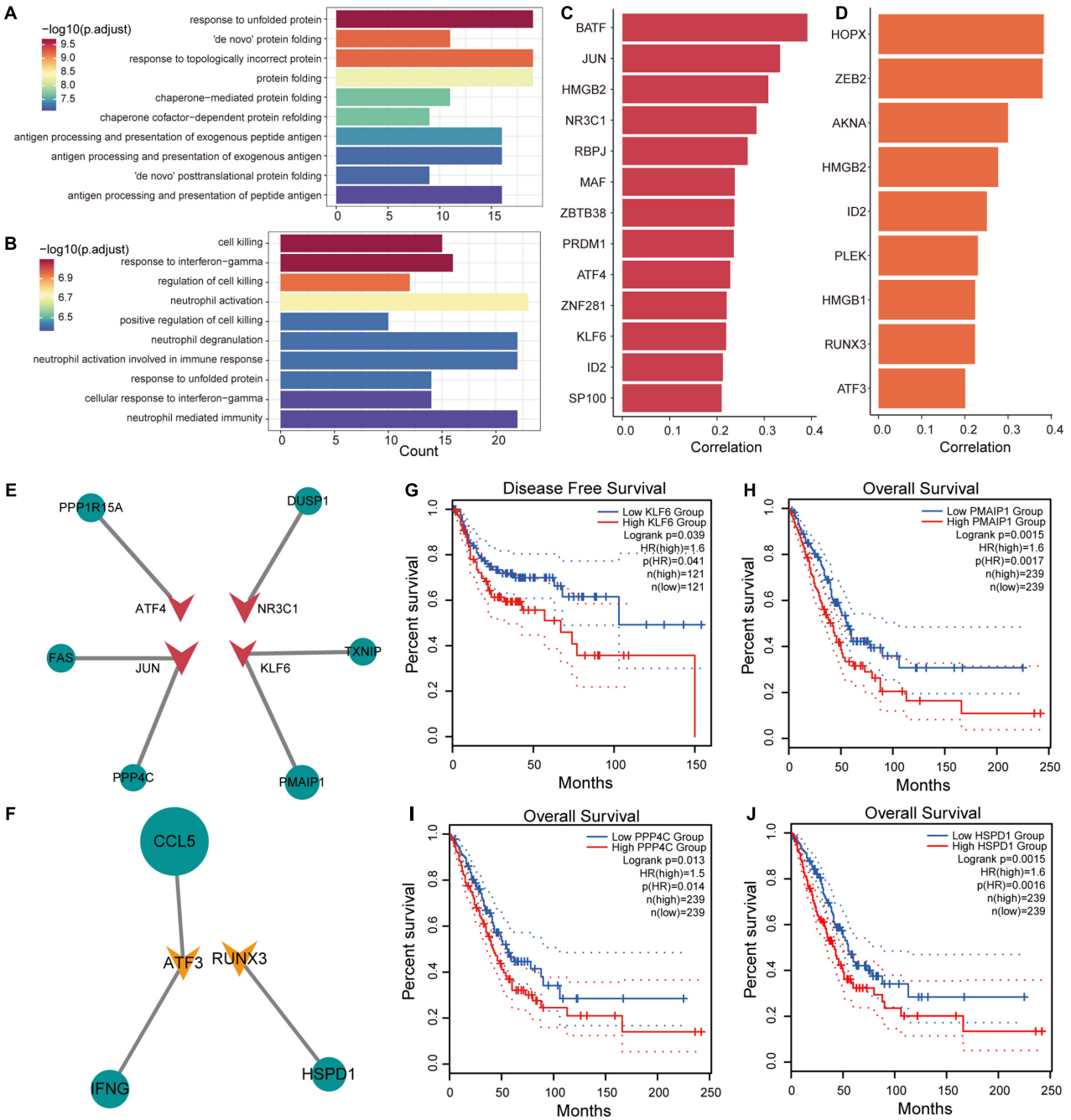


FIGURE 4 | Transcriptional regulation of genes associated with T-cell differentiation. **(A)** Enrichment results of upregulated genes associated with T/CD4+ differentiation in the gene ontology (GO) term. **(B)** Same as in **(A)** but for genes associated with T/CD8+ differentiation. **(C)** Transcription factor (TF) lists selected from genes positively associated with T/CD4+ differentiation and their Pearson correlation coefficients with pseudo-time. **(D)** Same as in **(C)** but for TF lists selected from genes associated with T/CD8+ differentiation. **(E)** The transcriptional regulatory network of TFs and target genes in genes related to T/CD4+ differentiation. The size of the nodes in the network represents the correlation coefficient, the circle represents the target gene, and the triangle represents the TF. **(F)** Same as in **(E)** but for genes related to T/CD8+ differentiation. **(G)** Survival curves comparing disease-free survival (DFS) of patients with high and low *KLF6* gene expression. **(H–J)** Survival curves comparing overall survival (OS) of patients with high and low *PMAIP1*, *PPP4C*, and *HSPD1* gene expression.

of cell cluster identification and T-cell differentiation trajectory identification in this study.

Identification of key factors that promote T-cell differentiation and the formation of specific metabolic mechanisms for T-cell is of substantial importance for the treatment of NSCLC. We focused on transcription factors (TFs) that regulate the expression efficiency of target genes. Using human TF data collected from AnimalTFDB (Hu et al.,

2019), 13 and 9 TFs were identified from genes associated with T/CD4+ differentiation and T/CD8+ differentiation (**Figures 4C,D**). In combination with other genes associated with T-cell differentiation, we constructed transcriptional regulatory networks associated with T/CD4+ and T/CD8+ differentiation (**Figures 4E,F**), respectively. We found that high expression of TF KLF6 was associated with poor disease-free survival (DFS) in patients with NSCLC and that the high expression of its target gene, *PMAIP1*, was associated with poor OS (**Figures 4G,H**). Previous studies have confirmed that KLF6 is a key TF that participates in the activation of T cells (Palau et al., 2013), suggesting that KLF6 can be used as a potential target for regulating the differentiation of T cells and assisting in immunotherapy. The variation in KLF6 expression was also associated with metastatic potential and poor prognosis of patients with NSCLC (DiFeo et al., 2009; Zhang et al., 2019a). PPP4C is required for extracellular skeleton composition in cell metastasis (Martin-Granados et al., 2008), which was highly expressed in CD4+ T cells and was associated with poor OS (**Figure 4I**). High expression of HSPD1 in CD8+ T cells was strongly associated with poorer patient OS (**Figure 4J**). Moreover, high expression of HSPD1 was shown to inhibit E-cadherin expression to promote cell metastasis (Kang et al., 2019). The HSP60 protein encoded by *HSPD1* could be used for exosomal antigen presentation to dendritic cells (DCs) to inhibit the differentiation of CD4+ T cells (Cui et al., 2019), suggesting that inhibiting the differentiation of CD4+ T cells by regulating the expression of KLF6 can be used as a strategy to alleviate immunosuppression. Additionally, we have precisely defined the functions of target genes that are positively associated with CD4+ and CD8+ T-cell differentiation using Metascape (Zhou et al., 2019) (**Supplementary Tables 3, 4**). Taken together, these results suggest that the key genes that drive T-cell differentiation can stimulate tumor invasion and metastasis and could be used as potential therapeutic targets.

Identifying Prognostic Markers for NSCLC in Combination With Public Data

The immune microenvironment of tumors is closely related to tumor development (Lei et al., 2020). As an important component of the immune microenvironment, the dynamics of T cells at the molecular and cellular levels significantly affect tumor development and metastasis. It is intriguing to identify the markers associated with the prognosis of patients with NSCLC from genes related to T-cell differentiation. Next, we used lasso regression to screen for genes that are strongly associated with patient prognosis of LUAD, and 8 genes associated with patient prognosis were identified (**Figure 5A**). The multivariate Cox regression models were used to fit these eight feature genes, four of which, *HERPUD1*, *MAP3K8*, *GAPDH*, and *DNAJB4*, were significantly associated with the risk of death in patients ($p < 0.05$; **Supplementary Figure 5**). Nomograms were used to predict the probability of death at 1, 2, and 4 years (**Figure 5B**). The results of the calibration curve showed the strong stability of the risk prediction model (**Figure 5C**). To identify the best predictive time points for the risk prediction model, we split the 4-year period into six time periods and evaluated the prediction results using the ROC curve. We found that the risk prediction results reached a maximum AUC value of 0.71, in the fourth year (1460 days; **Figure 5D**). Furthermore, we used regression coefficients for HERPUD1, MAP3K8, GAPDH, and DNAJB4 to construct risk score models as follows: risk score = −0.197*MAP3K8 −0.261*HERPUD1+ 0.185*GAPDH+ 0.191*DNAJB4 and calculated the risk score for each LUAD tumor sample. The samples were divided into two categories (high-risk and low-risk) based on the median of risk scores, and we found that the high-risk samples were associated with poorer OS in LUAD (**Figure 5E**). The four prognostic markers, HERPUD1, MAP3K8, GAPDH, and DNAJB4, have also been used to predict survival in patients with LUSC. In LUSC, the high-risk samples also showed poorer OS (**Figure 5F**). Additionally, we combined the four prognostic markers identified and samples of NSCLC obtained from the GEO database to reconstruct the risk score model. We found that high-risk scores were significantly associated with poorer survival, which was similar to previous predictions (**Supplementary Figure 6**). These results suggest that HERPUD1, MAP3K8, GAPDH, and DNAJB4 can be used as prognostic markers in NSCLC. By combining clinical information from the LUAD sample with the risk score, we found that patients with stage IV cancer had a significantly higher risk score (**Figure 5G**). Although smoking could increase the risk of lung cancer (Bade and Dela Cruz, 2020), the history of smoking does not show a positive correlation with the prognostic risk (**Figure 5H**), indicating that specific tobacco tolerance is caused by individual genetic differences.

DISCUSSION

In this study, we used a computational pipeline to reveal the biological information contained in the single-cell RNA profiles of NSCLC. By clustering cells from the tumor tissues of four patients, individual differences in the metabolic landscapes of malignant epithelial cells were revealed. We characterized the metabolic characteristics of the malignant epithelial cells. GSEA revealed that OXPHOS and glycolysis are the major contributors to the intra-tissue metabolic heterogeneity of malignant cells. We constructed the differentiation trajectory of T cells using the monocle tool and revealed that T-cell subtypes have different metabolic features from normal tissues to adapt to the TME. In the process of T-cell differentiation, we found that the key genes that drive T-cell differentiation could serve as potential therapeutic targets. Finally, we constructed a survival risk prediction model and identified HERPUD1, MAP3K8, GAPDH, and DNAJB4 as prognostic markers for NSCLC.

The TME includes not only the tumor cells but also the surrounding fibroblasts, immune and inflammatory cells, glial cells, and other cells (Arneth, 2019; Hinshaw and Shevde, 2019). Tumor proliferation and metastasis are drove by malignant cells and other stromal cells. Therefore, we focused on malignant epithelial cells and T cells, which are important components of immune cells. Although numerous studies have

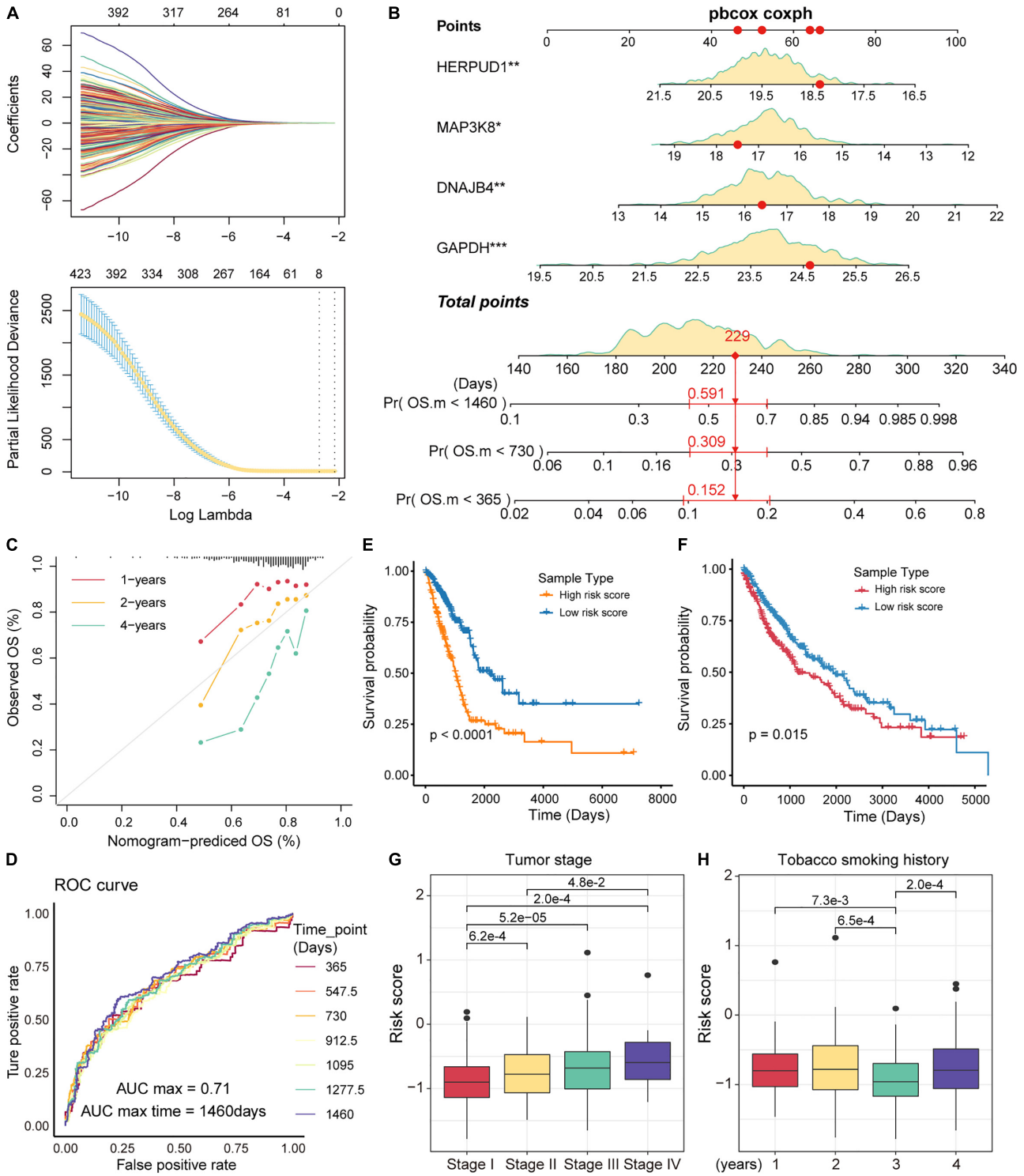


FIGURE 5 | Identification of prognostic markers in non-small cell lung cancer (NSCLC). **(A)** The curves reflect the relationship between the regression coefficient and λ value. The features were selected by the lasso regression model. **(B)** Nomogram for survival risk prediction of 1, 2, and 4 years. The model contains four feature genes (prognostic markers). **(C)** Calibration curve of the nomogram. **(D)** Receiver operating characteristic (ROC) plots of the outcomes predicted by the risk regression model for seven time points. The different colored curves represent specific time points. **(E)** Kaplan-Meier (KM) curves for the survival time (days) of lung adenocarcinoma (LUAD) samples in high- and low-risk categories. The log-rank test was used to calculate statistical significance. **(F)** Same as in **(E)** but for the survival of lung squamous cell carcinoma (LUSC) samples. **(G)** Box plots of risk scores for different tumor stage samples. The rank-sum test was used to statistically measure differences between groups. **(H)** Same as in **(G)** but for samples with different smoking histories.

explored the features of immune cells and signaling pathways in the NSCLC microenvironment at a single-cell resolution (Guo et al., 2018; Clarke et al., 2019; Maynard et al., 2020), few studies have delved into the metabolic programming of malignant cells and the driving mechanism of T-cell differentiation (Kim et al., 2020). In this study, we characterized the metabolic characteristics of malignant cells and T-cell subtypes. Considering that numerous genes appear repeatedly in multiple metabolic pathways, we did not use traditional gene set variation analysis (GAVA) (Hanzelmann et al., 2013) to calculate metabolic pathway activity. The weighted relative pathway activity algorithm can highlight the activity features of each pathway to avoid the effect of under- and over-expression of overlapping genes in multiple pathways.

Tumors are formed when normal cells continue to proliferate in an uncontrolled manner due to the loss of cell cycle control (Radomska et al., 2019). Tumors develop different directions of mutations and malignant evolution due to different cancer-causing agents. Therefore, we aimed to identify the individual differences in the metabolic activity of malignant epithelial cells. Malignant cells are influenced by their tissue environment and show significant heterogeneity in mitochondrial activity. The differentiation process of T cells is accompanied by metabolic reprogramming. We found that Tregs have higher glycolytic activity than Ths cells in tumor tissues, which is opposite to that in normal tissues. Tregs play a crucial role in maintaining immune tolerance, and their abnormal expression can lead to autoimmune diseases (Barbi et al., 2014). The reprogramming of energy metabolism in Treg cells may be an important contributor to immune dysfunction in tumor tissues. We found that genes related to T-cell differentiation can serve as prognostic markers for patients with NSCLC, which may be due to the remodeling of T-cell molecular mechanisms that affect tumor proliferation and metastasis.

In conclusion, this study describes the metabolic and immune landscapes of NSCLC at the single-cell level. Marker genes associated with patient prognosis were identified through the construction of survival risk models. Although we have only analyzed NSCLC in terms of metabolic pathways and T-cell differentiation, this study provides insights into tumor proliferation and prognosis of patients with NSCLC. The findings of this study may provide theoretical guidance for the diagnosis and treatment of NSCLC.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

PW, FW, and YZ conceived and designed the experiments. FW, YZ, and YH analyzed the data. XL and YQ collected the data. FW, MX, and QX validated the methods and data. FW and YZ wrote the manuscript. All authors have read and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (32070622), Natural Science Foundation of Heilongjiang Province (LH2019H042), Health Commission Foundation of Heilongjiang Province (2018241), Top Young Talent Project of Harbin Medical University Cancer Hospital (BJQN2020-02), and Nn10 program of Harbin Medical University Cancer Hospital (Nn10py2017-04).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.702112/full#supplementary-material

REFERENCES

Angelin, A., Gil-de-Gomez, L., Dahiya, S., Jiao, J., Guo, L., Levine, M. H., et al. (2017). Foxp3 reprograms T cell metabolism to function in low-glucose, high-lactate environments. *Cell Metab.* 25, 1282–1293. doi: 10.1016/j.cmet.2016.12.018

Arneth, B. (2019). Tumor microenvironment. *Medicina* 56:15.

Bade, B. C., and Dela Cruz, C. S. (2020). Lung cancer 2020: epidemiology, etiology, and prevention. *Clin. Chest Med.* 41, 1–24.

Barbi, J., Pardoll, D., and Pan, F. (2014). Treg functional stability and its responsiveness to the microenvironment. *Immunol. Rev.* 259, 115–139. doi: 10.1111/imr.12172

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995.

Belle, J. I., and DeNardo, D. G. (2019). A Single-Cell Window into Pancreas Cancer Fibroblast Heterogeneity. *Cancer Discov* 9, 1001–1002. doi: 10.1158/2159-8290.cd-19-0576

Bian, X., Xiao, Y. T., Wu, T., Yao, M., Du, L., Ren, S., et al. (2019). Microvesicles and chemokines in tumor microenvironment: mediators of intercellular communications in tumor progression. *Mol. Cancer* 18:50.

Boroughs, L. K., and DeBerardinis, R. J. (2015). Metabolic pathways promoting cancer cell survival and growth. *Nat. Cell Biol.* 17, 351–359. doi: 10.1038/ncb3124

Buck, M. D., Sowell, R. T., Kaech, S. M., and Pearce, E. L. (2017). Metabolic instruction of immunity. *Cell* 169, 570–586. doi: 10.1016/j.cell.2017.04.004

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550. doi: 10.1038/nature13385

Carmona-Fontaine, C., Deforet, M., Akkari, L., Thompson, C. B., Joyce, J. A., and Xavier, J. B. (2017). Metabolic origins of spatial organization in the tumor microenvironment. *Proc. Natl. Acad. Sci. U.S.A.* 114, 2934–2939. doi: 10.1073/pnas.1700600114

Clarke, J., Panwar, B., Madrigal, A., Singh, D., Gujar, R., Wood, O., et al. (2019). Single-cell transcriptomic analysis of tissue-resident memory T cells in human lung cancer. *J. Exp. Med.* 216, 2128–2149.

Cui, X., Liu, Y., Wang, S., Zhao, N., Qin, J., Li, Y., et al. (2019). Circulating exosomes activate dendritic cells and induce unbalanced CD4+ T cell differentiation in hashimoto thyroiditis. *J. Clin. Endocrinol. Metab.* 104, 4607–4618. doi: 10.1210/jc.2019-00273

DeBerardinis, R. J., and Chandel, N. S. (2016). Fundamentals of cancer metabolism. *Sci. Adv.* 2:e1600200. doi: 10.1126/sciadv.1600200

DiFeo, A., Martignetti, J. A., and Narla, G. (2009). The role of KLF6 and its splice variants in cancer therapy. *Drug Resist. Updat.* 12, 1–7. doi: 10.1016/j.drup.2008.11.001

Goldstraw, P., Ball, D., Jett, J. R., Le Chevalier, T., Lim, E., Nicholson, A. G., et al. (2011). Non-small-cell lung cancer. *Lancet* 378, 1727–1740.

Gomes-Silva, D., Srinivasan, M., Sharma, S., Lee, C. M., Wagner, D. L., Davis, T. H., et al. (2017). CD7-edited T cells expressing a CD7-specific CAR for the therapy of T-cell malignancies. *Blood* 130, 285–296. doi: 10.1182/blood-2017-01-761320

Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., et al. (2018). Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* 24, 978–985.

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 46, D380–D386.

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

Hellmann, M. D., Paz-Ares, L., Bernabe Caro, R., Zurawski, B., Kim, S. W., Carcereny Costa, E., et al. (2019). Nivolumab plus ipilimumab in advanced non-small-cell lung cancer. *N. Engl. J. Med.* 381, 2020–2031.

Hinshaw, D. C., and Shevde, L. A. (2019). The tumor microenvironment innately modulates cancer progression. *Cancer Res.* 79, 4557–4566. doi: 10.1158/0008-5472.can-18-3962

Hu, H., Miao, Y. R., Jia, L. H., Yu, Q. Y., Zhang, Q., and Guo, A. Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.* 47, D33–D38.

Huang, A. C., Postow, M. A., Orlowski, R. J., Mick, R., Bengsch, B., Manne, S., et al. (2017). T-cell invigoration to tumour burden ratio associated with anti-PD-1 response. *Nature* 545, 60–65. doi: 10.1038/nature22079

Imanishi, T., and Saito, T. (2020). T Cell Co-stimulation and functional modulation by innate signals. *Trends Immunol.* 41, 200–212. doi: 10.1016/j.it.2020.01.003

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361.

Kang, B. H., Shu, C. W., Chao, J. K., Lee, C. H., Fu, T. Y., Liou, H. H., et al. (2019). HSPD1 repressed E-cadherin expression to promote cell invasion and migration for poor prognosis in oral squamous cell carcinoma. *Sci. Rep.* 9:8932.

Kim, N., Kim, H. K., Lee, K., Hong, Y., Cho, J. H., Choi, J. W., et al. (2020). Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* 11:2285.

Lei, X., Lei, Y., Li, J. K., Du, W. X., Li, R. G., Yang, J., et al. (2020). Immune cells within the tumor microenvironment: biological functions and roles in cancer immunotherapy. *Cancer Lett.* 470, 126–133. doi: 10.1016/j.canlet.2019.11.009

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Liu, X., Cooper, D. E., Cluntun, A. A., Warmoes, M. O., Zhao, S., Reid, M. A., et al. (2018). Acetate production from glucose and coupling to mitochondrial metabolism in mammals. *Cell* 175, 502–513. doi: 10.1016/j.cell.2018.08.040

Lun, A. T., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* 5:2122. doi: 10.12688/f1000research.9501.2

Ma, C., Kesarwala, A. H., Eggert, T., Medina-Echeverz, J., Kleiner, D. E., Jin, P., et al. (2016). NAFLD causes selective CD4(+) T lymphocyte loss and promotes hepatocarcinogenesis. *Nature* 531, 253–257. doi: 10.1038/nature16969

Maimela, N. R., Liu, S., and Zhang, Y. (2019). Fates of CD8+ T cells in tumor microenvironment. *Comput. Struct. Biotechnol. J.* 17, 1–13. doi: 10.1016/j.csbj.2018.11.004

Martin-Granados, C., Philp, A., Oxenham, S. K., Prescott, A. R., and Cohen, P. T. (2008). Depletion of protein phosphatase 4 in human cells reveals essential roles in centrosome maturation, cell migration and the regulation of Rho GTPases. *Int. J. Biochem. Cell Biol.* 40, 2315–2332. doi: 10.1016/j.biocel.2008.03.021

Maynard, A., McCoach, C. E., Rotow, J. K., Harris, L., Haderk, F., Kerr, D. L., et al. (2020). Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing. *Cell* 182, 1232–1251.

Michalek, R. D., Gerriets, V. A., Jacobs, S. R., Macintyre, A. N., MacIver, N. J., Mason, E. F., et al. (2011). Cutting edge: distinct glycolytic and lipid oxidative metabolic programs are essential for effector and regulatory CD4+ T cell subsets. *J. Immunol.* 186, 3299–3303. doi: 10.4049/jimmunol.1003613

Miller, A., Nagy, C., Knapp, B., Laengle, J., Ponweiser, E., Groeger, M., et al. (2017). Exploring metabolic configurations of single cells within complex tissue microenvironments. *Cell Metab.* 26, 788–800. doi: 10.1016/j.cmet.2017.08.014

O'Shea, J. J., and Paul, W. E. (2010). Mechanisms underlying lineage commitment and plasticity of helper CD4+ T cells. *Science* 327, 1098–1102. doi: 10.1126/science.1178334

Palau, N., Julia, A., Ferrandiz, C., Puig, L., Fonseca, E., Fernandez, E., et al. (2013). Genome-wide transcriptional analysis of T cell activation reveals differential gene expression associated with psoriasis. *BMC Genomics* 14:825. doi: 10.1186/1471-2164-14-825

Pavlova, N. N., and Thompson, C. B. (2016). The emerging hallmarks of cancer metabolism. *Cell Metab.* 23, 27–47. doi: 10.1016/j.cmet.2015.12.006

Png, Y. T., Vinanica, N., Kamiya, T., Shimasaki, N., Coustan-Smith, E., and Campana, D. (2017). Blockade of CD7 expression in T cells for effective chimeric antigen receptor targeting of T-cell malignancies. *Blood Adv.* 1, 2348–2360. doi: 10.1182/bloodadvances.2017009928

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171, 1611–1624. doi: 10.1016/j.cell.2017.10.044

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14, 979–982. doi: 10.1038/nmeth.4402

Radomska, K. J., Coulpier, F., Gresset, A., Schmitt, A., Debbiche, A., Lemoine, S., et al. (2019). Cellular origin, tumor progression, and pathogenic mechanisms of cutaneous neurofibromas revealed by mice with Nf1 knockout in boundary cap cells. *Cancer Discov.* 9, 130–147. doi: 10.1158/2159-8290.cd-18-0156

Reina-Campos, M., Moscat, J., and Diaz-Meco, M. (2017). Metabolism shapes the tumor microenvironment. *Curr. Opin. Cell Biol.* 48, 47–53. doi: 10.1016/j.ceb.2017.05.006

Reznik, E., Luna, A., Aksoy, B. A., Liu, E. M., La, K., Ostrovnaya, I., et al. (2018). A landscape of metabolic variation across tumor types. *Cell Syst.* 6, 301–313. doi: 10.1016/j.cels.2017.12.014

Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., et al. (2015). Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348, 124–128.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502. doi: 10.1038/nbt.3192

Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., et al. (2015). Single-Cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* 17, 360–372. doi: 10.1016/j.stem.2015.07.013

Song, Q., Hawkins, G. A., Wudel, L., Chou, P. C., Forbes, E., Pullikuth, A. K., et al. (2019). Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. *Cancer Med.* 8, 3072–3085. doi: 10.1002/cam4.2113

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., et al. (2015). Model of human colorectal tumor growth. *Nat. Genet.* 47, 209–216.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338.

Topalian, S. L., Hodi, F. S., Brahmer, J. R., Gettinger, S. N., Smith, D. C., McDermott, D. F., et al. (2012). Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* 366, 2443–2454.

Topalian, S. L., Taube, J. M., Anders, R. A., and Pardoll, D. M. (2016). Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat. Rev. Cancer* 16, 275–287. doi: 10.1038/nrc.2016.36

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA Cancer J. Clin.* 65, 87–108. doi: 10.3322/caac.21262

Vafaee, F., Krycer, J. R., Ma, X., Burykin, T., James, D. E., and Kuncic, Z. (2016). ORTI: an open-access repository of transcriptional interactions for interrogating mammalian gene expression data. *PLoS One* 11:e0164535. doi: 10.1371/journal.pone.0164535

van Meerbeeck, J. P., Fennell, D. A., and De Ruysscher, D. K. (2011). Small-cell lung cancer. *Lancet* 378, 1741–1755.

Xiao, Z., Dai, Z., and Locasale, J. W. (2019). Metabolic landscape of the tumor microenvironment at single cell resolution. *Nat. Commun.* 10:3763.

Yasunaga, M. (2020). Antibody therapeutics and immunoregulation in cancer and autoimmune disease. *Semin. Cancer Biol.* 64, 1–12. doi: 10.1016/j.semcancer.2019.06.001

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). Clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zenatti, P. P., Ribeiro, D., Li, W., Zuurbier, L., Silva, M. C., Paganin, M., et al. (2011). Oncogenic IL7R gain-of-function mutations in childhood T-cell acute lymphoblastic leukemia. *Nat. Genet.* 43, 932–939.

Zhang, N., Yan, Q. Q., Lu, L., Shao, J. B., and Sun, Z. G. (2019a). The KLF6 splice variant KLF6-SV1 promotes proliferation and invasion of non-small cell lung cancer by up-regultating PI3K-AKT signaling pathway. *J. Cancer* 10, 5324–5331. doi: 10.7150/jca.34212

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., et al. (2019b). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 47, D721–D728.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Wang, Zhang, Hao, Li, Qi, Xin, Xiao and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 06 July 2021
doi: 10.3389/fcell.2021.700061

DeepGP: An Integrated Deep Learning Method for Endocrine Disease Gene Prediction Using Omics Data

*Ningyi Zhang[1], Haoyan Wang[1], Chen Xu[2], Liyuan Zhang[1] and Tianyi Zang[1]**

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] Center for Bioinformatics, Harbin Institute of Technology, Harbin, China

Endocrinology is the study focusing on hormones and their actions. Hormones are known as chemical messengers, released into the blood, that exert functions through receptors to make an influence in the target cell. The capacity of the mammalian organism to perform as a whole unit is made possible based on two principal control mechanisms, the nervous system and the endocrine system. The endocrine system is essential in regulating growth and development, tissue function, metabolism, and reproductive processes. Endocrine diseases such as diabetes mellitus, Grave's disease, polycystic ovary syndrome, and insulin-like growth factor I deficiency (IGFI deficiency) are classical endocrine diseases. Endocrine dysfunction is also an increasing factor of morbidity in cancer and other dangerous diseases in humans. Thus, it is essential to understand the diseases from their genetic level in order to recognize more pathogenic genes and make a great effort in understanding the pathologies of endocrine diseases. In this study, we proposed a deep learning method named DeepGP based on graph convolutional network and convolutional neural network for prioritizing susceptible genes of five endocrine diseases. To test the performance of our method, we performed 10-cross-validations on an integrated reported dataset; DeepGP obtained a performance of the area under the curve of ∼83% and area under the precision-recall curve of ∼65%. We found that type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM) share most of their associated genes; therefore, we should pay more attention to the rest of the genes related to T1DM and T2DM, respectively, which could help in understanding the pathogenesis and pathologies of these diseases.

Keywords: endocrine disease, Graves' disease, T2DM, PCOS, T1DM, IGF-I, deep learning methods

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Tao Huang,
Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, China
Xiaofang Zhao,
Institute of Computing Technology, Chinese Academy of Sciences (CAS), China

***Correspondence:**
Tianyi Zang
tianyi.zang@hit.edu.cn

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *25 April 2021*
Accepted: *31 May 2021*
Published: *06 July 2021*

Citation:
Zhang N, Wang H, Xu C, Zhang L and Zang T (2021) DeepGP: An Integrated Deep Learning Method for Endocrine Disease Gene Prediction Using Omics Data. Front. Cell Dev. Biol. 9:700061. doi: 10.3389/fcell.2021.700061

INTRODUCTION

Endocrine diseases fall into broad categories of hormone over- or underproduction, modulate tissue response to hormones, or tumors caused by endocrine tissue (Belfiore and LeRoith, 2018). Hormones synthesized and released by the endocrine glands exert their functions by regulating the biological process of cells. There are several examples of common endocrine diseases: type I/II diabetes mellitus, Graves' disease (GD), polycystic ovary syndrome (PCOS), and insulin-like growth factor I (IGFI) deficiency, etc. To date, genome-wide association studies (GWAS) have reported numerous gene regions associated with different endocrine diseases. The aim of GWAS

analysis is to determine how the combined allele frequency of multiple susceptibility genes can affect autoimmunity and/or disease risk.

Graves' disease is an organ-specific autoimmune thyroid disease, resulting from excessive secretion of thyroid hormones by thyroid tissue (Dvornikova et al., 2020). The pathogenesis of GD is mediated by the production of antibodies to TSH receptors, which provide increased secretion of thyroid hormones and a rapid growth of the thyroid after stimulation (Smith et al., 2018; Soh and Aw, 2019). Since, GD is a hereditary and polygenic transmission disease (Perricone and Shoenfeld, 2019). It has been identified that associations between CTLA-4, FOXP3, TLR class polymorphism, and a number of pathological conditions develop in GD (Xiao et al., 2015; Fathima et al., 2019).

Diabetes mellitus, such as type 1 diabetes mellitus (TIDM) and type 2 diabetes mellitus (T2DM) are also a typical group of endocrine diseases. But basic pathogenic differences exist in these two types of diabetes mellitus. T1DM is immune mediated while T2DM is mediated by metabolic mechanisms (Eizirik et al., 2020). Glucagon secretion is observed to be reduced in patients with T1DM, with an increasing risk of insulin-induced hypoglycemia, but it is enhanced in T2DM, exacerbating the effects of reduced insulin release and action on glucose of blood levels (Gromada et al., 2018). Recent studies have detected several novel and promising TDM-susceptible genes, such as GCKR, SLC30A8, TLR4, and FTO (Ehrmann et al., 1999; Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, 2004; Day et al., 2015).

Polycystic ovary syndrome is a common endocrinopathy among women, with symptoms including irregular menstrual cycles, hyperandrogenism and polycystic ovarian morphology (Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, 2004). It is also accompanied with obesity and insulin resistance, increasing the risk of diabetes, metabolic syndrome, and other cardiovascular diseases (Ehrmann et al., 1999). There are several published causal variants from GWAS studies associated with PCOS (Day et al., 2015; Hayes et al., 2015). Despite the detrimental impact of the disorder on women's health, the etiology remains poorly understood.

Insulin-like growth factor I production is mainly mediated by growth hormones (GH); both GH and IGF-I have an anabolic effect on skeletal muscle and bone. Despite the effects on growing process in childhood, GH also plays an important role in the regulation of metabolism, body composition, and mood, which persist into adult life (Gazzaruso et al., 2014). However, low serum levels of IGF-I have been detected in patients suffering from chronic liver disease and malnutrition despite normal or elevated GH secretion (Cuneo et al., 1995).

Though there have been numerous susceptible loci identified by GWAS that are associated with endocrine diseases, etiology and pathology is still unclear. Analogous to other complex traits, common susceptibility loci identified by GWAS account for only a small proportion of the genetic heritability of the traits. As GWAS were designed to detect the common allelic variants with a minor allele frequencies of 2 to 5%, the variants occurred less frequently but with greater effect sizes being ignored which may account for the observed deficit in heritability (Manolio et al., 2009). Since SNPs detected by GWAS may not be the real causative regions, the SNPs related to them may be the real causative genes of complex diseases due to the theory of linkage disequilibrium (LD). Therefore, we take the expression level of genes regulated by SNPs into account to reduce the impact of LD. eQTL research plays an important role in prioritizing SNP loci in GWAS susceptible regions (Barral et al., 2012). Previous studies usually investigate susceptible genes of complex diseases based on the regulation function of SNPs on gene expression. MR analysis, for example, is proposed to explain the causative effect on a phenotype of gene expression based on the regulation of genetic variants (Freeman et al., 2013). However, it is rarely available in practice to obtain such a large sample size of different types of data, such as phenotype, genome-wide SNP genotype, and gene expression data, to perform a MR analysis. To overcome this, a SMR method was proposed which integrates summary-level data from independent GWAS with eQTL data to identify disease susceptible genes (Zhu et al., 2016).

Machine learning methods have been widely used in prediction problems, such as support vector machine (SVM), network embedding algorithms (such as node2vec), network diffusion algorithms (such as Laplacian heat diffusion, random walk with restart), etc. Combined with multiple biomolecular features, novel biomarkers, genes, and proteins can be predicted (Chen et al., 2018, 2019; Zhang et al., 2020). Nowadays, deep learning methods have also been utilized in bioinformatics. Liu et al. (2017) applied convolutional neural network (CNN) model to identify cell cycle-regulated genes. Graph convolutional network (GCN) was utilized to predict disease-related metabolites in the study by Zhao et al. (2020). Most machine learning and deep learning methods focused on feature extraction and selection. However, no computational method has been developed to predict the susceptible genes of endocrine diseases based on integrated omics data to eliminate the LD disequilibrium bias.

In this study, we developed "DeepGP," a method to prioritize susceptible genes of endocrine diseases based on deep learning approaches. First, we obtained curated disease-gene associations of five endocrine diseases from disGeNET database; susceptible regions and expression level data were downloaded from GWAS catalog and GTEx database, respectively. After mapping the genes to susceptible loci based on position information of genes, the feature vector of each gene was composed of two types of features, a phenotype-based feature derived from GWAS dataset and a transcriptome-based feature derived from eQTL data. Disease similarity network can be obtained from our previous work, which can represent disease features. GCN was then utilized to decipher the integrated feature representations of the gene. Finally, the classification of candidate genes was performed by CNN.

MATERIALS AND METHODS

Work Frame

DeepGP contains three main parts, data preprocessing (feature extraction), feature reconstruction based on GCN, and endocrine disease-related gene prediction based on CNN. In the feature extraction process, we obtained endocrine disease-related gene

information from DisGeNET (Bauer-Mehren et al., 2010), GWAS Catalog (Buniello et al., 2019), and GTEx Portal databases (Carithers and Moore, 2015). After extracting the features of genes and diseases, we built a heterogeneous network composed of genes and diseases. We then utilized the GCN method to reconstruct the integrated gene features to obtain a more precise feature representation of each gene. In the disease gene prediction part, CNN is used to prioritize the causative genes related to endocrine diseases based on a comprehensive feature representation of disease-gene pairs. The workflow of DeepGP is shown in **Figure 1**.

Data Collection and Preprocessing

Genome-wide association studies have identified thousands of genetic variants that are associated with diseases and traits of medical importance in humans. However, the genes detected from the SNPs identified by GWAS which are pathogenic on diseases remain largely unknown due to the complicated LD between SNPs. Intuitively, genes closest to the top associated variants in position are the most likely causative genes. However, there have been studies reporting that causal genes are distinct from the nearest genes (Zhu et al., 2016). Studies have verified that gene expression may be influenced by different genetic variants among genes with different genotypes of the genetic variants, which means the phenotypes can be influenced by genetic variants through regulating the expression of their target genes. Therefore, we used a GCN network embedding method based on two types of omics data to obtain the comprehensive feature vector of each gene.

We first collected endocrine diseases from the "Endocrine diseases" chapter according to the International Classification of Diseases 11th Revision released by the World Health Organization. The major glands of the endocrine system include the pineal gland, pituitary gland, pancreas, ovaries, testes, thyroid gland, parathyroid gland, hypothalamus, and adrenal glands. Considering that some of the endocrine diseases have not been previously widely investigated, we chose five of the endocrine diseases, GD (which is mainly related to thyroid gland), T1DM/T2DM (which is mainly related to pancreas), PCOS (which is mainly related to ovaries), and IGFI deficiency (which is mainly related to pituitary gland). We then collected curated disease genes from the DisGeNET database (Piñero et al., 2016), causative loci from the GWAS database (Buniello et al., 2019), and expression level data from the GTEx database (Carithers and Moore, 2015). In addition, gene–gene interaction network was downloaded from HumanNet v2.0, where the correlation scores between a pair of genes were calculated. We then utilized an R package named biomaRt to obtain detailed information, such as gene location, chromosome number, and start and end position of each gene. BiomaRt can also be used to transform different gene IDs, such as gene symbol, Ensembl ID, and entrez ID, from different databases. After mapping the genes to the susceptible loci identified by GWAS and eQTL, genes with at least one susceptible SNP were kept. Finally, we obtained 7,406 genes, including 4,212 known causal genes and 3,194 candidate genes obtained from HumanNet v2.0.

Therefore, each gene has a feature of 2 × 25D based on GWAS and eQTL summary data according to five diseases. For the phenotype-based feature of each gene, we used the *p*-value of top 5 related SNPs of each disease to indicate the gene feature vector, for genes with less than five related SNPs, the feature vector is filled with 1. Thus, the phenotype-based feature of each gene could be denoted as:

$$G_S^i = [P_S^1 P_S^2, \cdots, P_S^{25}] \tag{1}$$

where P_S^i denotes the *p*-value of SNPs mapped by gene locations. Transcriptome-based feature of each gene can be extracted by the same method:

$$G_S^T = [P_T^1, P_T^2, \cdots, P_T^{25}] \tag{2}$$

where P_T^i denotes the *p*-value of susceptible loci consistent with that from the phenotype-based feature vector of each gene obtained from eQTL data. Thus, an initial integrated 2 × 25D feature vector of each gene is constructed.

Feature Reconstruction by GCN

In this section, we introduced a network-embedding algorithm based on GCN in order to present a new representation of gene features. GCN is a graph deep learning method based on node features and network architecture to classify the nodes

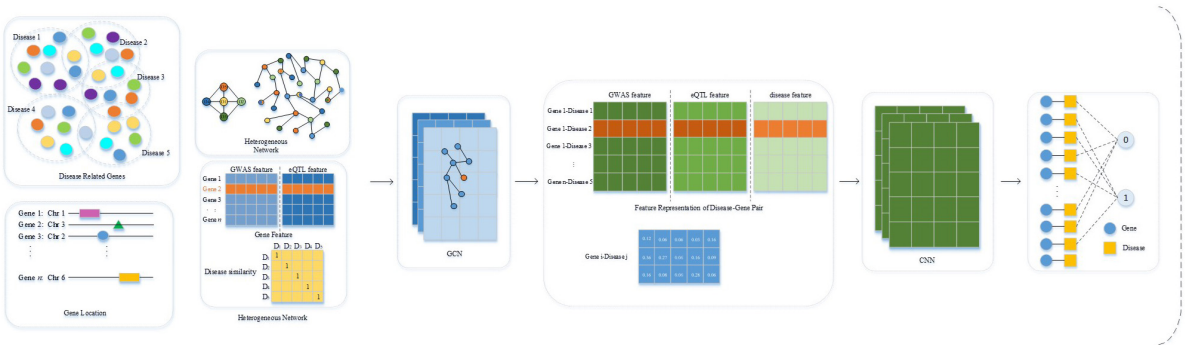


FIGURE 1 | Workflow of DeepGP.

of a network. Although GCNs have been successfully applied in other domains, to our knowledge, this is the first time that GCN was utilized to represent latent gene features from several omics datasets and network properties, while also being capable of disentangling the underlying molecular mechanisms driving the etiology of endocrine diseases.

Considering a graph $G = (V, E, W)$, where V denotes the nodes of the network, E denotes the edges of the network, and W the weight matrix encoding the interacting weight between nodes, which is obtained from HumanNet v2.0 as the gene interactions between gene pairs, the feature matrix of the nodes can be denoted as $X \in R^{N\times F}$. Thus, the eigenvectors of the graph Laplacian L can be denoted as:

$$\mathrm{L} = D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}} \tag{3}$$

where the adjacency matrix $\tilde{A} = A + I$ has added self-connections since gene nodes should contain both gene interaction and gene itself information and D is the degree matrix.

Finally, we can define a propagation rule for each layer:

$$H^{l+1} = (LH^{l}XW^{l}) \tag{4}$$

$$\mathrm{ReLU}\,(x) = \begin{cases} x, & x>0 \\ 0, & x\leq 0 \end{cases} \tag{5}$$

where σ denotes a nonlinearity, such as the Rectified Linear Unit activation function. The input of the first layer is X, which includes the gene expression feature and gene variation feature of five interested traits, so H^0=X. Therefore, the feature of the gene network could be extracted by formula (4). Since we combined two omics types of data, each gene feature could be represented as a 2 × 25D vector. Thus, we have a feature representation of each gene; the feature can be denoted as: $g_i = \begin{bmatrix} P^1_{i,s}, P^2_{i,s}, P^3_{i,s}, \cdots, P^{25}_{i,s} \\ P^1_{i,T}, P^2_{i,T}, P^3_{i,T}, \cdots, P^{25}_{i,T} \end{bmatrix}$.

Causal Gene Prediction With CNN

After obtaining the best combination of initial feature representation by GCN, we constructed disease features based on disease similarity matrix calculated by the method ImpAESim. Each disease feature can be denoted as a 1 × 5D vector:

$$D_i = \{S_{i,j}\}, \mathrm{j} = 1, 2, 3, 4, 5 \tag{6}$$

$S_{i,j}$ denotes the similarity between D_i and D_j, $S_{i,j}$ is 1 if $i = j$. We then combined the gene feature and disease feature as a 3 × 5D feature of disease–gene pair.

We then trained a CNN model to predict causal genes based on the gene features derived from GCN. Analogous to other machine learning methods, CNN consists of a training step where the estimation of network parameters from a given training dataset is learned, and the testing step utilized the well-trained network to predict outputs of new testing dataset (Min et al., 2017). Since our feature format of each gene–disease pair is 3 × 5D, which can be regarded as an image with three channels, in this work, the structure of the CNN section is shown in **Figure 2**. The CNN section includes four parts as follows: convolution layer, max-pooling layer, fully connected layer, and an output layer. Convolution layer is responsible for extracting the subspace features of the input. Max-pooling layer is used for dimension reduction to discard the redundant information. The final fully connected layer connects all the nodes and the output layer applying sigmoid as the activation function to solve the binary classification problem.

The tanh function is the activation function in each convolutional layer.

$$\tanh\,(x) = \frac{\sinh x}{\cosh x} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{7}$$

The sigmoid function is used as the activation function in the output layer.

$$\delta\,(x) = 1/(1 + e^{-x}) \tag{8}$$

Since our disease gene prediction can be treated as a binary classification problem, we chose the binary cross-entropy function as the loss function to assess the probability of the output.

$$\mathrm{loss} = -\sum_{i\ 1}^{n} y_i' \log\,(y_i) + (1 - y_i')\log(1 - y_i') \tag{9}$$

$$\frac{\partial \mathrm{loss}}{\partial y} = -\sum_{i=1}^{n} \frac{y_i'}{y_i} - \frac{1 - y_i'}{1 - y_i'} \tag{10}$$

According to formulas (7, 8), loss is 0 as long as y_i' is equal to y_i .

As a result, each disease–gene pair was assigned a correlation score with a range of [0, 1], where 1 denotes the pair having the strongest association and 0 means no association.

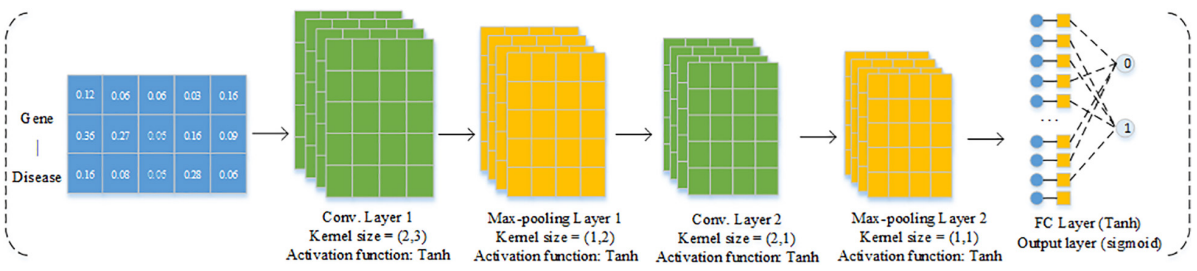


FIGURE 2 | Structure of CNN model.

Training Steps

According to the 4,212 curated disease-related genes to five endocrine diseases, there are 6,258 positive disease–gene pairs derived from the curated disease–gene associations, and 14,802 (4,212 × 5–6,258) pairs are not reported to be associated; we randomly selected 6,258 pairs to construct the negative samples. However, the sample size of different diseases are extremely unbalanced, as shown in **Table 1**; the sample size ranges from 28 IGFI–gene pairs to 3,058 T2D–gene pairs due to the insufficiency of studies related to these diseases, which may have a serious negative impact on the classification performance. From the analysis for summary data derived from GWAS and eQTL, T1D, and T2D share as much as 950 genes in total, with 679/1,629 genes merely related to T1D and 2,108/3,058 genes merely related to T2D.

After obtaining the new dataset that consisted of 6,258 positive samples and 6,258 negative samples, we conducted a 10-cross-validation on this new dataset to test the performance. First, the dataset is randomly divided into 10 groups, then 10 times of iterations were performed based on nine of 10 groups as training set and one group as test set, which made sure that each group can be used as an independent test set.

RESULTS

Performance Evaluation on Predicting Disease–Gene Associations

The area under the curve (AUC) and the area under the precision-recall curve (AUPR) are used to assess DeepGP. The AUC and AUPR of each iteration in 10-cross-validation process are shown in **Table 2**. As a result, DeepGP achieved a mean AUC of 0.845 and a mean AUPR of 0.833, which have shown better and stable in disease–gene prediction.

TABLE 1 | Number of curated disease genes.

Disease	T2D	T1D	PCOS	GD	IGFI deficiency
No. of samples	3,058	1,629	974	568	28

TABLE 2 | AUC and AUPR of DeepGP in 10 times 10 cross-validation.

	1	2	3	4	5	6	7	8	9	10	Average
AUC	0.832	0.845	0.821	0.854	0.864	0.855	0.831	0.861	0.856	0.831	0.845
AUPR	0.827	0.837	0.816	0.845	0.838	0.825	0.816	0.858	0.842	0.826	0.833

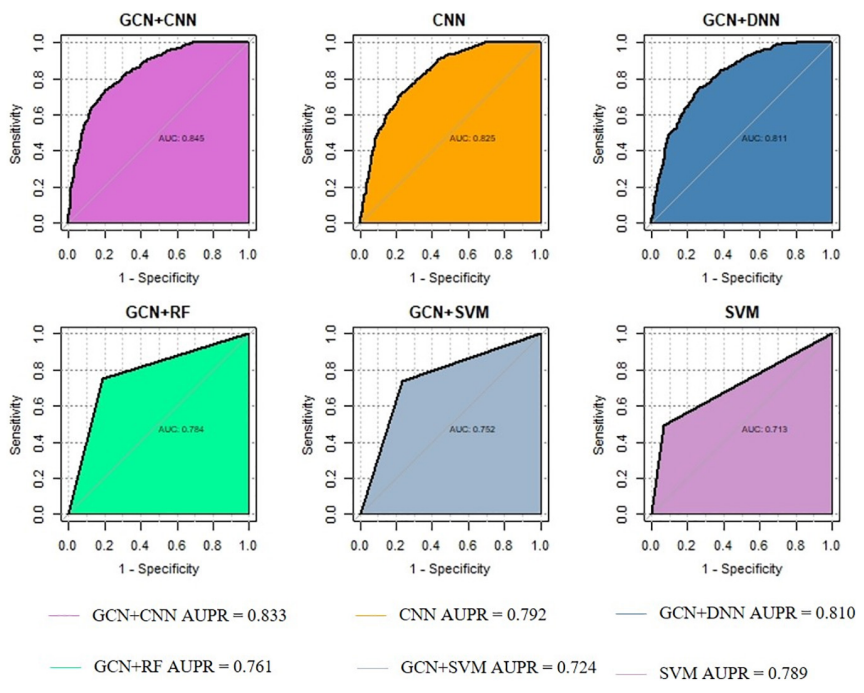


FIGURE 3 | Performance comparison among different methods.

Comparison Experiments With Classic Methods

We evaluated the performance of DeepGP and classic machine learning methods, such as SVM, random forest (RF), Naïve Bayes, and deep neural network (DNN) for predicting disease–gene associations. To validate the performance of extracting best combination of gene features by GCN, we alternatively only used CNN for feature extraction and prediction. Then, to assess the effect of convolution layers in CNN, we applied a typical deep learning method DNN and two classic machine learning methods SVM and RF. As a contrast, we also only used a SVM model for the classification task. Therefore, we compared the performance of five methods with DeepGP: CNN, GCN-DNN, GCN-SVM, GCN-RF, and SVM.

The method of training and testing was performed the same as DeepGP. As shown in **Figure 3**, comparing with other classic machine learning methods, DeepGP achieves the highest performance according to both AUC and AUPR. CNN and DNN achieved the second and third highest AUROC which infers deep learning methods are better than classic machine learning methods in this disease–gene prediction task; however, it can also be inferred that convolution layers are essential. In addition, the performance was improved after feature encoding by GCN.

It has been shown that the depth of CNN models can affect the classification performance. To assess the influence of the depth of CNN models, we compared DeepGP with a shallower CNN model, denoted as GCN+CNN (B), consisting of one set of one convolution layer and one max-pooling layer. Since the feature dimension of each disease–gene pair is 3×5, we also enlarged the kernel size to detect the effect of dimension reduction. The performance is shown in **Figure 4**. GCN+CNN (A) denotes the structure of DeepGP, GCN+CNN (B) denotes the shallower CNN model, and GCN+CNN (C) denotes the structure of CNN with enlarged kernel size. As a result, DeepGP achieved the best performance comparing with other methods even though the feature dimension of each sample is low. Hence, according to the above results, we conclude that our proposed DeepGP is competitive against other methods.

Validation of Prediction Results

After verifying the effectiveness of DeepGP based on the comparison experiments. We conducted the disease–gene prediction process among all the unknown disease–gene pairs. We set a threshold of 0.5 as default to screen the scores and identified 7,702 of 14,903 pairs to be true. Among the predicted associations, 971 novel genes were identified to be associated with GD, 3,142 novel genes associated with PCOS, and 2,437 and 1,154 novel genes associated with T1DM and T2DM, respectively. Due to deficiency of studies focused on IGFI deficiency, only 28 of 4,212 were reported as curated genes related to the trait. The highest score of IGFI gene pairs was 0.33, which is under the threshold, and was predicted to be have no

TABLE 3 | Top 5 related genes with four diseases.

GD	PCOS	T1D	T2D
CCL27	RBM14	ADM2 (+)	UPK3B (+)
CXCL16	miR-1307	Enho (+)	miR-592
BECN1	CMKLR1	FUT6 (+)	NELFCD (+)
PROX1	AKT3	FUT7	miR-589 (+)
PTX3	GCGR	ATRNL1 (+)	Linc00641 (+)

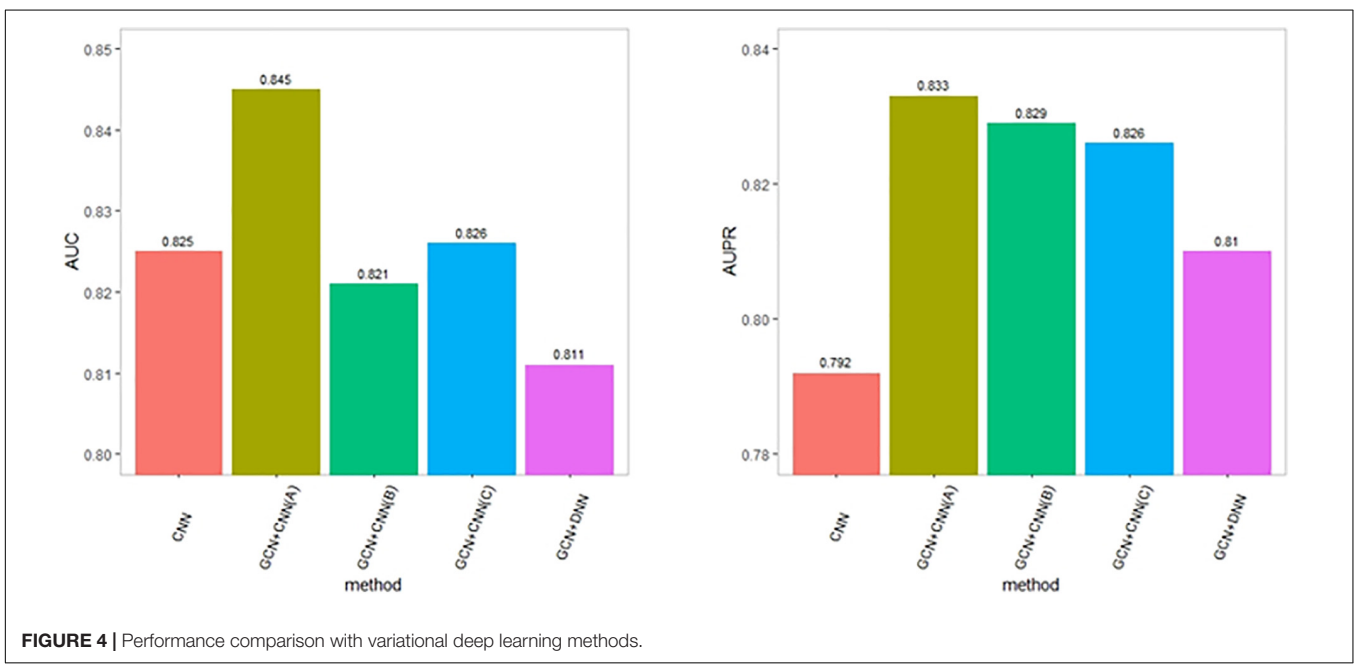


FIGURE 4 | Performance comparison with variational deep learning methods.

associations by DeepGP. For the rest of the diseases, 772 of 3,778 genes were shared.

Case Study

Graves' Disease

Graves' disease is mainly mediated by T cells which produce cytokines and chemokines in abnormal amounts. CCL27 is reported to be associated with serum chemokine concentrations detected in GD, which might be a good biomarker for GD (Hiratsuka et al., 2015). CXCL16 have been demonstrated to bind to the unique receptor CXCR6, which is expressed on a subset of multiple types of T cells, and it has been implicated in the pathogenesis of atherosclerosis and GD (Günther et al., 2012). It has been identified that loci at BECN1 can be denoted as a significant extract differentially methylated region for Graves' orbitopathy under the context of GD (Shi et al., 2019).

Polycystic Ovary Syndrome

RBM14 and miR-1307 have been reported to be up-regulated in PCOS patients in the study of Xu et al. (2015) and Che et al. (2020). Besides, CMKLR1, known as chemerin chemokine-like receptor 1, is the receptor of chemerin which is expressed at both mRNA and protein levels in human granulosa cells, and it has been reported to vary in women with PCOS (Bongrani et al., 2019).

Diabetes Mellitus

Due to the pathophysiological characteristics and many other potential etiopathogenesis factors T1DM and T2DM share, there are genes linked to both diseases such as *GLIS3* (Mahajan et al., 2014), *EIL2AK3* (also named as *PERK*; Delépine et al., 2000), etc. According to the result obtained from our method, we identified 2,582 genes related to T1D and 1,153 genes related to T2D, with a number of 474 shared genes. We then searched the top-ranked predicted genes related to T1D and T2D. According to the work of Ahmed et al., ADM2 is found to be preferentially up-regulated by bacteroides dorei (BD), which is a bacteria increased significantly at the time of onset of T1D. Enho is an energy homeostasis-associated gene that can produce a regulatory peptides named adropin, which has been identified to be strongly associated with type 1 diabetes in children (Polkowska et al., 2019). *FUT7* gene has been demonstrated to be linked with an antigen termed bile salt-dependent lipase which is reported to be associated with type 1 diabetes (Panicot et al., 1999). UPK3B is regarded as a mesothelial-like cell marker of a major adipocyte progenitor cell subpopulations which may induce adipocyte dysfunction in visceral adipose tissue in type 2 diabetes (Strieder-Barboza et al., 2020). miR-592 is reported to be associated with T2D due to its background of insulin resistance by Song et al. (2019). NELFCD has been identified to be matched with risk haplotypes across five FDM-risk haplotype, which is further identified in a common T2D gene, *ANK1*.

Therefore, we can illustrate that the predicted disease-gene pairs are reliable from the case studies mentioned above; top related genes with diseases are shown in **Table 3**. The "+" means the gene identified with T1D (T2D) is also related to T2D (T1D) according to the DisGeNET database.

DISCUSSION

In summary, we proposed a disease gene prediction method based on integrated deep learning models. We construct gene features considering both biological process and "linkage disequilibrium" theory. Identifying disease genes merely based on disease-susceptible loci identified by GWAS studies may be inaccurate due to LD. It has been indicated that genetic variants can affect the phenotype by regulating the gene expression level. In this study, eQTL data are also utilized to extract gene features. To fully use the underlying information contained in a gene interaction network, GCN was applied to extract comprehensive gene features based on the constructed gene network. Therefore, our method exploits the predictive power derived from complementary data types and the underlying network simultaneously. Disease features are derived based on disease similarity, which is calculated by the method named ImpAESim. Finally, we combined gene and disease features as a disease–gene pair feature. CNN was then used to classify the disease–gene pairs as a binary classification task.

As a result, DeepGP achieved an average AUC of 0.845 and an average AUPR of 0.833 after a 10 times 10-cross-validation based on the constructed training set, which is superior to other classic machine learning and deep learning methods. We then used the well-trained model to predict the novel disease-gene associations based on the disease–gene pairs which have not been reported to be associated before. We verified the prediction results based on a case study. Most of the top disease-related genes have reported evidences to illustrate that the genes may have associations with the diseases. In addition, we also identified 474 genes shared by T1DM and T2DM which may be helpful in designing therapeutic methods for diabetes mellitus patients. Therefore, the novel disease genes identified by DeepGP provide a strong support for the feasibility of extracting diagnostic markers for future validation and shed light on new strategies for the diagnosis and treatment of endocrine diseases.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material; further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

NZ did most of the work and wrote the manuscript. HW, CX, and LZ assisted in completing the experiments. TZ proposed the idea of the manuscript. HW made a great contribution in modifying the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Barral, S., Bird, T., Goate, A., Farlow, M. R., Diaz-Arrastia, R., Bennett, D. A., et al. (2012). Genotype patterns at PICALM, CR1, BIN1, CLU, and APOE genes are associated with episodic memory. *Neurology* 78, 1464–1471. doi: 10.1212/wnl.0b013e3182553c48

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics* 26, 2924–2926. doi: 10.1093/bioinformatics/btq538

Belfiore, A., and LeRoith, D. (2018). *Principles of Endocrinology and Hormone Action*. New York, NY: Springer.

Bongrani, A., Mellouk, N., Rame, C., Cornuau, M., Guérif, F., Froment, P., et al. (2019). Ovarian expression of adipokines in polycystic ovary syndrome: a role for chemerin, omentin, and apelin in follicular growth arrest and ovulatory dysfunction? *Int. J. Mol. Sci.* 20:3778. doi: 10.3390/ijms20153778

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Res.* 47, D1005–D1012.

Carithers, L. J., and Moore, H. M. (2015). *The Genotype-Tissue Expression (GTEx) Project*. New Rochelle, NY: Mary Ann Liebert, Inc.

Che, Q., Liu, M., Zhang, D., Lu, Y., Xu, J., Lu, X., et al. (2020). Long noncoding RNA HUPCOS promotes follicular fluid androgen excess in PCOS patients via aromatase inhibition. *J. Clin. Endocrinol. Metab.* 105, 1086–1097. doi: 10.1210/clinem/dgaa060

Chen, L., Zhang, Y.-H., Huang, G., Pan, X., Huang, T., and Cai, Y. D. (2019). Inferring novel genes related to oral cancer with a network embedding method and one-class learning algorithms. *Gene Ther.* 26, 465–478. doi: 10.1038/s41434-019-0099-y

Chen, L., Zhang, Y.-H., Zhang, Z., Huang, T., and Cai, Y. D. (2018). Inferring novel tumor suppressor genes with a protein-protein interaction network and network diffusion algorithms. *Mol. Ther. Methods Clin. Dev.* 10, 57–67. doi: 10.1016/j.omtm.2018.06.007

Cuneo, R. C., Hickman, P. E., Wallace, J. D., Teh, B. T., Ward, G., Veldhuis, J. D., et al. (1995). Altered endogenous growth hormone secretory kinetics and diurnal GH—binding protein profiles in adults with chronic liver disease. *Clin. Endocrinol.* 43, 265–275. doi: 10.1111/j.1365-2265.1995.tb02031.x

Day, F. R., Hinds, D. A., Tung, J. Y., Stolk, L., Styrkarsdottir, U., Saxena, R., et al. (2015). Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat. Commun.* 6, 1–7.

Delépine, M., Nicolino, M., Barrett, T., Golamaully, M., Lathrop, G. M., and Julier, C. (2000). EIF2AK3, encoding translation initiation factor 2-α kinase 3, is mutated in patients with Wolcott-Rallison syndrome. *Nat. Genet.* 25, 406–409. doi: 10.1038/78085

Dvornikova, K. A., Bystrova, E. Y., Platonova, O. N., and Churilov, L. P. (2020). Polymorphism of toll-like receptor genes and autoimmune endocrine diseases. *Autoimmun. Rev.* 19:102496. doi: 10.1016/j.autrev.2020.102496

Ehrmann, D. A., Barnes, R. B., Rosenfield, R. L., Cavaghan, M. K., and Imperial, J. (1999). Prevalence of impaired glucose tolerance and diabetes in women with polycystic ovary syndrome. *Diabetes Care* 22, 141–146. doi: 10.2337/diacare.22.1.141

Eizirik, D. L., Pasquali, L., and Cnop, M. (2020). Pancreatic β-cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nat. Rev. Endocrinol.* 16, 349–362. doi: 10.1038/s41574-020-0355-7

Fathima, N., Narne, P., and Ishaq, M. (2019). Association and gene–gene interaction analyses for polymorphic variants in CTLA-4 and FOXP3 genes: role in susceptibility to autoimmune thyroid disease. *Endocrine* 64, 591–604. doi: 10.1007/s12020-019-01859-3

Freeman, G., Cowling, B. J., and Schooling, C. M. (2013). Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *Int. J. Epidemiol.* 42, 1157–1163. doi: 10.1093/ije/dyt110

Gazzaruso, C., Gola, M., Karamouzis, I., Giubbini, R., and Giustina, A. (2014). Cardiovascular risk in adult patients with growth hormone (GH) deficiency and following substitution with GH—an update. *J. Clin. Endocrinol. Metab.m* 99, 18–29. doi: 10.1210/jc.2013-2394

Gromada, J., Chabosseau, P., and Rutter, G. A. (2018). The α-cell in diabetes mellitus. *Nat. Rev. Endocrinol.* 14, 694–704.

Günther, C., Carballido-Perrig, N., Kaesler, S., Carballido, J. M., and Biedermann, T. (2012). CXCL16 and CXCR6 are upregulated in psoriasis and mediate cutaneous recruitment of human CD8+ T cells. *J. Invest. Dermatol.* 132, 626–634. doi: 10.1038/jid.2011.371

Hayes, M. G., Urbanek, M., Ehrmann, D. A., Armstrong, L. L., Lee, J. Y., Sisk, R., et al. (2015). Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. *Nat. Commun.* 6, 1–13.

Hiratsuka, I., Itoh, M., Yamada, H., Yamamoto, K., Tomatsu, E., Makino, M., et al. (2015). Simultaneous measurement of serum chemokines in autoimmune thyroid diseases: possible role of IP-10 in the inflammatory response. *Endocr. J.* 62, EJ15–EJ0448.

Liu, C., Cui, P., and Huang, T. (2017). Identification of cell cycle-regulated genes by convolutional neural network. *Comb. Chem. High Throughput Screen.* 20, 603–611.

Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* 46, 234–244. doi: 10.1038/ng.2897

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.

Min, S., Lee, B., and Yoon Deep, S. (2017). learning in bioinformatics. *Brief. Bioinformatics* 18, 851–869.

Panicot, L., Mas, E., Thivolet, C., and Lombardo, D. (1999). Circulating antibodies against an exocrine pancreatic enzyme in type 1 diabetes. *Diabetes* 48, 2316–2323. doi: 10.2337/diabetes.48.12.2316

Perricone, C., and Shoenfeld, Y. (2019). *Mosaic of Autoimmunity: the Novel Factors of Autoimmune Diseases*. Cambridge, MA: Academic Press.

Piñero, J., Bravo, À, Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45:gkw943.

Polkowska, A., Pasierowska, I. E., Pasławska, M., Pawluczuk, E., and Bossowski, A. (2019). Assessment of serum concentrations of adropin, afamin, and neudesin in children with type 1 diabetes. *BioMed. Res. Int.* 2019:6128410.

Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil. Steril.* 81, 19–25. doi: 10.1016/j.fertnstert.2003.10.004

Soh, S-B, and Aw, T-C (2019). Laboratory testing in thyroid conditions-pitfalls and clinical utility. *Ann. Lab. Med.* 39:3. doi: 10.3343/alm.2019.39.1.3

Shi, T.-T., Hua, L., Xin, Z., Li, Y., Liu, W., and Yang, Y. L. (2019). Identifying and validating genes with DNA methylation data in the context of biological network for Chinese patients with Graves' orbitopathy. *Int. J. Endocrinol.* 2019:6212681.

Smith, M. J., Rihanek, M., Coleman, B. M., Gottlieb, P. A., Sarapura, V. D., and Cambier, J. C. (2018). Activation of thyroid antigen-reactive B cells in recent onset autoimmune thyroid disease patients. *J. Autoimmun.* 89, 82–89. doi: 10.1016/j.jaut.2017.12.001

Song, Y., Wu, L., Li, M., Xiong, X., Fang, Z., Zhou, J., et al. (2019). Down-regulation of MicroRNA-592 in obesity contributes to hyperglycemia and insulin resistance. *EBioMedicine* 42, 494–503. doi: 10.1016/j.ebiom.2019.03.041

Strieder-Barboza, C., Flesher, C. G., Geletka, L. M., Orourke, RW., and Lumeng, CN. (2020). 1973-P: single-Nuclei transcriptomics of human adipose tissue identify distinct adipocyte progenitor subpopulations in type 2 diabetes. *Am. Diabetes Assoc.* 69(Suppl. 1).

Xiao, W., Liu, Z., Lin, J., Li, J., Wu, K., Ma, Y., et al. (2015). Polymorphisms in TLR1, TLR6 and TLR10 genes and the risk of Graves' disease. *Autoimmunity* 48, 13–18.

Xu, B., Zhang, Y.-W., Tong, X.-H., and Liu, Y. S. (2015). Characterization of microRNA profile in human cumulus granulosa cells: identification of microRNAs that regulate Notch signaling and are associated with PCOS. *Mol. Cell. Endocrinol.* 404, 26–36. doi: 10.1016/j.mce.2015.01.030

Zhang, H., Wang, S., and Huang, T. (2020). Identification of chronic hypersensitivity pneumonitis biomarkers with machine learning and differential co-expression analysis. *Curr. Gene Ther.* 21.

Zhao, T., Hu, Y., and Cheng, L. (2020). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinformatics* 20:bbaa212.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48, 481–487. doi: 10.1038/ng.3538

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Zhang, Wang, Xu, Zhang and Zang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 16 July 2021
doi: 10.3389/fcell.2021.686909

Multi-Omics Data Integration Analysis of an Immune-Related Gene Signature in LGG Patients With Epilepsy

Quan Cheng[1,2†], Weiwei Duan[3†], Shiqing He[4], Chen Li[5], Hui Cao[6], Kun Liu[7], Weijie Ye[8], Bo Yuan[7] and Zhiwei Xia[9]*

[1] Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha, China, [2] National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China, [3] Department of Neurology, Xiangya Hospital, Central South University, Changsha, China, [4] Department of Neurosurgery, Affiliated Nanhua Hospital, Hengyang Medical College, University of South China, Hengyang, China, [5] Department of Rehabilitation Medicine, Hunan Provincial People's Hospital, Hunan Normal University, Changsha, China, [6] Department of Psychiatry, The Second People's Hospital of Hunan Province, The Hospital of Hunan University of Chinese Medicine, Changsha, China, [7] Department of Cerebrovascular Surgery, The Second People's Hospital of Hunan Province, The Hospital of Hunan University of Chinese Medicine, Changsha, China, [8] Department of Clinical Pharmacology, Xiangya Hospital, Central South University, Changsha, China, [9] Department of Neurology, Hunan Aerospace Hospital, Changsha Medical University, Changsha, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Jun Liu,
Yuebei People's Hospital, China
Jukun Song,
Guizhou Provincial People's Hospital, China
Weifeng Hong,
The First Affiliated Hospital of Guangdong Pharmaceutical University, China

***Correspondence:**
Zhiwei Xia
xiazhiwei2011@gmail.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: 28 March 2021
Accepted: 22 June 2021
Published: 16 July 2021

Citation:
Cheng Q, Duan W, He S, Li C, Cao H, Liu K, Ye W, Yuan B and Xia Z (2021) Multi-Omics Data Integration Analysis of an Immune-Related Gene Signature in LGG Patients With Epilepsy. Front. Cell Dev. Biol. 9:686909. doi: 10.3389/fcell.2021.686909

Background: The tumor immune microenvironment significantly affects tumor occurrence, progression, and prognosis, but its impact on the prognosis of low-grade glioma (LGG) patients with epilepsy has not been reported. Hence, the purpose of this study is to explore its effect on LGG patients with epilepsy.

Methods: The data of LGG patients derived from the TCGA database. The level of immune cell infiltration and the proportion of 22 immune cells were evaluated by ESTIMATE and CIBERSORT algorithms, respectively. The Cox and LASSO regression analysis was adopted to determine the DEGs, and further established the clustering and risk score models. The association between genomic alterations and risk score was investigated using CNV and somatic mutation data. GSVA was adopted to identify the immunological pathways, immune infiltration and inflammatory profiles related to the signature genes. The Tumor Immune Dysfunction and Exclusion (TIDE) algorithm and GDSC database were used to predict the patient's response to immunotherapy and chemotherapy, respectively.

Results: The prognosis of LGG patients with epilepsy was associated with the immune score. Three prognostic DEGs (ABCC3, PDPN, and INA) were screened out. The expression of signature genes was regulated by DNA methylation. The clustering and risk score models could stratify glioma patients into distinct prognosis groups. The risk score was an independent predictor in prognosis, with a high risk-score indicating a poor prognosis, more malignant clinicopathological and genomic aberration features. The nomogram had the better predictive ability. Patients at high risk had a higher level of macrophage infiltration and increased inflammatory activities associated with T cells and macrophages. While the higher percentage of NK CD56bright cell and more

active inflammatory activity associated with B cell were present in the low-risk patients. The signature genes participated in the regulation of immune-related pathways, such as IL6-JAK-STAT3 signaling, IFN-α response, IFN-γ response, and TNFA-signaling-via-NFKB pathways. The high-risk patients were more likely to benefit from anti-PD1 and temozolomide (TMZ) treatment.

Conclusion: An immune-related gene signature was established based on ABCC3, PDPN, and INA, which can be used to predict the prognosis, immune infiltration status, immunotherapy and chemotherapy response of LGG patients with epilepsy.

Keywords: glioma, epilepsy, immune, immunotherapy, prognosis

INTRODUCTION

As the most common brain malignant tumor, patients with glioma are often complicated with epilepsy (Goldstein and Feyissa, 2018; Jones et al., 2019; Liu B. et al., 2019). Compared with high-grade glioma, patients with low-grade glioma (LGG) have a significantly higher incidence of epilepsy (Huberfeld and Vecht, 2016), which may be related to the slower growth rate and higher IDH-1 mutation frequency of LGG (Cohen et al., 2013; Chen et al., 2018). The effect of epilepsy on the prognosis of LGG is currently unclear. Several studies have suggested that epilepsy is a positive prognostic factor in LGG patients (Blümcke et al., 2004; Krzyszkowski et al., 2004; Danfors et al., 2009). However, Lote et al. (1998) found that epilepsy had no effect on the prognosis after analyzing 379 LGG patients. This indicates that there are other factors affecting the prognosis of LGG patients with epilepsy.

In addition to tumor cells, non-tumor cells including immune cells are also important components of tumors. Immune cells infiltrating tumor tissues, such as peripheral macrophages, microglia, granulocytes, T lymphocytes, and myeloid-derived suppressor cells (Gieryng et al., 2017), create a tumor immune microenvironment that affects tumor development, metastasis, and response to treatment (Ma et al., 2018). However, its influence on LGG patients with epilepsy has not been established.

Hence, we explored the immune microenvironment of LGG patients with epilepsy and evaluated its impact on prognosis through the integration analysis of multi-omics data. We developed a gene signature related to immune infiltration, which can be used to predict the prognosis and therapeutic response of LGG patients with epilepsy.

MATERIALS AND METHODS

Data Acquisition

The RNA transcript data (Workflow Type: HTSeq-Counts), corresponding clinical information, DNA methylation (Platform: Illumina human methylation 450), somatic mutation (Workflow Type: MuTect2), and copy number variation (CNV) data of 473 LGG patients (WHO grade II and grade III) were downloaded from the TCGA database[1]. Among them, there were 297 patients with epilepsy and 176 patients without epilepsy. The sample screening process was shown in **Supplementary Figure 1**.

Inference of Immune Infiltration Level and Immune Cell Abundance

The immune score was calculated by the "ESTIMATE" algorithm (version 1.0.13) (Yoshihara et al., 2013) to infer the level of immune infiltration. The proportions of 22 immune cells were inferred by the "CIBERSORT" algorithm (Newman et al., 2015) using LM22, a leukocyte gene signature matrix[2].

Screening of Prognostic Signature Genes

The "DESeq2" R package (version 1.30.1) (Love et al., 2014) was used to search for differentially expressed genes (DEGs), |log(2) fold change| >2 and $p < 0.05$ as a threshold. The Cox regression analysis was adopted to find the DEGs with independent prognostic value. Finally, the DEGs with the most prognostic value and their regression coefficients (β) were obtained by the LASSO regression analysis.

Consensus Clustering Analysis

The consensus clustering analysis was conducted using the "ConsensusClusterPlus" R package (version 1.54.0) (Wilkerson and Hayes, 2010). The clustering run was permuted by varying the category number k from 2 to 10. The optimal *k* value was determined based on good consistency within the clusters and a relatively small incremental change in the area under the CDF curve.

Establishment and Evaluation of Risk Score Model

The risk score model was established as follows: $risk\ score = \Sigma_{i=1}^{n}(\mathrm{Geneexpr}_i \times \beta_i)$. The effectiveness of the risk model was evaluated by comparing the prognosis of high- and low-risk patients and calculating the AUC value of the ROC curve. The correlation between risk score and clinicopathological characteristics was further analyzed. The R package "rms" (version 6.1-0) was used to construct a nomogram.

[1]https://portal.gdc.cancer.gov/

[2]http://cibersort.stanford.edu/

Analysis of Genetic Alterations

The somatic mutation data were processed by the "maftools" R package (version 2.6.05) (Mayakonda et al., 2018), and the genes with the highest mutation frequency were screened and presented. Genomic Identification of Significant Targets in Cancer (GISTIC) 2.0 (Mermel et al., 2011) was used to analyze CNV data in reference to the Human genome reference consortium h19 derived from GISTIC 2·0. The threshold copy number at alteration peaks was obtained from GISTIC analysis.

The GSVA Analysis

The gene set variation analysis (GSVA) (Hänzelmann et al., 2013) was used to establish each sample's immune signature and the correlation analysis was further carried out. The correlogram and heatmap were drawn using the "corrgram" (version 1.13) and "pheatmap" R packages (version 1.0.12), respectively. The immunological pathways associated with signature genes were identified by GSVA based on the hallmark gene sets (Liberzon et al., 2015).

Therapy Response Prediction

The patients' response to immune checkpoint blocking therapy was predicted by the Tumor Immune Dysfunction and Exclusion (TIDE) (Jiang et al., 2018) algorithm. The patients' sensitivity to temozolomide (TMZ) therapy was predicted using the data derived from the GDSC database[3]. The "pRRophetic" R package (version 0.5) (Geeleher et al., 2014) was used for the estimation of the half-maximal inhibitory concentration (IC50), which represents the drug response.

Statistical Analysis

The statistical analysis was finished in R software (version 3.5.3). Statistical significance between groups was determined using the Student's *t*-test. Pearson rank was adopted in the correlation analysis. The "survival" R package (version 3.2-7) and Kaplan-Meier method were used for survival analysis. The AUC value of ROC curve was calculated by the "survival ROC" R package (version 1.0.3). $P < 0.05$ was considered to be statistically significant.

RESULTS

The Level of Immune Infiltration Affected the Prognosis

The characteristics of all included patients were summarized in **Supplementary Table 1**. The workflow designed for the study was shown in **Figure 1**. According to the immune score and whether they had epilepsy, all patients were divided into four groups: C1(high immune score, seizure), C2 (high immune score, non-seizure), C3 (low immune score, seizure), and C4 (low immune score, non-seizure). Here, the immune scores that were >0 were considered high while immune scores that were <0 were considered low (Wen et al., 2020). Through survival analysis, we found that a high immune score suggested a poor prognosis in LGG patients with epilepsy ($p = 0.0025$, **Figure 2A**). The baseline characteristics of LGG patients with epilepsy were shown in **Supplementary Table 2**. We further compared the abundance of 22 immune cells between high and low immune score groups in LGG patients with epilepsy. The results showed that the patients with high immune score had a higher percentage of macrophages (M2), and a higher proportion of naive B cells, activated dendritic cells, neutrophils, activated memory $CD4^+$ T cells, and resting memory $CD4^+$ T cells also appeared in the high immune score group ($p < 0.05$, **Figure 2B**). While the patients in the low immune score group had higher percentages of memory B cells, naive $CD4^+$ T cells, resting mast cells, $CD8^+$ T cells, and follicular T helper cells ($p < 0.05$, **Figure 2B**). Further Cox regression analysis indicated that activated memory $CD4^+$ T cells and activated dendritic cells were negative factors for prognosis ($p < 0.05$, **Supplementary Table 3**).

Three Prognostic Signature Genes Were Identified

A total of 12278 DEGs were found between the high and low immune score groups in LGG patients with epilepsy. Furthermore, 51 DEGs with independent prognostic value were identified by Cox regression analysis ($p < 0.05$). Finally, three genes (ABCC3, INA, PDPN) with the most prognostic value were obtained using the LASSO regression (**Figures 3A,B**) and their coefficients (ABCC3: 0.0314, INA: −0.0018, PDPN: 0.0957) were shown in **Figure 3C**. Consequently, a prognostic immune-related gene signature was established. We found that the high immune score group had higher expression levels of ABCC3 and PDPN, while the low immune score group showed higher expression of INA ($p < 0.001$, **Figures 3D–F**). The high expression of ABCC3 (HR 3.6, $p < 0.001$) and PDPN (HR 3.73, $p < 0.001$) was associated with poor prognosis (**Figures 3G,H**), while the high expression of the INA (HR 0.34, $p < 0.001$) was a positive prognostic factor (**Figure 3I**). This indicates that PDPN and ABCC3 are risk genes, and INA is a protective gene for LGG patients with epilepsy.

We also explored the relationship between the methylation levels of signature genes and their expression levels in LGG patients with epilepsy. We found that the expression of ABCC3 ($R = -0.51$, $p < 0.001$) and PDPN ($R = -0.56$, $p < 0.001$) genes was negatively correlated with methylation levels ($p < 0.001$), while the INA ($R = 0.35$, $p < 0.001$) was positively correlated (**Figure 3J** and **Supplementary Figure 2**). This suggests that the hypomethylation of ABCC3 and PDPN is present in LGG patients with epilepsy who have high immune score and poor prognosis.

Clustering Model Based on Signature Genes Identified Two Subgroups With Distinct Outcomes

The samples of LGG patients with epilepsy were grouped using consensus clustering analysis to preliminarily explore the prognostic value of signature genes. In consensus clustering analysis, $k = 2$ was determined as the optimal selection,

[3]https://www.cancerrxgene.org

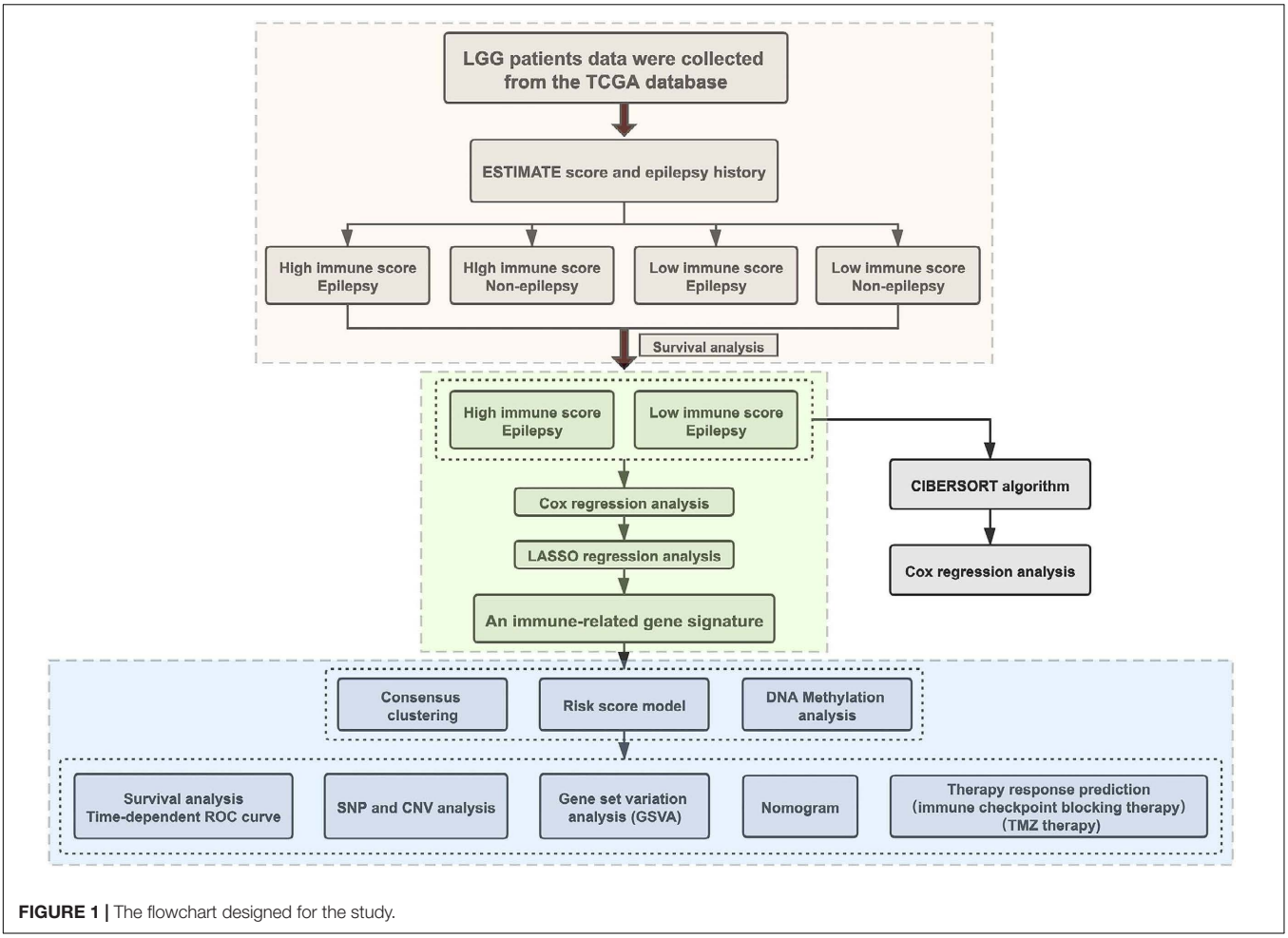


FIGURE 1 | The flowchart designed for the study.

which had the small incremental change in the area under CDF curve while keeping the maximal consensus within clusters (**Supplementary Figures 3A–C**). Therefore, we could obtain two patient subgroups, cluster1 and cluster2. And the principal component analysis (PCA) could obviously distinguish the two clusters, which also proved the rationality of the clustering (**Supplementary Figure 3D**). A significant difference in prognosis was found between cluster1 and cluster2. The overall survival (OS), disease specific survival (DSS), and progression free interval (PFI) of patients in cluster1 were all better than those in cluster 2 ($p < 0.0001$, **Supplementary Figure 3E**).

Risk Score Model Based on Signature Genes Had Excellent Predictive Ability in Prognosis

We established a risk score model based on the signature genes. According to the median value of patients' risk score calculated by the risk score model, the LGG patients with epilepsy were divided into high- and low-risk groups. The prognosis of patients was significantly different between the two groups, and the OS, DSS, and PFI of patients with high risk were significantly worse ($p < 0.0001$, **Figures 4A–C**). The AUC values of 3-year and 5-year ROC curves of the risk model were 0.846 and 0.759 in OS, 0.863 and 0.760 in DSS, 0.734 and 0.711 in PFI, respectively (**Figures 4D–F**). The above results show that the risk model has a good prognosis prediction ability.

Risk Score Was Correlated With Clinicopathological and Genomic Aberration Features

We conducted a subgroup analysis to explore the correlation between clinicopathological features and risk score. We found significant differences in risk scores within the subgroups distinguished by MGMT promoter status, IDH status, subtype, WHO grade, and 1p19q codel status ($p < 0.001$), while no significant differences were found within gender and age subgroups ($p > 0.05$). Patients with unmethylated MGMT promoter, 1p/19q non-codel, classical, and mesenchymal subtype, IDH-wildtype, and higher WHO grade had higher risk scores (**Supplementary Figure 4**).

Somatic mutations were observed in 137 (94.48%) of the high-risk samples and in 142 (97.93%) samples of the low-risk group. The mutation frequency of IDH1, CIC, and IDH2 genes

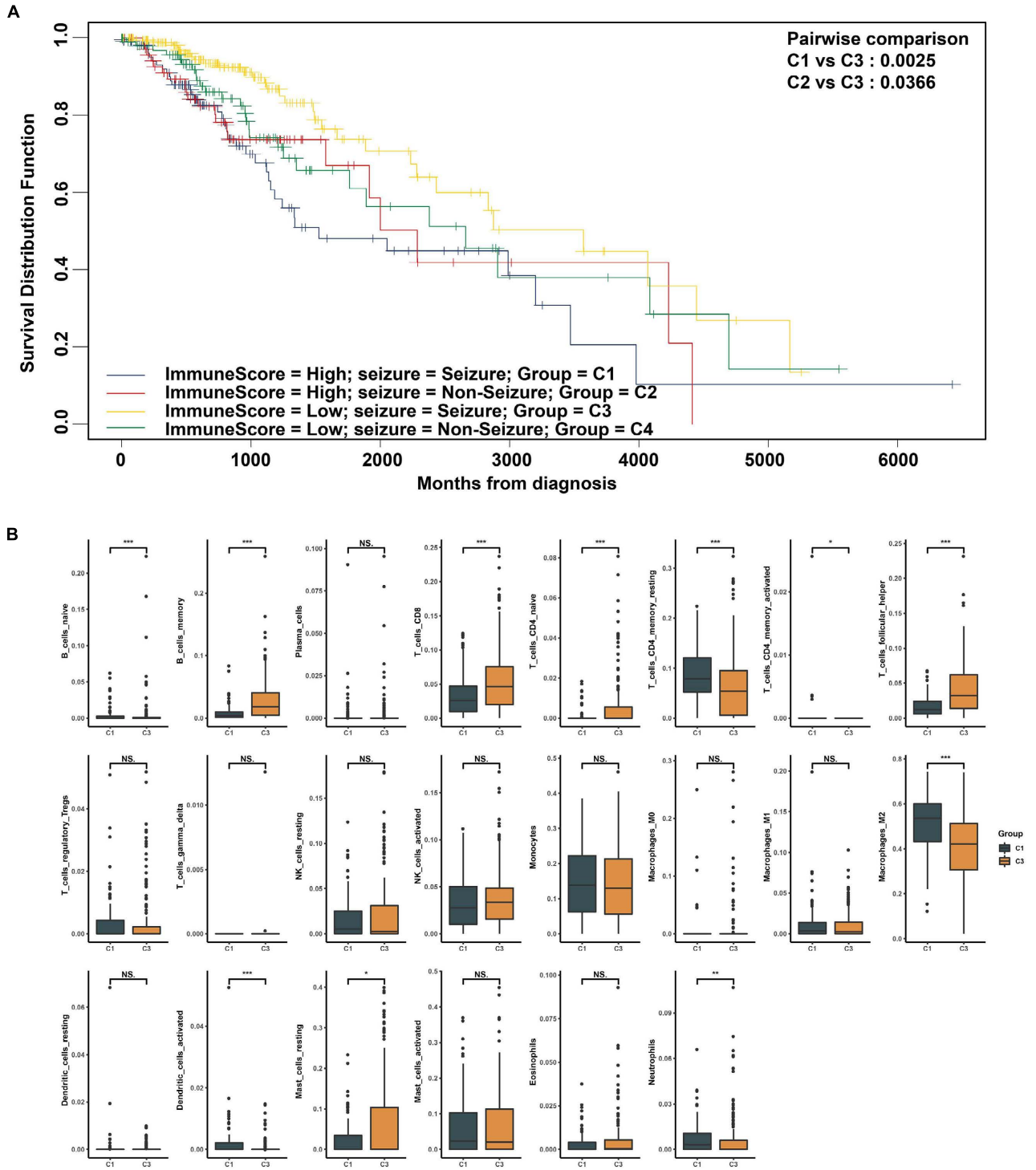


FIGURE 2 | (A) Comparison of survival rate of low-grade glioma (LGG) subgroups differentiated by seizure and immune score. **(B)** Comparison of 22 immune cell components between high and low immune score groups in LGG patients with epilepsy. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

in high-risk patients was lower than that in low-risk patients (IDH1: 65% vs. 88%, $p < 0.001$; CIC: 10% vs. 33%, $p < 0.001$; IDH2 0% vs. 9%, $p < 0.001$), while high-risk patients had a higher frequency of mutations in EGFR, NF1, and PTEN (EGFR: 14% vs. 0%, $p < 0.001$; NF1: 11% vs. 1%, $p < 0.001$; PTEN 10% vs. 1%, $p < 0.001$) (**Figures 5A,B** and **Supplementary Figure 5**).

In the analysis of CNV data, several significant amplification regions containing multiple oncogenes, such as 12q14.1(CDK4),

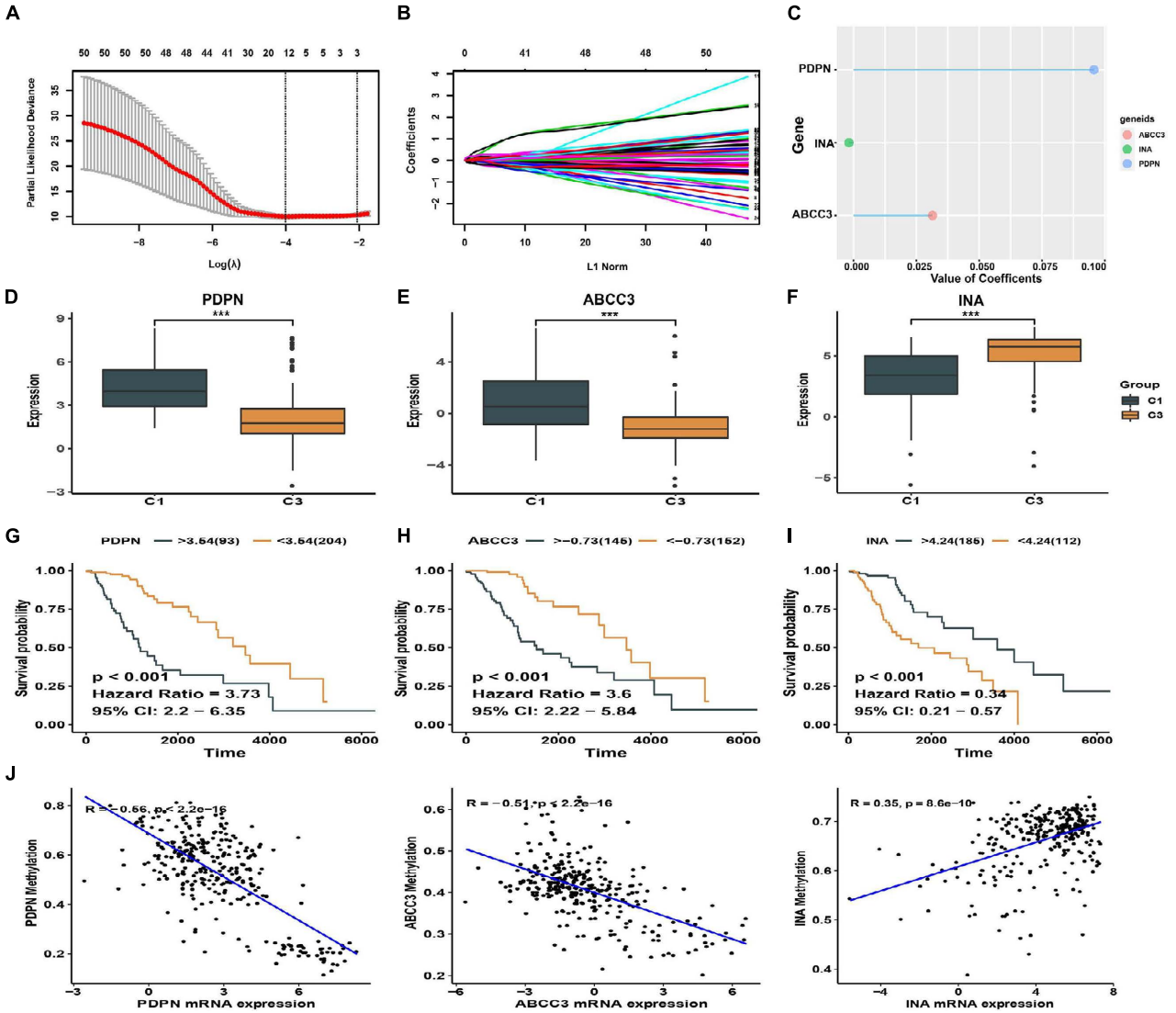


FIGURE 3 | Identification of signature genes. **(A)** The cross-validation plot of LASSO regression [the dashed lines signify the optimal log (λ) value]. **(B)** LASSO coefficients of 51 DEGs with independent prognostic value. **(C)** LASSO coefficients of the signature genes. **(D–F)** Expression difference of ABCC3, PDPN, INA between high and low immune score groups. **(G–I)** Kaplan-Meier plot showing different prognosis based on the differential expression of ABCC3, PDPN, and INA. **(J)** Correlation between DNA methylation and signature genes expression. ***$p < 0.001$.

7p11.2 (EGFR), 4q12 (PDGFRA), and 1q32.1 (PIK3C2B), were identified in high-risk patients (**Figure 5C**). In addition, focal deletion peak 9p21.3 (CDKN2A, CDKN2B) was detected in patients with high risk (**Figure 5C**). In the low-risk group, the focal amplification and deletion peaks also appeared, but they had significantly lower *G* values (**Figure 5D**).

The Risk Score Was an Independent Predictor for Prognosis

We comprehensively analyzed the effect of risk score and clinicopathological characteristics on a clinical outcome to assess whether the risk score has an independent prognostic value. In univariate Cox regression analysis, age, WHO grade, IDH status, subtype, and risk score were identified as survival-related factors ($p < 0.001$, **Supplementary Table 4**). Further multivariate Cox regression analysis indicated that age and risk score were independent prognostic factors. Age $\geq$ 45 (HR 3.29, $p < 0.001$) was associated with a worse prognosis, while low risk score (HR 0.40, $p = 0.026$) was a benign prognostic factor (**Figure 6A**).

In addition, we constructed a nomogram that integrated the clinicopathological features and risk score to improve clinical feasibility (**Figure 6B**). The predicted 3/5-year OS were close to the actually observed ones, as shown in the calibration curve (**Figure 6C**). According to the median value of patients' points calculated by the nomogram, we could also obtain two groups of patients, the high-risk group and the low-risk group. Survival analysis suggested that high-risk patients had worse OS than

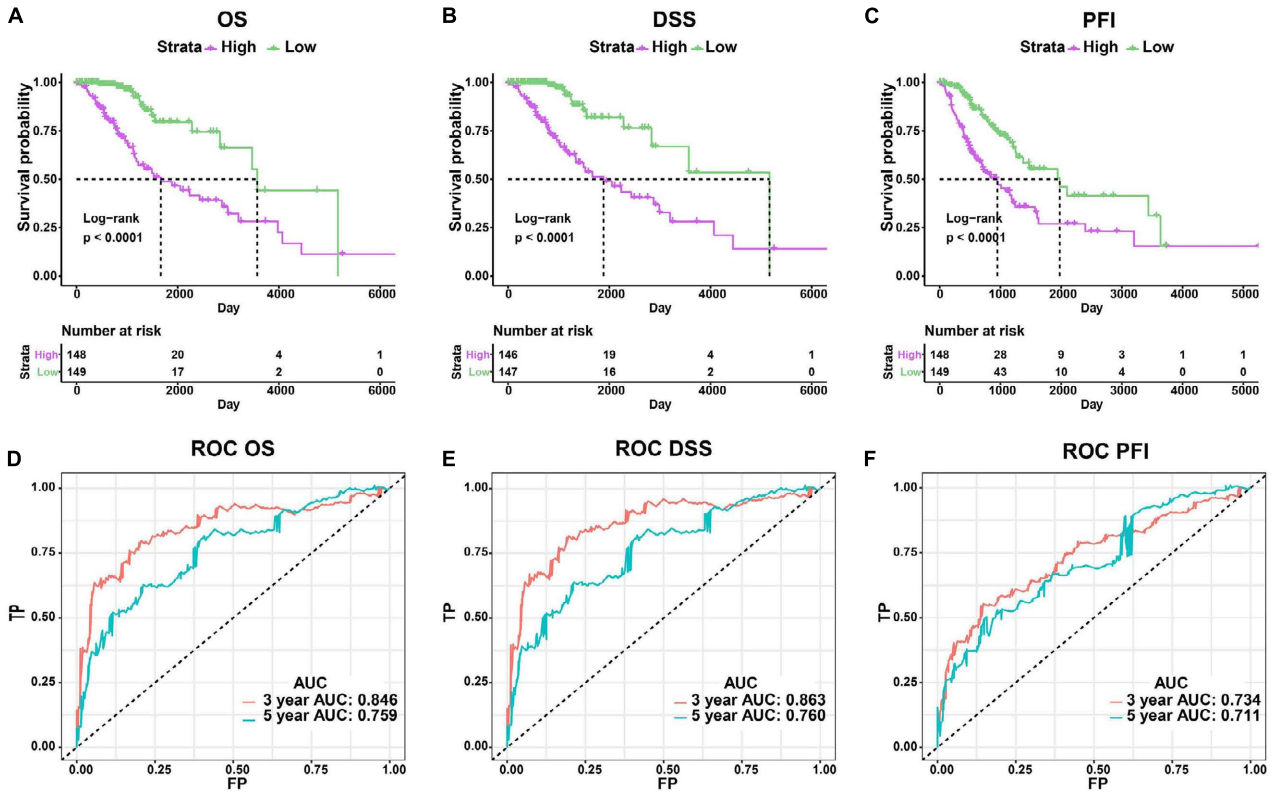


FIGURE 4 | Evaluation of the risk score model. **(A–C)** Comparison of clinical outcomes between high- and low-risk groups using OS, DSS, and PFI as the endpoints. **(D–F)** ROC curves and AUC values of risk model.

low-risk patients (p = 0.00037, **Figure 6D**). The AUC values of ROC curves from the nomogram were 0.917 at predicted 3-year OS and 0.783 at 5-year OS, both of which were better than those obtained by using the risk model alone (**Figure 6E**). Therefore, the nomogram has better predictive power compared with the risk model.

Immunological Pathways, Immune Infiltration and Inflammatory Profiles Related to Signature Genes

We explored the immunological activities related to the signature genes. We first carried out the GSVA based on specific marker genes of 24 immune cells and seven metagenes representing different types of inflammation and immune responses (Rody et al., 2009). Further correlation analysis showed that macrophages infiltration was positively correlated with risk score (p = 0.002; **Figures 7A,B** and **Supplementary Table 5**). While B cells, T cells, T helper (Th) cells, follicular helper T cell (TFH), Th17 cells, Th1 cells, Th2 cells, gamma delta T Cells (Tgd), central memory T cell (Tcm), effector memory T cell (Tem), NK CD56dim cells, NK CD56bright cells, mast cells, dendritic cell (DC), activated DC (aDC), immature DC (iDC), cytotoxic cells, and neutrophils infiltration were negatively correlated with risk score, especially the NK CD56bright cells (p < 0.05; **Figures 7A,B** and **Supplementary Table 5**). In terms of inflammatory activities, MHC-II, STAT1, LCK, MHC-I, HCK, and Interferon were positively correlated with risk score, while IgG, a marker for B cells, was negatively correlated (p < 0.001; **Figures 7C,D** and **Supplementary Table 6**).

We also conducted GSVA to identify the immunological pathways associated with the signature genes. Compared with the low-risk group, the IL6-JAK-STAT3 signaling, IFN-α response, IFN-γ response, and TNFA-signaling-via-NFKB pathways were significantly active in patients with high-risk (p < 0.05, **Supplementary Figure 6**).

Risk Score Was Correlated With Immunotherapy and Chemotherapy Response

Cancer growth and progression are associated with immunosuppression, and it is currently believed that the immune checkpoint pathways are the major mechanism of tumor immune resistance (Pardoll, 2012). The immune checkpoint blockade has become a promising immunotherapy method for various tumors including glioma (Pardoll, 2012; Hodges et al., 2017). In this study, we found that immune checkpoint molecules CD724 (PD-L1), CTL4, LAG3, and PDCD1 (PD1) were expressed at higher levels in high-risk patients (p < 0.01, **Figure 8A**). We further predicted the patients' response to anti-CTLA4 and

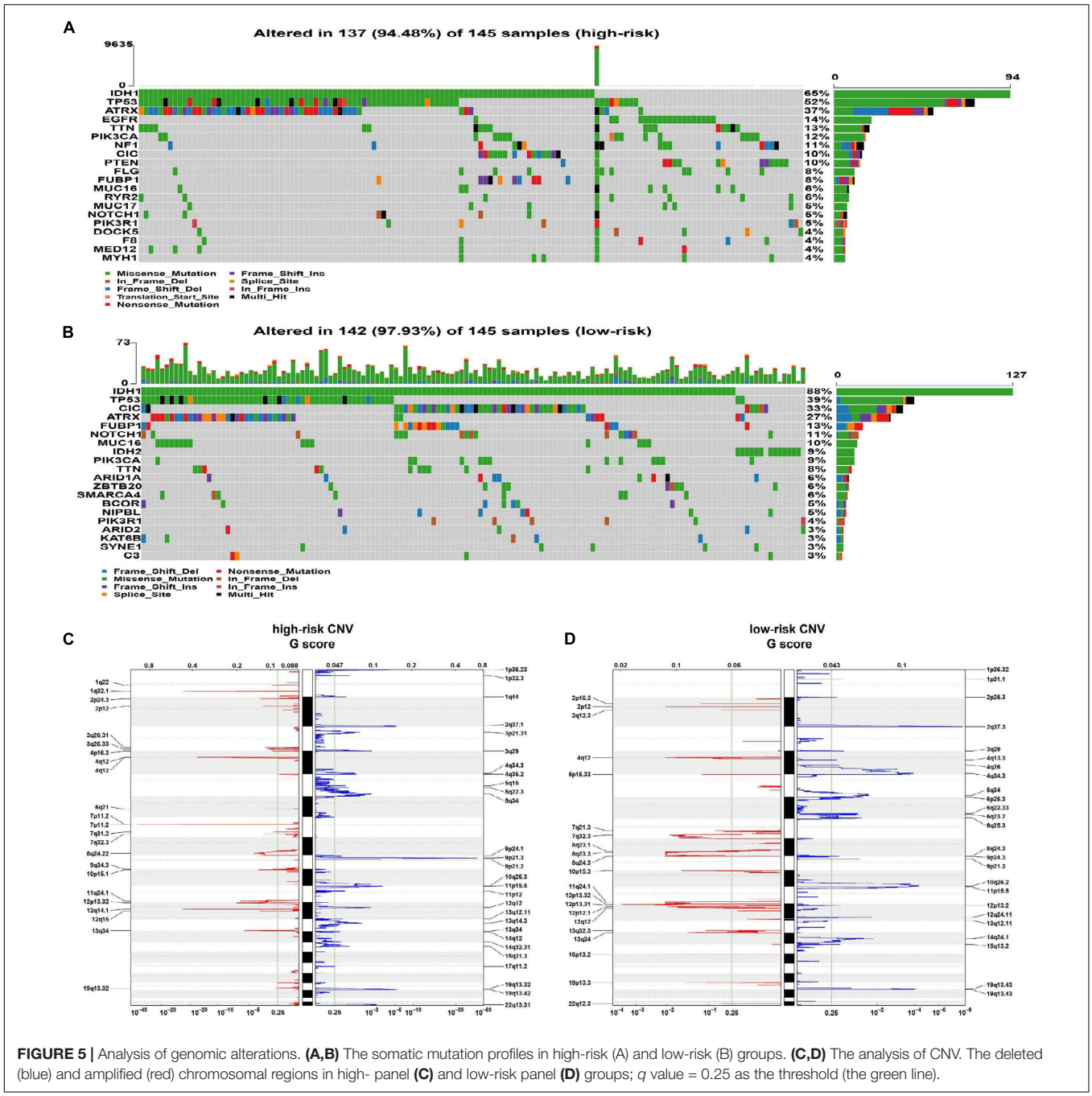


FIGURE 5 | Analysis of genomic alterations. **(A,B)** The somatic mutation profiles in high-risk (A) and low-risk (B) groups. **(C,D)** The analysis of CNV. The deleted (blue) and amplified (red) chromosomal regions in high- panel **(C)** and low-risk panel **(D)** groups; *q* value = 0.25 as the threshold (the green line).

anti-PD1 therapy and found that high-risk patients were more likely to respond to anti-PD1 therapy than low-risk patients (p = 0.01, **Figure 8B**). No significant difference was observed in response to anti-CTLA4 therapy between the two groups ($p > 0.05$, **Figure 8B**). Moreover, a higher tumor mutation burden (TMB) was observed in high-risk patients ($p < 0.01$, **Figure 8C**).

Meanwhile, we also predicted the patients' sensitivity to TMZ therapy, which is currently the most commonly used chemotherapeutics for glioma. We trained a predictive model using cell line data derived from the GDSC database. With this predictive model, the IC50 value of each patient's TMZ was estimated. We found that the estimated IC50 value in low-risk patients was higher compared with high-risk patients ($p < 0.001$, **Figure 8D**). This indicates that patients with high risk may be more sensitive to TMZ therapy.

DISCUSSION

In this study, we explored the effect of the immune microenvironment on the prognosis of LGG patients with

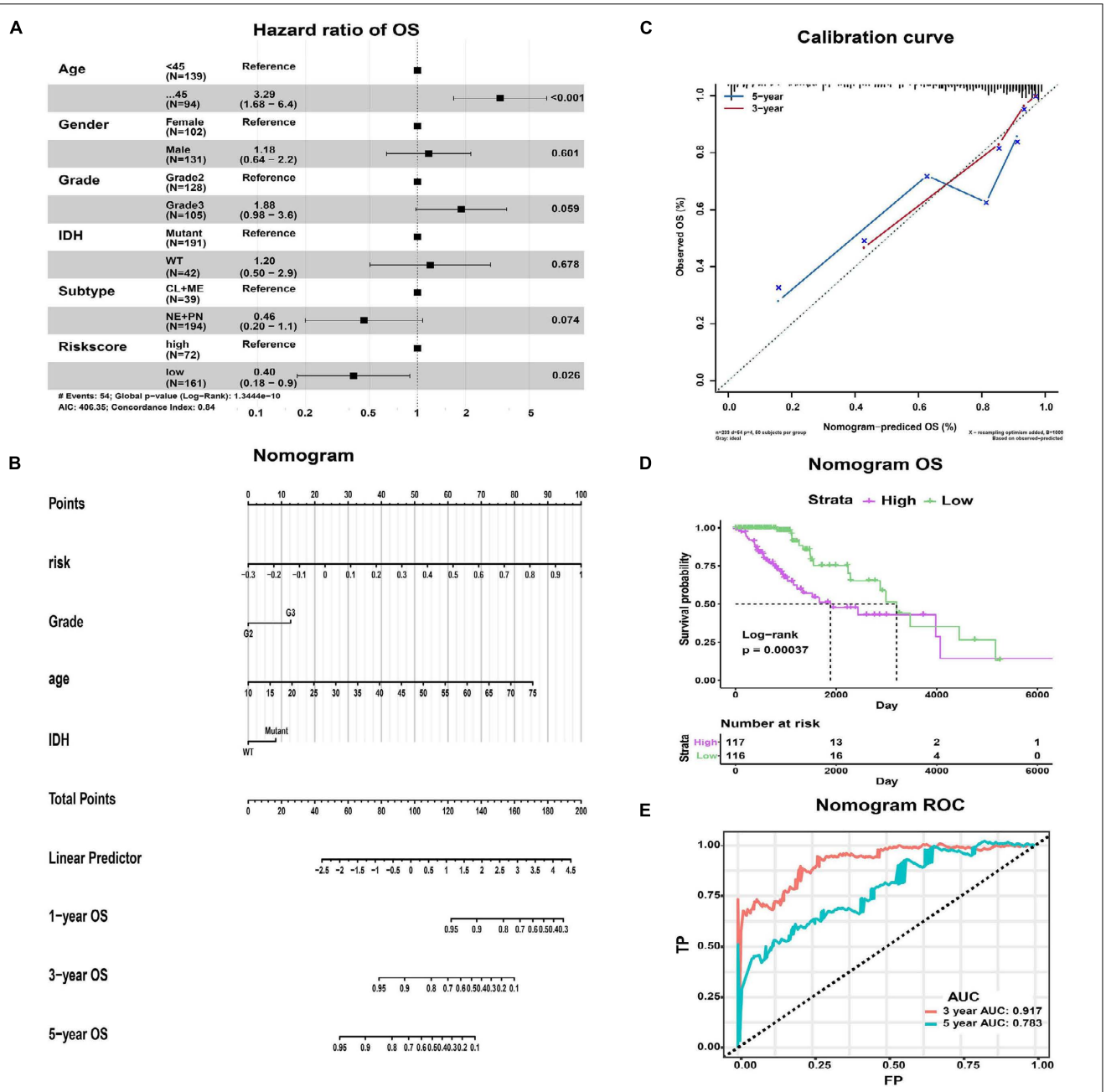


FIGURE 6 | The nomogram predicting the OS of LGG patients with epilepsy. **(A)** Forest map showing the results of multivariate Cox regression analysis. **(B)** The nomogram was used by summing the points of each variable. **(C)** The calibration curves showing the predicted and actual observed OS rates. **(D)** K-M plot showing the difference in OS between high- and low-risk groups. **(E)** The ROC curves and AUC values of the nomogram.

epilepsy. We found that the immune score was associated with survival, and patients with a high immune score had a poor prognosis. Moreover, the difference of immune cell infiltration between different immune score groups was observed, and a higher proportion of M2 macrophages appeared in high immune score patients. The M2 macrophage infiltration subverts the host's adaptive immune response and fosters a tumor milieu ripe for angiogenesis, migration, and metastasis in the glioma (Michelson et al., 2016; Liu et al., 2018). These findings indicate that the immune microenvironment affects the prognosis of LGG patients with epilepsy.

Further, three prognostic DEGs (ABCC3, INA, and PDPN) were identified between high and low immune score groups in LGG patients with epilepsy. Hence, a prognostic immune-related gene signature was established. The ATP binding cassette subfamily C member 3 (ABCC3) gene belongs to the ABCC gene family encoding the ABC transporters and participates in transporting ATP-dependent substances across lipid membranes

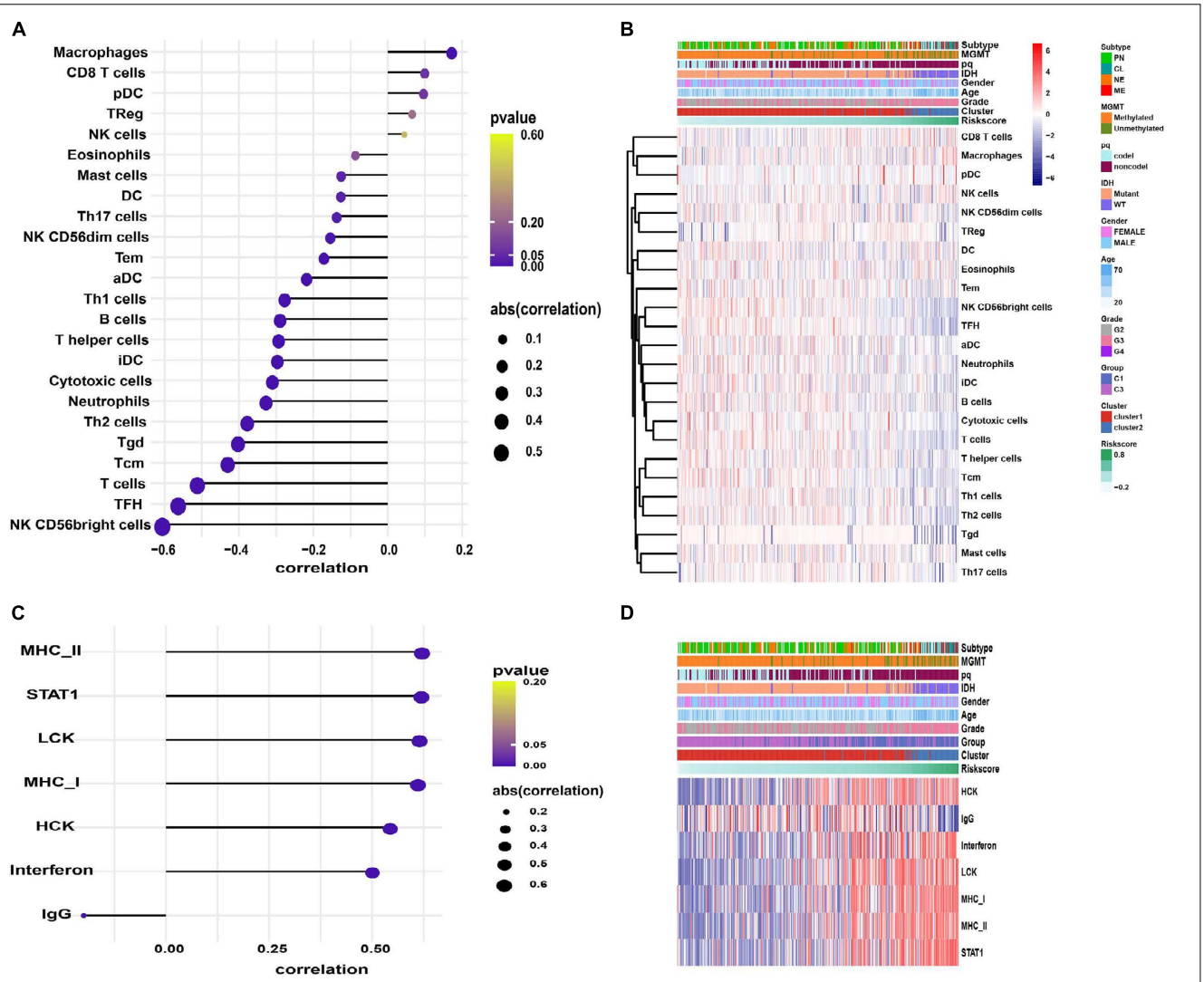


FIGURE 7 | The GSVA results. **(A,C)** Correlograms showing the correlation between immune cell infiltration, inflammation activities, and risk score. **(B,D)** Heatmaps displaying the activities of immune cells panel **(B)** and inflammation panel **(D)**.

(Chen and Tiwari, 2011; Mao et al., 2019). A high expression level of the ABCC3 gene is associated with various tumors' poor prognosis, such as gastric cancer, breast cancer, and pancreatic cancer (Adamska et al., 2019; Mao et al., 2019; Wang et al., 2020). Podoplanin (PDPN) is a glycoprotein receptor that affects cell behavior by interacting with other proteins and has various physiological functions (Krishnan et al., 2018). The PDPN is crucial for the development of the lymphatic system, lungs, and heart (Astarita et al., 2012). It is also a malignant tumor promoter, which promotes tumor migration, invasion, lymphangiogenesis, and angiogenesis (Astarita et al., 2012; Quintanilla et al., 2019); Its up-regulation in tumor tissues is correlated with the poor prognosis of mesothelioma, melanoma, and squamous cell carcinoma (Watanabe et al., 2020). Alpha-internexin (INA) gene encodes a neurofilament interacting protein and is involved in neuronal development and axonal outgrowth (Ducray et al., 2011; Rajmohan et al., 2020). The reduced/loss of its expression is related to tumor invasion and suggests a poor prognosis in patients with gastroenteropancreatic neuroendocrine neoplasms (GEP-NENs) (Wang et al., 2018). In this study, the high expression of ABCC3 and PDPN genes was found in the high immune score group and was related to the poor prognosis of LGG patients with epilepsy. And the high expression of INA gene was observed in patients with low immune score and suggests a better prognosis. We also found that the expression level of signature genes was regulated by DNA methylation in LGG patients with epilepsy. DNA methylation is associated with gene silencing and tumorigenesis, and tumors are often accompanied by hypomethylation of oncogenes and hypermethylation of tumor suppressor genes (Morgan et al., 2018). This also provides new potential targets for the regulation of glioma epigenetics.

We established a risk score model based on the above signature genes, and the high risk-score suggested a poor prognosis. The risk score was determined as a factor with an

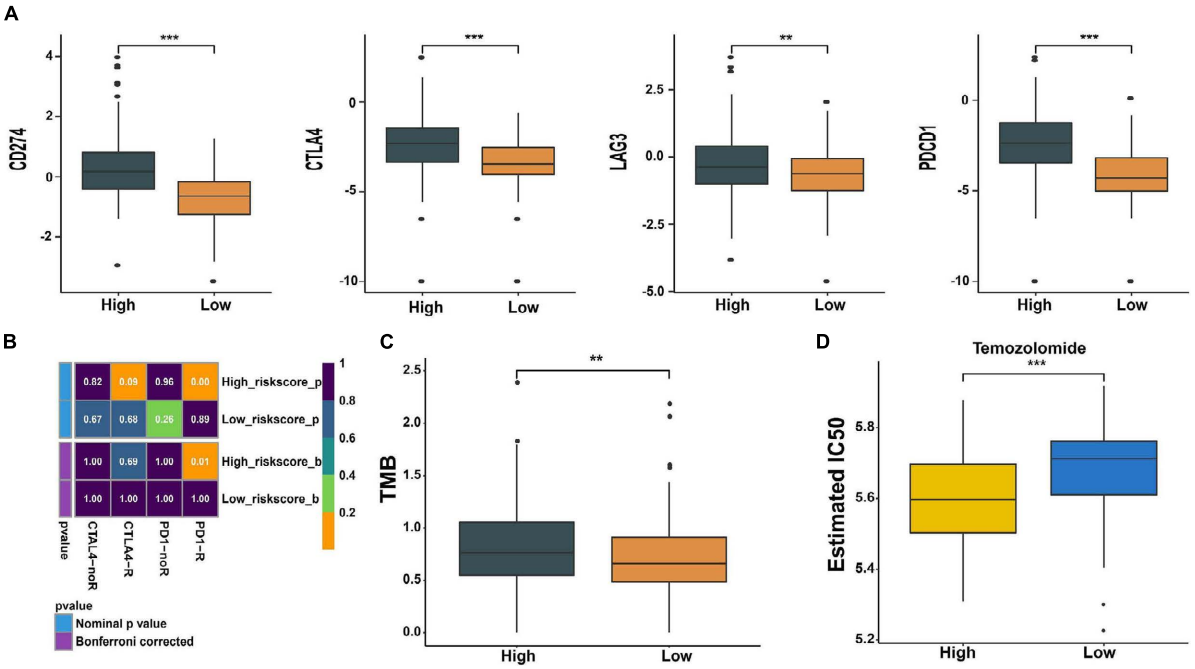


FIGURE 8 | Prediction of response to immunotherapy and chemotherapy. **(A)** Expression of immune checkpoints in high- and low-risk groups. **(B)** The analysis of response to immunotherapy. **(C)** Comparison of TMB between the two groups. **(D)** The estimated IC50 value of temozolomide (TMZ) in the two groups. $^{**}p < 0.01$, $^{***}p < 0.001$.

independent prognostic value. The 1p/19q non−codel, IDH-wildtype, classical and mesenchymal subtype, MGMT promotor unmethylated, and higher WHO grade, which are the malignant clinicopathological features in glioma (Verhaak et al., 2010; Weller et al., 2012), were associated with a higher risk score. In addition, the frequency of IDH1, CIC, and IDH2 mutations suggested a positive prognosis (Appin and Brat, 2014; Miller et al., 2017; Liu Z. et al., 2019) were lower in high-risk patients, while the malignancy driving mutations (EGFR, NF1, and PTEN) (Appin and Brat, 2014; Asif et al., 2019) were more frequent. And the deletion peaks of tumor suppressor genes (CDKN2A, CDKN2B) and amplification peaks of oncogenes (EGFR, CDK4, PDGFRA, PIK3C2B) (Rao et al., 2010) were also detected in patients with high risk. These results suggest that this gene signature is a reliable prognostic predictor for LGG patients with epilepsy. Additionally, the nomogram constructed by integrating the clinicopathological features and risk score had a better prognostic prediction ability, and further improves the clinical practicality.

The relationship between the immune microenvironment and gene signature was further explored. The high- and low-risk patients had different immune cell infiltration and inflammatory profiles. Macrophage infiltration level was higher in high-risk patients, while low-risk patients had higher B cells, T cells, Tgd, Th1 cells, T helper cells, Th2 cells, TFH, Tem, Tcm, Th17 cells, NK CD56bright cells, NK CD56dim cells, mast cells, iDC, aDC, DC, cytotoxic cells, and neutrophils infiltration, especially the NK CD56bright cells. Macrophages are recruited to the glioma environment and create a supportive stroma for neoplastic cell invasion and expansion through the release of various growth factors and cytokines, and the number of macrophages increases with increasing malignancy grade (Hambardzumyan et al., 2016). The CD56bright cells are a functional subset of NK cells, which are characterized by low cytotoxicity and high cytokine production, and play an immunoregulatory role (Shevtsov and Multhoff, 2016). The infiltration of NK CD56bright cells is negatively correlated with the malignant degree of gastric carcinomas, lung tumors, and papillary thyroid cancer (Gogali et al., 2013). However, the effect of NK CD56bright cells on glioma has not been sufficiently explored. The infusion of IL15, which is required for survival, expansion, and activation of NK cells, significantly expands the subpopulation of CD56 bright NK Cells compared with the CD56dim NK Cells (Dubois et al., 2017), another functional subset of NK cells with high cytotoxic activity and low cytokine release properties (Shevtsov and Multhoff, 2016). Moreover, infusion of IL-15 in glioma model mice significantly increases the infiltration of NK cells into tumor and reduces tumor growth (Garofalo et al., 2015). Combined with the results of this study, we believe that NK CD56bright cells play an important role in glioma immunity, and its infiltration improves the prognosis. In addition, the high-risk patients had more active inflammatory responses related to T cells and macrophages, while patients with low risk had more active inflammatory response associated with B cells. We further explored the mechanism of signature genes involved in the regulation of the immune microenvironment, and found that the signature genes participated in the regulation of IL6-JAK-STAT3 signaling, IFN-α response, IFN-γ response, and TNFA-signaling-via-NFKB

pathways, which are onco-immunological pathways (Wang et al., 2009; De Simone et al., 2015; Mostofa et al., 2017; Silginer et al., 2017; Zhu et al., 2020).

Immune checkpoint blockade is a promising treatment for glioma patients. However, not all patients can benefit from it (Hodges et al., 2017). Therefore, we wonder whether this gene signature can be used as a predictive marker for the clinical response to immune checkpoint blocking therapy in LGG patients with epilepsy. The results showed that high-risk patients were more likely to benefit from anti-PD1 treatment. This may be related to their high expression of PD1/PD-L1, and the active IL6-STAT3 signaling and IFN-γ response pathways in high-risk patients can promote the expression of PD1/PD-L1 (Zhang et al., 2016; Qian et al., 2018; Litak et al., 2019). Also, a higher TMB was found in high-risk patients, and high TMB is currently considered to indicate a relatively better response rate to immunotherapy (Hellmann et al., 2018). We also found that high-risk patients might be more sensitive to TMZ therapy, which may be related to their active IFN pathway down-regulating MGMT level and increasing the sensitivity of glioma cells to TMZ (Shen et al., 2015).

In general, we established an immune-related gene signature, which can be used to predict the prognosis of LGG patients with epilepsy, and infer their immune infiltration status, response to immunotherapy and chemotherapy. Moreover, this study also deepens our understanding of the immune microenvironment in glioma. Of course, there are some limitations in this study. First of all, only TCGA data was used for analysis in this paper, and the sample size was relatively limited. It may be necessary to further expand the sample size and verify the results of this study in other databases. Secondly, due to the difficulty in obtaining clinical samples from LGG patients with epilepsy, no further *in vitro* experiments were conducted in this study to verify the results. In the end, the specific mechanisms of signature genes involved in glioma immunity need further exploration. The *p* values obtained from the analysis in this paper were summarized in **Supplementary Table 7**.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/.

AUTHOR CONTRIBUTIONS

ZX and QC conceived and designed the research. QC and WD drafted and revised the manuscript. ZX, QC, and WD analyzed the data. KL, CL, SH, WY, BY, and HC acquired and interpreted the data. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Science Foundation of AMHT Group (No. 2020YK10), the Key Research and Development Programs of Hunan Provincal Science and Technology Department (No. 2017SK2081), China Postdoctoral Science Foundation (No. 2018M633002), Hunan Provincial Natural Science Foundation of China (No. 2018JJ3838), Hunan Provincial Health Committee Foundation of China (C2019186), and Xiangya Hospital Central South University Postdoctoral Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.686909/full#supplementary-material

REFERENCES

Adamska, A., Domenichini, A., Capone, E., Damiani, V., Akkaya, B. G., Linton, K. J., et al. (2019). Pharmacological inhibition of ABCC3 slows tumour progression in animal models of pancreatic cancer. *J. Exp. Clin. Cancer Res.* 38:312. doi: 10.1186/s13046-019-1308-7

Appin, C. L., and Brat, D. J. (2014). Molecular genetics of gliomas. *Cancer J.* 20, 66–72. doi: 10.1097/PPO.0000000000000020

Asif, S., Fatima, R., Krc, R., Bennett, J., and Raza, S. (2019). Comparative proteogenomic characterization of glioblastoma. *CNS Oncol.* 8:CNS37. doi: 10.2217/cns-2019-0003

Astarita, J. L., Acton, S. E., and Turley, S. J. (2012). Podoplanin: emerging functions in development, the immune system, and cancer. *Front. Immunol.* 3:283. doi: 10.3389/fimmu.2012.00283

Blümcke, I., Luyken, C., Urbach, H., Schramm, J., and Wiestler, O. D. (2004). An isomorphic subtype of long-term epilepsy-associated astrocytomas associated with benign prognosis. *Acta Neuropathol.* 107, 381–388. doi: 10.1007/s00401-004-0833-3

Chen, D. Y., Chen, C. C., Crawford, J. R., and Wang, S. G. (2018). Tumor-related epilepsy: epidemiology, pathogenesis and management. *J. Neurooncol.* 139, 13–21. doi: 10.1007/s11060-018-2862-0

Chen, Z.-S., and Tiwari, A. K. (2011). Multidrug resistance proteins (MRPs/ABCCs) in cancer chemotherapy and genetic diseases. *FEBS J.* 278, 3226–3245. doi: 10.1111/j.1742-4658.2011.08235.x

Cohen, A. L., Holmen, S. L., and Colman, H. (2013). IDH1 and IDH2 mutations in gliomas. *Curr. Neurol. Neurosci. Rep.* 13:345. doi: 10.1007/s11910-013-0345-4

Danfors, T., Ribom, D., Berntsson, S. G., and Smits, A. (2009). Epileptic seizures and survival in early disease of grade 2 gliomas. *Eur. J. Neurol.* 16, 823–831. doi: 10.1111/j.1468-1331.2009.02599.x

De Simone, V., Franzè, E., Ronchetti, G., Colantoni, A., Fantini, M. C., Di Fusco, D., et al. (2015). Th17-type cytokines, IL-6 and TNF-α synergistically activate STAT3 and NF-kB to promote colorectal cancer cell growth. *Oncogene* 34, 3493–3503. doi: 10.1038/onc.2014.286

Dubois, S., Conlon, K. C., Müller, J. R., Hsu-Albert, J., Beltran, N., Bryant, B. R., et al. (2017). IL15 infusion of cancer patients expands the subpopulation of cytotoxic CD56 NK cells and increases NK-cell cytokine release capabilities. *Cancer Immunol. Res.* 5, 929–938. doi: 10.1158/2326-6066.CIR-17-0279

Ducray, F., Mokhtari, K., Crinière, E., Idbaih, A., Marie, Y., Dehais, C., et al. (2011). Diagnostic and prognostic value of alpha internexin expression in a series of 409 gliomas. *Eur. J. Cancer* 47, 802–808. doi: 10.1016/j.ejca.2010.11.031

Garofalo, S., D'Alessandro, G., Chece, G., Brau, F., Maggi, L., Rosa, A., et al. (2015). Enriched environment reduces glioma growth through immune and non-immune mechanisms in mice. *Nat. Commun.* 6:6623. doi: 10.1038/ncomms7623

Geeleher, P., Cox, N., and Huang, R. S. (2014). pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS One* 9:e107468. doi: 10.1371/journal.pone.0107468

Gieryng, A., Pszczolkowska, D., Walentynowicz, K. A., Rajan, W. D., and Kaminska, B. (2017). Immune microenvironment of gliomas. *Lab Invest.* 97, 498–518. doi: 10.1038/labinvest.2017.19

Gogali, F., Paterakis, G., Rassidakis, G. Z., Liakou, C. I., and Liapi, C. (2013). CD3(-)CD16(-)CD56(bright) immunoregulatory NK cells are increased in the tumor microenvironment and inversely correlate with advanced stages in patients with papillary thyroid cancer. *Thyroid* 23, 1561–1568. doi: 10.1089/thy.2012.0560

Goldstein, E. D., and Feyissa, A. M. (2018). Brain tumor related-epilepsy. *Neurol. Neurochir. Pol.* 52, 436–447. doi: 10.1016/j.pjnns.2018.06.001

Hambardzumyan, D., Gutmann, D. H., and Kettenmann, H. (2016). The role of microglia and macrophages in glioma maintenance and progression. *Nat. Neurosci.* 19, 20–27. doi: 10.1038/nn.4185

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

Hellmann, M. D., Ciuleanu, T.-E., Pluzanski, A., Lee, J. S., Otterson, G. A., Audigier-Valette, C., et al. (2018). Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *N. Engl. J. Med.* 378, 2093–2104. doi: 10.1056/NEJMoa1801946

Hodges, T. R., Ott, M., Xiu, J., Gatalica, Z., Swensen, J., Zhou, S., et al. (2017). Mutational burden, immune checkpoint expression, and mismatch repair in glioma: implications for immune checkpoint immunotherapy. *Neuro Oncol.* 19, 1047–1057. doi: 10.1093/neuonc/nox026

Huberfeld, G., and Vecht, C. J. (2016). Seizures and gliomas–towards a single therapeutic approach. *Nat. Rev. Neurol.* 12, 204–216. doi: 10.1038/nrneurol.2016.26

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24, 1550–1558. doi: 10.1038/s41591-018-0136-1

Jones, D. T. W., Bandopadhayay, P., and Jabado, N. (2019). The power of human cancer genetics as revealed by low-grade gliomas. *Annu. Rev. Genet.* 53, 483–503. doi: 10.1146/annurev-genet-120417-031642

Krishnan, H., Rayes, J., Miyashita, T., Ishii, G., Retzbach, E. P., Sheehan, S. A., et al. (2018). Podoplanin: an emerging cancer biomarker and therapeutic target. *Cancer Sci.* 109, 1292–1299. doi: 10.1111/cas.13580

Krzyszkowski, T., Czepko, R., and Betlej, M. (2004). [Prognostic value of epileptic seizures in patients with cerebral gliomas]. *Ann. Acad. Med. Stetin.* 50, 35–40.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Litak, J., Mazurek, M., Grochowski, C., Kamieniak, P., and Roliński, J. (2019). PD-L1/PD-1 Axis in glioblastoma multiforme. *Int. J. Mol. Sci.* 20:5347. doi: 10.3390/ijms20215347

Liu, B., Liu, J., Liu, K., Huang, H., Li, Y., Hu, X., et al. (2019). A prognostic signature of five pseudogenes for predicting lower-grade gliomas. *Biomed. Pharmacother.* 117:109116. doi: 10.1016/j.biopha.2019.109116

Liu, Z., Kuang, W., Zhou, Q., and Zhang, Y. (2018). TGF-β1 secreted by M2 phenotype macrophages enhances the stemness and migration of glioma cells via the SMAD2/3 signalling pathway. *Int. J. Mol. Med.* 42, 3395–3403. doi: 10.3892/ijmm.2018.3923

Liu, Z., Liu, H., Liu, Z., and Zhang, J. (2019). Oligodendroglial tumours: subventricular zone involvement and seizure history are associated with CIC mutation status. *BMC Neurol.* 19:134. doi: 10.1186/s12883-019-1362-y

Lote, K., Stenwig, A. E., Skullerud, K., and Hirschberg, H. (1998). Prevalence and prognostic significance of epilepsy in patients with gliomas. *Eur. J. Cancer* 34, 98–102. doi: 10.1016/s0959-8049(97)00374-2

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Ma, Q., Long, W., Xing, C., Chu, J., Luo, M., Wang, H. Y., et al. (2018). Cancer stem cells and immunosuppressive microenvironment in glioma. *Front. Immunol.* 9:2924. doi: 10.3389/fimmu.2018.02924

Mao, X., He, Z., Zhou, F., Huang, Y., and Zhu, G. (2019). Prognostic significance and molecular mechanisms of adenosine triphosphate-binding cassette subfamily C members in gastric cancer. *Medicine (Baltimore)* 98:e18347. doi: 10.1097/MD.0000000000018347

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* 28, 1747–1756. doi: 10.1101/gr.239244.118

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12:R41. doi: 10.1186/gb-2011-12-4-r41

Michelson, N., Rincon-Torroella, J., Quiñones-Hinojosa, A., and Greenfield, J. P. (2016). Exploring the role of inflammation in the malignant transformation of low-grade gliomas. *J. Neuroimmunol.* 297, 132–140. doi: 10.1016/j.jneuroim.2016.05.019

Miller, J. J., Shih, H. A., Andronesi, O. C., and Cahill, D. P. (2017). Isocitrate dehydrogenase-mutant glioma: evolving clinical and therapeutic implications. *Cancer* 123, 4535–4546. doi: 10.1002/cncr.31039

Morgan, A. E., Davies, T. J., and Mc Auley, M. T. (2018). The role of DNA methylation in ageing and cancer. *Proc. Nutr. Soc.* 77, 412–422. doi: 10.1017/S0029665118000150

Mostofa, A. G. M., Punganuru, S. R., Madala, H. R., Al-Obaide, M., and Srivenugopal, K. S. (2017). The Process And Regulatory Components Of Inflammation In Brain Oncogenesis. *Biomolecules* 7:34. doi: 10.3390/biom7020034

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* 12, 252–264. doi: 10.1038/nrc3239

Qian, J., Wang, C., Wang, B., Yang, J., Wang, Y., Luo, F., et al. (2018). The IFN-γ/PD-L1 axis between T cells and tumor microenvironment: hints for glioma anti-PD-1/PD-L1 therapy. *J. Neuroinflammation* 15:290. doi: 10.1186/s12974-018-1330-2

Quintanilla, M., Montero-Montero, L., Renart, J., and Martín-Villar, E. (2019). Podoplanin in inflammation and cancer. *Int. J. Mol. Sci.* 20:707. doi: 10.3390/ijms20030707

Rajmohan, K. S., Sugur, H. S., Shwetha, S. D., Pandey, P., Arivazhagan, A., and Santosh, V. (2020). Alpha internexin: a surrogate marker for 1p/19q codeletion and prognostic marker in anaplastic (WHO grade III) gliomas. *Neurol. India* 68, 832–837. doi: 10.4103/0028-3886.293453

Rao, S. K., Edwards, J., Joshi, A. D., Siu, I. M., and Riggins, G. J. (2010). A survey of glioblastoma genomic amplifications and deletions. *J. Neurooncol.* 96, 169–179. doi: 10.1007/s11060-009-9959-4

Rody, A., Holtrich, U., Pusztai, L., Liedtke, C., Gaetje, R., Ruckhaeberle, E., et al. (2009). T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers. *Breast Cancer Res.* 11:R15. doi: 10.1186/bcr2234

Shen, D., Guo, C.-C., Wang, J., Qiu, Z.-K., Sai, K., Yang, Q.-Y., et al. (2015). Interferon-α/β enhances temozolomide activity against MGMT-positive glioma stem-like cells. *Oncol. Rep.* 34, 2715–2721. doi: 10.3892/or.2015.4232

Shevtsov, M., and Multhoff, G. (2016). Immunological and translational aspects of NK cell-based antitumor immunotherapies. *Front. Immunol.* 7:492. doi: 10.3389/fimmu.2016.00492

Silginer, M., Nagy, S., Happold, C., Schneider, H., Weller, M., and Roth, P. (2017). Autocrine activation of the IFN signaling pathway may promote immune escape in glioblastoma. *Neuro Oncol.* 19, 1338–1349. doi: 10.1093/neuonc/nox051

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Wang, H., Lathia, J. D., Wu, Q., Wang, J., Li, Z., Heddleston, J. M., et al. (2009). Targeting interleukin 6 signaling suppresses glioma stem cell survival and tumor growth. *Stem Cells* 27, 2393–2404. doi: 10.1002/stem.188

Wang, Y., Chen, Y., Li, X., Hu, W., Zhang, Y., Chen, L., et al. (2018). Loss of expression and prognosis value of alpha-internexin in gastroenteropancreatic

neuroendocrine neoplasm. *BMC Cancer* 18:691. doi: 10.1186/s12885-018-4449-8

Wang, Y., Zhang, X., Zhao, B., Xu, Z., and Lv, Y. (2020). Suspension state promotes drug resistance of breast tumor cells by inducing ABCC3 overexpression. *Appl. Biochem. Biotechnol.* 190, 410–422. doi: 10.1007/s12010-019-03084-0

Watanabe, N., Kidokoro, M., Tanaka, M., Inoue, S., Tsuji, T., Akatuska, H., et al. (2020). Podoplanin is indispensable for cell motility and platelet-induced epithelial-to-mesenchymal transition-related gene expression in esophagus squamous carcinoma TE11A cells. *Cancer Cell Int.* 20:263. doi: 10.1186/s12935-020-01328-2

Weller, M., Stupp, R., Hegi, M. E., van den Bent, M., Tonn, J. C., Sanson, M., et al. (2012). Personalized care in neuro-oncology coming of age: why we need MGMT and 1p/19q testing for malignant glioma patients in clinical practice. *Neuro Oncol.* 14(Suppl. 4), iv100–iv108. doi: 10.1093/neuonc/nos206

Wen, P., Gao, Y., Chen, B., Qi, X., Hu, G., Xu, A., et al. (2020). Pan-cancer analysis of radiotherapy benefits and immune infiltration in multiple human cancers. *Cancers (Basel)* 12:957. doi: 10.3390/cancers12040957

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26, 1572–1573. doi: 10.1093/bioinformatics/btq170

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612. doi: 10.1038/ncomms3612

Zhang, N., Zeng, Y., Du, W., Zhu, J., Shen, D., Liu, Z., et al. (2016). The EGFR pathway is involved in the regulation of PD-L1 expression via the IL-6/JAK/STAT3 signaling pathway in EGFR-mutated non-small cell lung cancer. *Int. J. Oncol.* 49, 1360–1368. doi: 10.3892/ijo.2016.3632

Zhu, C., Chen, X., Guan, G., Zou, C., Guo, Q., Cheng, P., et al. (2020). IFI30 is a novel immune-related target with predicting value of prognosis and treatment response in glioblastoma. *Onco Targets Ther.* 13, 1129–1143. doi: 10.2147/OTT.S237162

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a shared affiliation, though no other collaboration, with several of the authors WD, QC, and WY at the time of the review.

Copyright © 2021 Cheng, Duan, He, Li, Cao, Liu, Ye, Yuan and Xia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 16 July 2021
doi: 10.3389/fcell.2021.707938

Identification of Disease-Related 2-Oxoglutarate/Fe (II)-Dependent Oxygenase Based on Reduced Amino Acid Cluster Strategy

Jian Zhou[1†], Suling Bo[2†], Hao Wang[1], Lei Zheng[1], Pengfei Liang[1] and Yongchun Zuo[1]*

[1] State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, College of Life Sciences, Inner Mongolia University, Hohhot, China, [2] College of Computer and Information, Inner Mongolia Medical University, Hohhot, China

The 2-oxoglutarate/Fe (II)-dependent (2OG) oxygenase superfamily is mainly responsible for protein modification, nucleic acid repair and/or modification, and fatty acid metabolism and plays important roles in cancer, cardiovascular disease, and other diseases. They are likely to become new targets for the treatment of cancer and other diseases, so the accurate identification of 2OG oxygenases is of great significance. Many computational methods have been proposed to predict functional proteins to compensate for the time-consuming and expensive experimental identification. However, machine learning has not been applied to the study of 2OG oxygenases. In this study, we developed OGFE_RAAC, a prediction model to identify whether a protein is a 2OG oxygenase. To improve the performance of OGFE_RAAC, 673 amino acid reduction alphabets were used to determine the optimal feature representation scheme by recoding the protein sequence. The 10-fold cross-validation test showed that the accuracy of the model in identifying 2OG oxygenases is 91.04%. Besides, the independent dataset results also proved that the model has excellent generalization and robustness. It is expected to become an effective tool for the identification of 2OG oxygenases. With further research, we have also found that the function of 2OG oxygenases may be related to their polarity and hydrophobicity, which will help the follow-up study on the catalytic mechanism of 2OG oxygenases and the way they interact with the substrate. Based on the model we built, a user-friendly web server was established and can be friendly accessed at http://bioinfor.imu.edu.cn/ogferaac.

Keywords: 2-oxoglutarate/Fe (II)-dependent oxygenase, reduced amino acid cluster, machine learning, anova, incremental feature selection, 10-fold cross-validation test

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Leyi Wei,
Shandong University, China
Shihua Zhang,
Wuhan University of Science and Technology, China

****Correspondence:***
Yongchun Zuo
yczuo@imu.edu.cn

[†]These authors share first authorship

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *11 May 2021*
Accepted: *10 June 2021*
Published: *16 July 2021*

Citation:
Zhou J, Bo S, Wang H, Zheng L, Liang P and Zuo Y (2021) Identification of Disease-Related 2-Oxoglutarate/Fe (II)-Dependent Oxygenase Based on Reduced Amino Acid Cluster Strategy. Front. Cell Dev. Biol. 9:707938. doi: 10.3389/fcell.2021.707938

INTRODUCTION

2-Oxoglutarate/Fe (II)-dependent (2OG) oxygenases (EC:1.14.11), generally using nonheme iron as an active-site cofactor, promote oxidative decarboxylation of the substrate to produce carbon dioxide and succinic acid (Hausinger, 2004; Hewitson et al., 2005; Islam et al., 2018). 2OG oxygenases, which can catalyze many different oxidation reactions, are a superfamily with members

widely distributed in animals, plants, and microorganisms. In animals, their catalytic range includes hydroxylation and N-demethylation proceeding *via* hydroxylation; in plants and microbes, they affect a wider range, including hydroxylation, ring formations, cleavage, oxidation, rearrangements, desaturations, and halogenations (Farrow and Facchini, 2014; Kawai et al., 2014). The proteins of this superfamily can be divided into 2OG oxygenase domain-containing oxygenases and JmjC domain-containing oxygenases (Jia et al., 2017). **Figure 1** is a schematic diagram of the structure of 2OG oxygenases.

Due to the diversity of 2OG oxygenases and the wide range of binding substrates, these oxygenases play an important role in physiology and have high therapeutic value and therapeutic potential as targets in cancer and many other diseases (Rose et al., 2011). For example, the protein containing the JmjC domain (JMJD6) is located in the nucleus that catalyzes lysine hydroxylation and arginine demethylation of histone and non-histone peptides (Chang et al., 2007; Liu et al., 2013). JMJD6 promotes cell proliferation and migration *in vitro* and accelerates tumor growth *in vivo*, so it may become an attractive target for a new generation of anticancer drugs (Lin et al., 2006; Lee et al., 2012). Prolyl 4-hydroxylase (P4H) plays a vital role in the synthesis of collagen and the regulation of oxygen homeostasis. Collagen P4Hs are considered to be attractive targets for drug inhibitors and involved in the treatment of fibrotic diseases and cancer metastasis (Vasta and Raines, 2018). Hypoxia-inducible transcription factor-prolyl 4-hydroxylase inhibitors are believed to have beneficial effects in the treatment of diseases such as myocardial infarction, stroke, peripheral vascular disease, diabetes, and severe anemias (Myllyharju, 2008; Liao and Zhang, 2020). ALKB homologs (ALKBH) homologs can regulate the physiological and pathological processes of cardiovascular diseases (CVDs), which have great potential in the development of CVD drugs and are expected to become a potential target for the treatment of CVD (Xiao et al., 2020). The change in the catalytic activity or expression level of lysine demethylases (KDMs) is closely related to many diseases, including cancer genesis and progression, neurological disorders, inflammatory and immune disorders, metabolic diseases, and regenerative diseases. Modulators/inhibitors of KDMs may be used as new treatments for cancer and other diseases (Arifuzzaman et al., 2020). Therefore, it is particularly meaningful to predict 2OG oxygenases and find more potential 2OG oxygenases. Since the identification of 2OG oxygenase is time-consuming and expensive, machine learning is an effective and fast method to predict it.

In the past, many machine learning methods for the prediction of metal ion-binding proteins have achieved excellent results. For example, Lin et al. (2006) applied the sequence information used by support vector machine (SVM) to predict the metal ion-binding protein and got a relatively marvelous prediction result. Mohan et al. (2010) used a set of physicochemical parameters of metal ion-binding proteins encoded by the three genes *CzcA*, *CzcB*, and *CzcD* as the training set of the supervised classifier, establishing a model to identify metal ion-binding proteins from unknown proteins. Valasatava et al. (2016) developed MetalPredator, a web server used to predict iron–sulfur cluster-binding proteomes, and it featured an excellent performance in terms of precision and recall. Many studies have also achieved good results in the prediction of metal ion-binding sites, including iron ion-binding sites (Liu and Hu, 2011; Liou et al., 2014), zinc ion-binding sites (Shu et al., 2008; Chen et al., 2013; Yan et al., 2019), copper ion binding sites (Levy et al., 2009; Brylinski and Skolnick, 2011). The above indicate that machine learning is suitable for the application of metal ion-binding proteins (Valasatava et al., 2016). Not only that, studies have shown that using the reduced amino acid cluster (RAAC) strategy to predict the types of proteins can reduce noise and achieve higher accuracy (Zheng et al., 2019). In the prediction of human and nonhuman enzymes (Wang H. et al., 2021), ion channel-targeted conotoxins (Sun et al., 2020), plasmodium secretory protein (Zhang et al., 2020), and defensin peptides (Zuo et al., 2019), the method of reduced amino acid has shown superior performance.

In this study, we established a prediction model for 2OG oxygenases based on SVM, which can effectively identify 2OG oxygenases. A new feature representation scheme (amino acid reduction cluster) was involved in this work. The RAAC strategy can greatly decrease the complexity of protein sequences and extremely reduce the use of computer memory (Zuo et al., 2017; Zheng et al., 2019). The workflow of constructing the OGFE_RAAC is shown in **Figure 2**. Firstly, an objective dataset was established, which contains 734 2OG oxygenases and 385,381 non-2OG oxygenases from the InterPro database. Subsequently, reduced amino acid composition combined with K-mer strategy was used to represent sequence features, and the optimal one was selected from 673 reduction schemes (Zuo et al., 2015). At the same time, we obtained the best feature combination through analysis of variance (ANOVA) combined with incremental feature selection (IFS) and applied SVM to establish the model. The results of 10-fold cross-validation and independent test set showed that OGFE_RAAC could accurately predict 2OG oxygenases.

MATERIALS AND METHODS

Dataset

The 2OG oxygenase superfamily can be classified into 2OG oxygenase domain-containing oxygenases and JmjC domain-containing oxygenases, so we collected all the verified 734 proteins of these two domains in the IPR number (IPR005123 and IPR003347) of the InterPro public database as a positive sample. Concurrently, 385381 protein data verified by SwissProt were gathered as negative samples, which is the manual annotation and review part of UniProt. Then, CD-HIT (Huang et al., 2010) was used to remove sequences with a similarity of more than 50% (Zou et al., 2020), and 480 samples are selected as the training set (Fu et al., 2012). We chose 150 samples from the rest as the test set, and the dataset was named 2OG-SwissProt. For the purpose of getting a better model, we also used iron-binding protein as a negative sample to construct a dataset. We acquired 593 iron-binding proteins (GO:0005506, 2OG oxygenase proteins removed) from the InterPro public database and processed

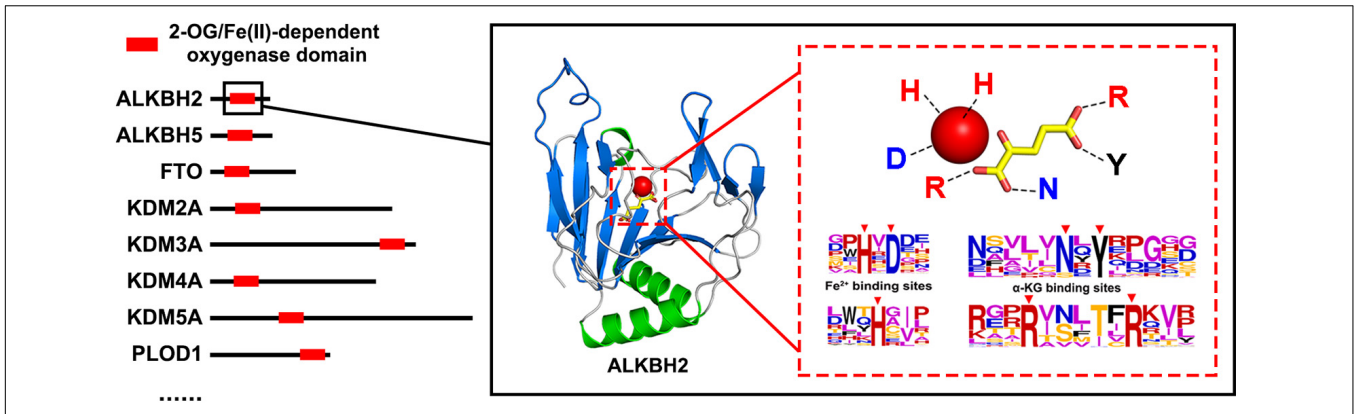


FIGURE 1 | Schematic diagram of the structure of 2-oxoglutarate/Fe (II)-dependent (2OG) oxygenase.

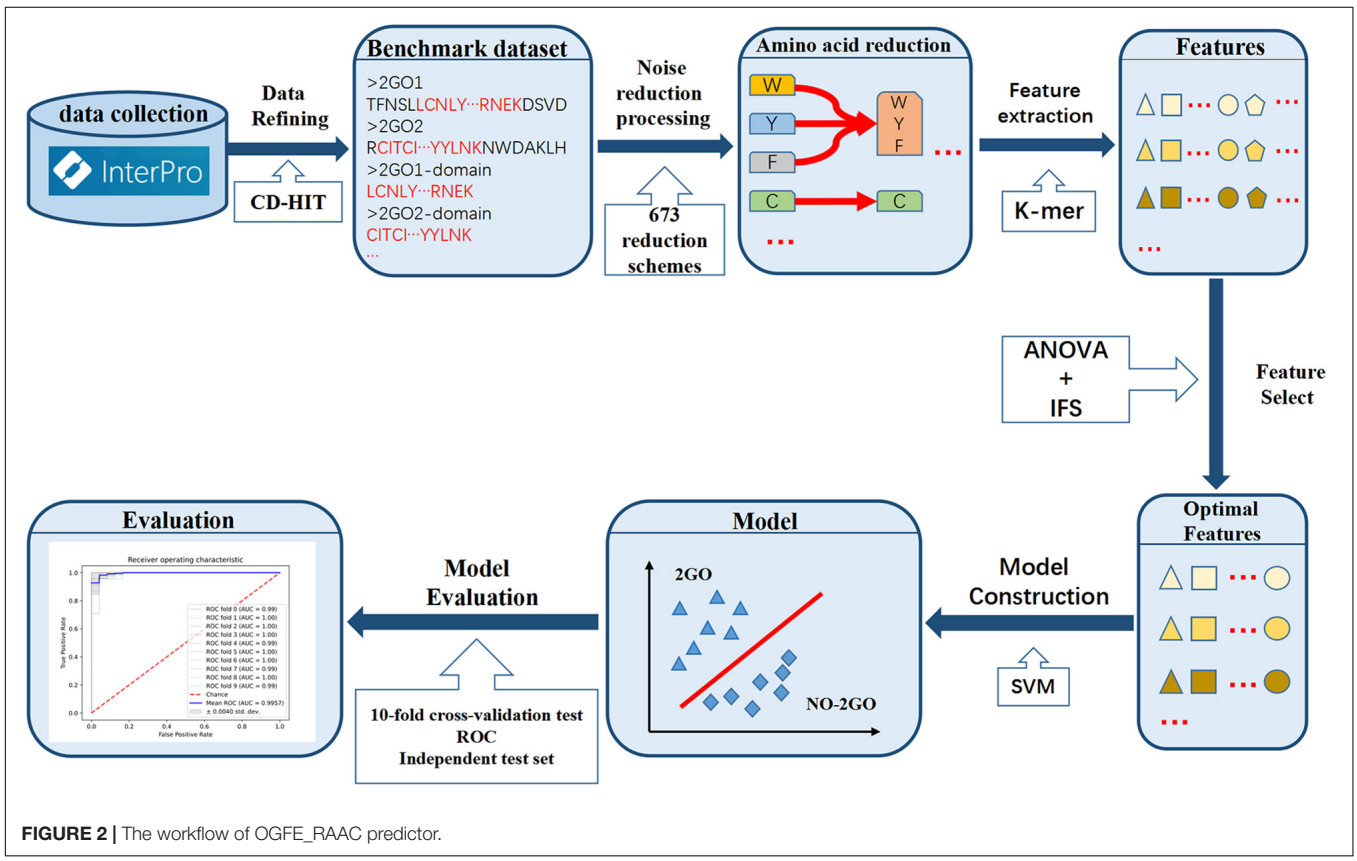


FIGURE 2 | The workflow of OGFE_RAAC predictor.

them in the same way as the 2OG-SwissProt dataset to obtain 471 training set samples and 159 test set samples; the dataset was named 2OG-Fe.

For further research, we manually extracted the domain sequences of 2OG oxygenase and iron-binding proteins. The processing method is the same as the above; in order to better verify the prediction results, we used CD-HIT processing sequence similarity less than 50% as the training set and the rest as the independent test set. Among them, 1,036 samples constitute an independent test set, 621 positive samples and 415 negative samples; 283 samples constitute a training set, 113 positive samples and 170 negative samples. This dataset was named 2OG-domain (**Table 1**).

Reduce Protein Sequence

Under normal circumstances, protein is composed of 20 natural amino acids. We combine amino acids with similar characteristics based on the physicochemical properties and atomic arrangement of amino acids. For instance, using fuzzy clustering technology and matrices cluster amino acids and interpret the sequence in a new encoding method (Georgiou et al., 2009; Zuo and Li, 2009). The strategy of RAACs can

TABLE 1 | Data composition of each dataset.

Dataset	Group	Training set	Test set
2OG-SwissProt	Positive	240	75
	Negative	240	75
2OG-Fe	Positive	240	75
	Negative	231	84
2OG-domain	Positive	113	621
	Negative	170	415

effectively reduce the complexity of the sequence and improve computational efficiency. In the study, we used 673 amino acid reduction schemes generated by 74 types to predict 2OG oxygenases, and each type has a reduced size of 2–19 (Zuo et al., 2019; Zheng et al., 2020).

Extract Features Based on K-mer

The typical K-mer (N-peptide) composition can effectively dig out the detailed information of the amino acid composition of the sequence (Zhu et al., 2019; Jaillard et al., 2020). We use K-mer ($K = 1, 2, 3$) to extract amino acid sequence information. Due to the limited memory, the maximum K value is 3, and a total of 20^K features can be obtained according to the original amino acid composition. The composition of K-mer ($K = 2$) can be expressed as follows:

$$P = R_1R_2R_3 \cdots R_{L-1}R_L \tag{1}$$

$$F = \left[d_1, d_2, \cdots d_{400}\right]^T \tag{2}$$

Here, R_i represents the i-th residue of the 2OG oxygenases. L represents the total length of the amino acid sequence. d_i ($i = 1$, 2,..., 400) is the i-th dipeptide in the 400-amino acid combination, and T means transposition operator. The d_i can be calculated as follows:

$$d_i = n_i \Big/ \sum_{i=1}^{400} n_i \tag{3}$$

Here, n_i denotes the number of the i-th dipeptide. Combined with RAAC strategy, the feature extraction method can be expressed as follows:

$$F = \left[P_{1,1}^1, P_{1,2}^2, \ldots, P_{i,j}^k, \ldots, P_{T,C}^N\right] \tag{4}$$

where $P_{i,j}^k$ denotes the method of the N-peptide with different RAAC descriptors (N-peptide). N denotes the N-peptide. T denotes the type of different amino acid alphabets, and C denotes the cluster of the reduced amino acid alphabet. The parameters of the above equation can be limited as follows:

$$\begin{cases} 1 \le k \le N, N = [1, 2, 3] \\ 1 \le i \le T, T = [1, 2, \ldots, 74] \\ 1 \le j \le C, C = [2, 3, \ldots, 19] \end{cases} \tag{5}$$

Support Vector Machine

Support vector machine is a machine learning model that classifies data according to supervised learning methods and has been widely used in bioinformatics (Beer, 2017; Huang et al., 2018; Manavalan et al., 2018; Meng et al., 2020; Tahir and Idris, 2020). There are four types of kernel function, including linear functions, polynomial functions, S-shaped functions, and radial basis functions (RBFs). In the past predictions of proteins, the RBF kernel function had better performance, and we have verified that the RBF kernel function has better performance in our model through the calculation and comparison of the four kernel functions. Accordingly, we used the SVM package with RBF kernel for the classifier, which can be obtained from https://www.csie.ntu.edu.tw/~cjlin/libsvm (Chang and Lin, 2011). The libsvm package provides a grid search program to optimize the parameters C and γ. The kernel parameter γ and the regularization parameter C are used to adjust the SVM model to obtain the best performance. The selection ranges of C and γ are as follows:

$$2^{-5} < C < 2^{15} \tag{6}$$

$$2^{-15} < \gamma < 2^3 \tag{7}$$

Feature Screening

The initial features extracted by K-mer are exclusive features, not the optimal combination of features (Zou et al., 2016; He et al., 2020). ANOVA is a popular feature selection method that can help us measure the weight value of each feature (Saeys et al., 2007; Tang et al., 2018). Then, we used IFS to determine the dimensionality of the best feature set according to the feature weights obtained by the ANOVA. The ANOVA equations are as follows:

$$F = \frac{S_x^2}{S_y^2} \tag{8}$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{9}$$

$$S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2 \tag{10}$$

where F is the variance value of the feature. S_x^2 is the sample variance between groups. S_y^2 denotes the sample variance within groups.

Performance Evaluation

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling (K-fold cross-validation) test, and jackknife test. However, among the three cross-validation methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors (Chou and Shen, 2008; Chou, 2011; Chou et al., 2012; Zhang et al., 2021). However, since the current study would involve feature selection as described above, to reduce the computational time, the 10-fold cross-validation test and independent dataset test would be adopted

as done by many investigators using SVM as the prediction engine. The performance can be measured in term of Sensitivity (Sn), Specificity (Sp), F1 score, Matthew's correlation coefficient (MCC), and Accuracy (Acc; Li et al., 2020; Shen and Zou, 2020; Yang et al., 2021), which are expressed as follows:

$$\mathrm{Sn} = \frac{TP}{TP + FN} \tag{11}$$

$$\mathrm{Sp} = \frac{TN}{TN + FP} \tag{12}$$

$$F1\ \mathrm{score} = \frac{2TP}{2TP + FP + FN} \tag{13}$$

$$\mathrm{Acc} = \frac{TP + TN}{TP + FN + TN + FP} \tag{14}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TN + FN)\,(TP + FN)\,(TN + FP)}} \tag{15}$$

where *TP*, *TN*, *FP*, and *FN* represent true-positive, true-negative, false-positive, and false-negative samples, respectively.

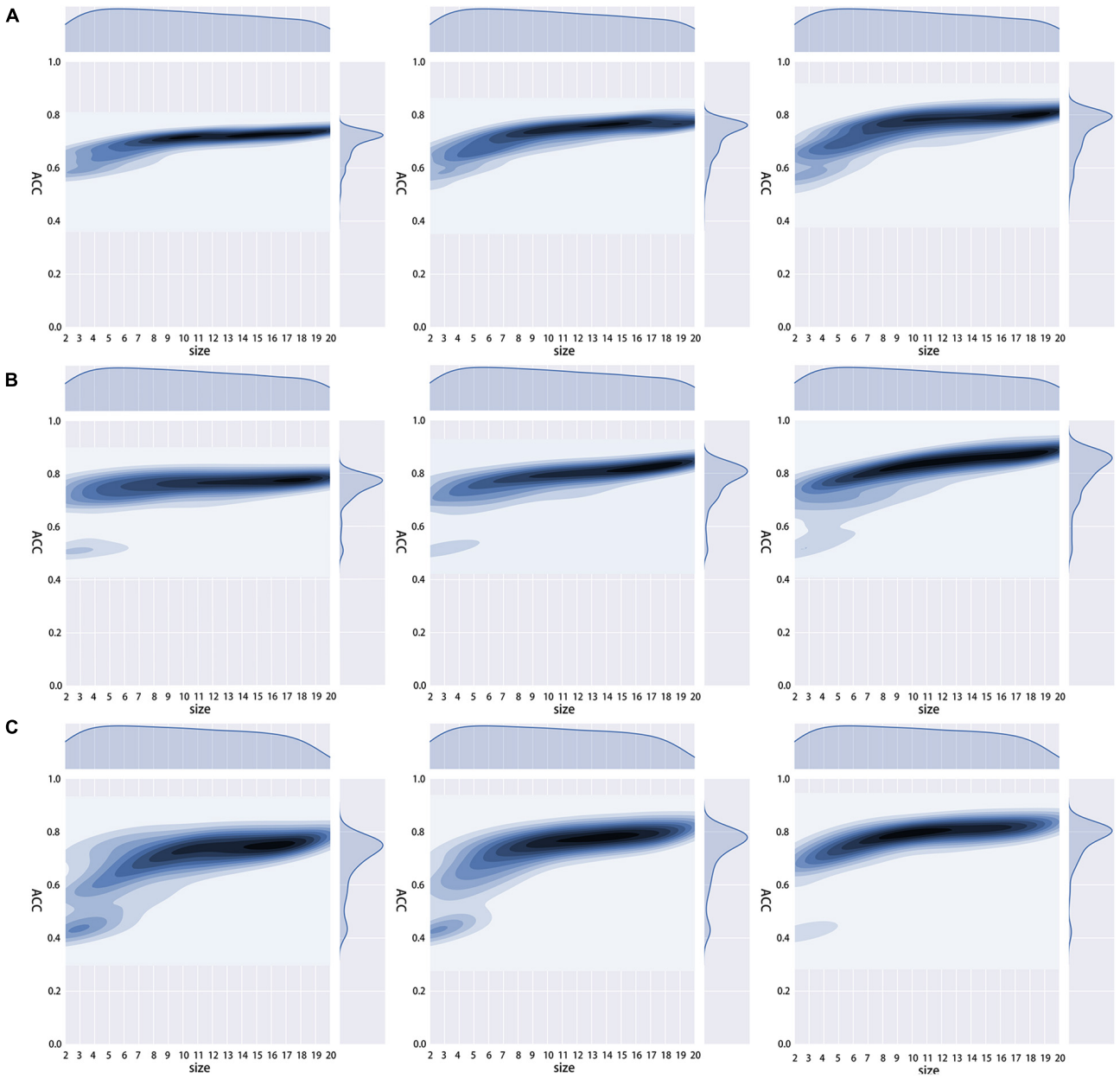


FIGURE 3 | Density distribution diagram of different *K* value accuracy rates. **(A–C)** are the density distribution diagrams of the 2OG-SwissProt set, 2OG-Fe set, and 2OG-domain set at different *K* values in 673 reduction schemes.

RESULTS

Predictive Performance of Different Reducing Amino Acid Cluster

To obtain the optimal amino acid reduction scheme and the appropriate K value (K = 1, 2, 3), we calculated the accuracy of the 673 reduction schemes mentioned in RAACBook (Zheng et al., 2019) with the different K values. We found that all three models showed the best performance at K = 3, and most of the reduction schemes had higher accuracy when K = 3 (**Figure 3**). We guessed that there would be more features when K = 3, and they would better reflect the properties of the protein and get a more accurate model.

After confirming that the model has better performance when K = 3, we then selected the best scheme from 673 RAAC schemes to construct the model. In the 2OG-SwissProt model, we tested each size of each reduction type and compared different reduction sizes of different reduction types (**Figure 4A**). We found that when t = 33 (**Table 2**), s = 15 (t represents the t-th reduction type in RAACBook; s represents the size of the RAAC), the highest accuracy rate is 83.75% (**Figure 4B**). In the prediction of the 2OG-Fe dataset, we were pleasantly surprised to find that the highest accuracy rate also appears in the reduction type 33, and the highest accuracy rate is 90.04% when s = 16 (**Supplementary Figure 1B**). There is also a very high accuracy rate at s = 15, reaching 88.76% (**Supplementary Figure 1A**). The reduction method of type 33 uses a database of aligned protein structures to propose a new clustering method based on the substitution scores, which aggregates 20 amino acids in two groups, namely, the hydrophobic groups and the polar groups (Li and Wang, 2007). Therefore, we speculated that the function of 2OG oxygenases may be related to its polarity and hydrophobicity.

To further prove that polarity and hydrophobicity may be related to the function of 2OG oxygenases, we manually extracted the 2OG oxygenase domain and JmjC domain sequences and other iron-binding domain sequences for prediction. Protein functions mainly through its domain region, and 2OG oxygenases also bind Fe(II) and 2-oxoglutarate in their domain position to perform their functions. Therefore, the region outside the domain may be noise information for feature extraction, and only using the domain sequence to extract features can better reflect the function of 2OG oxygenases (Shen and Zou, 2020). The result is the same as we expected, when t = 33 and s = 15, the highest accuracy rate is obtained (**Supplementary Figure 1B**). The same result is obtained with the complete sequence, which further proves that the polarity and hydrophobicity may be related to the function of 2OG oxygenases.

The functional domain of 2OG oxygenases contains Fe^{2+}-binding sites and α-ketoglutarate-binding sites, and their amino acid composition is almost completely conserved. The Fe^{2+}-binding motif (HXD-H) and α-KG-binding motif (N-Y-R-R) of the ALKBH family are entirely conserved in the homologs (Bjornstad et al., 2011; Fedeles et al., 2015; Alemu et al., 2016; Xu et al., 2021), and other 2OG oxygenases have similar structures (Bleijlevens et al., 2008; Islam et al., 2018;

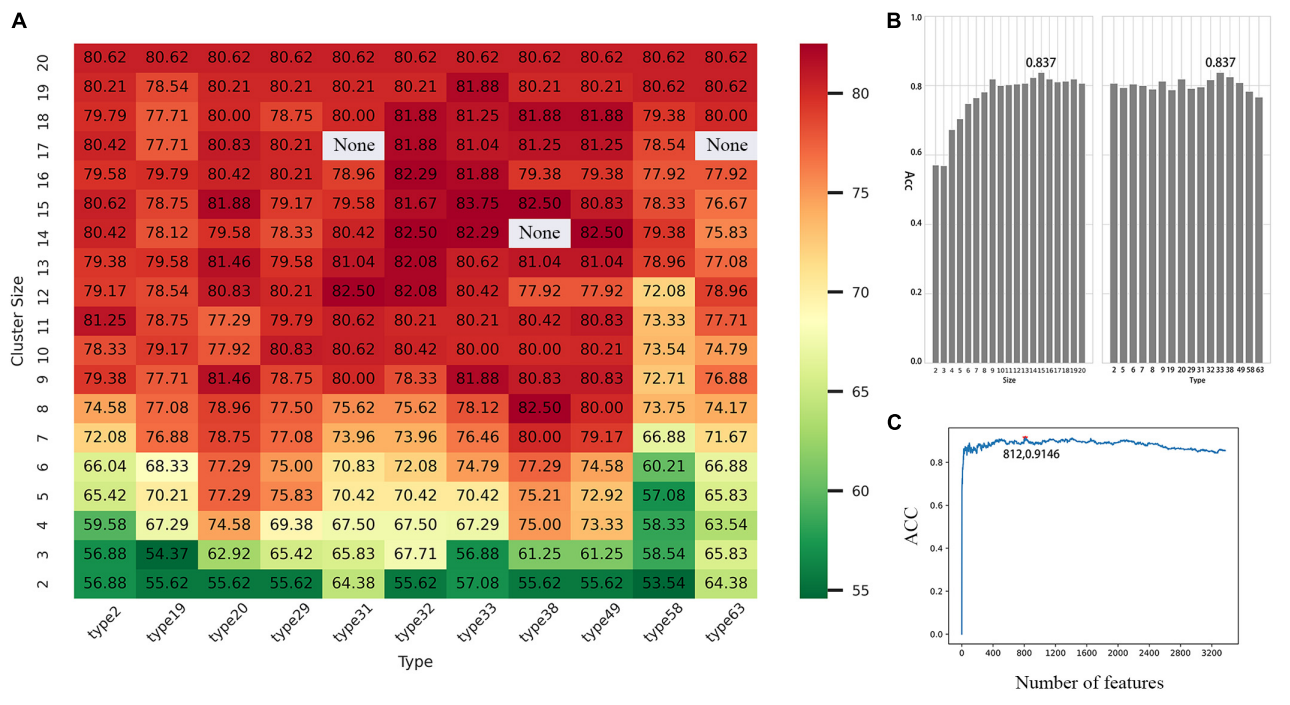


FIGURE 4 | Performance evaluation of different reduced amino acid clusters. **(A)** Heat map of accuracy distribution of different reduced amino acid clusters. **(B)** The accuracy rate of the reduced amino acid cluster (t = 33, s = 15) with the highest accuracy rate reaches 83.75%. **(C)** The incremental feature selection (IFS) curve shows that prediction accuracy is 91.46% when using 812 optimal features based on the tripeptide combination (t = 33, s = 15).

TABLE 2 | Cluster size of reduced amino acid alphabet of type 33.

Size	Reduced amino acid cluster
2	STANDGRQEKHPIVLMWYF-C
3	STANDGRQEKHP-IVLMWYF-C
4	STANDG-RQEKHP-IVLMWYF-C
5	STAND-G-RQEKHP-IVLMWYF-C
6	STAND-G-RQEK-HP-IVLMWYF-C
7	STA-ND-G-RQEK-HP-IVLMWYF-C
8	STA-ND-G-RQ-EK-HP-IVLMWYF-C
9	STA-ND-G-RQ-EK-HP-IVLM-WYF-C
10	ST-A-ND-G-RQ-EK-HP-IVLM-WYF-C
11	ST-A-ND-G-RQ-EK-H-P-IVLM-WYF-C
12	ST-A-N-D-G-RQ-EK-H-P-IVLM-WYF-C
13	ST-A-N-D-G-RQ-EK-H-P-IV-LM-WYF-C
14	S-T-A-N-D-G-RQ-EK-H-P-IV-LM-WYF-C
15	S-T-A-N-D-G-RQ-EK-H-P-IV-L-M-WYF-C
16	S-T-A-N-D-G-RQ-E-K-H-P-IV-L-M-WYF-C
17	S-T-A-N-D-G-RQ-E-K-H-P-IV-L-M-WY-F-C
18	S-T-A-N-D-G-R-Q-E-K-H-P-IV-L-M-WY-F-C
19	S-T-A-N-D-G-R-Q-E-K-H-P-I-V-L-M-WY-F-C

Wang et al., 2021). They all combine Fe^{2+} and α-ketoglutarate through conserved polar amino acid regions, which may be the reason why polarity is an essential feature of 2OG oxygenase identification. In addition, in the best reduction scheme, Phenylalanine (F), Tryptophan (W), and Tyrosine (Y) are recombined into a new letter, and these three amino acids are all aromatic amino acids. We speculate that the function of 2OG oxygenases may be related to the hydrophobicity of aromatic amino acids and the unique properties of its benzene ring.

TABLE 3 | The results of each evaluation index of the three models.

Model	Acc (%)	Sn (%)	SP (%)	MCC (%)	*F*1 score (%)	AUC (%)
2OG-SwissProt	91.04	93.33	88.75	82.34	91.26	97.15
2OG-Fe	97.23	97.92	96.53	94.48	97.31	99.57
2OG-domain	97.87	98.23	97.65	95.60	97.37	99.89

Acc, accuracy; AUC, area under the curve; MCC, Matthew's correlation coefficient; Sn, sensitivity; and Sp, specificity.

Feature Selection

Although we can get more features when $K = 3$, not every feature can be helpful to the prediction of 2OG oxygenases; some features may even become noise information and affect the final result. Therefore, we used ANOVA combined with IFS to select the best feature combination. Through 10-fold cross-validation, the 2OG-SwissProt model achieves an optimal performance of 91.46% with 812 feature combinations (**Figure 4C**); the 2OG-Fe model achieves an optimal performance of 96.61% with 1,181 feature combinations (**Supplementary Figure 1C**); 2OG-domain

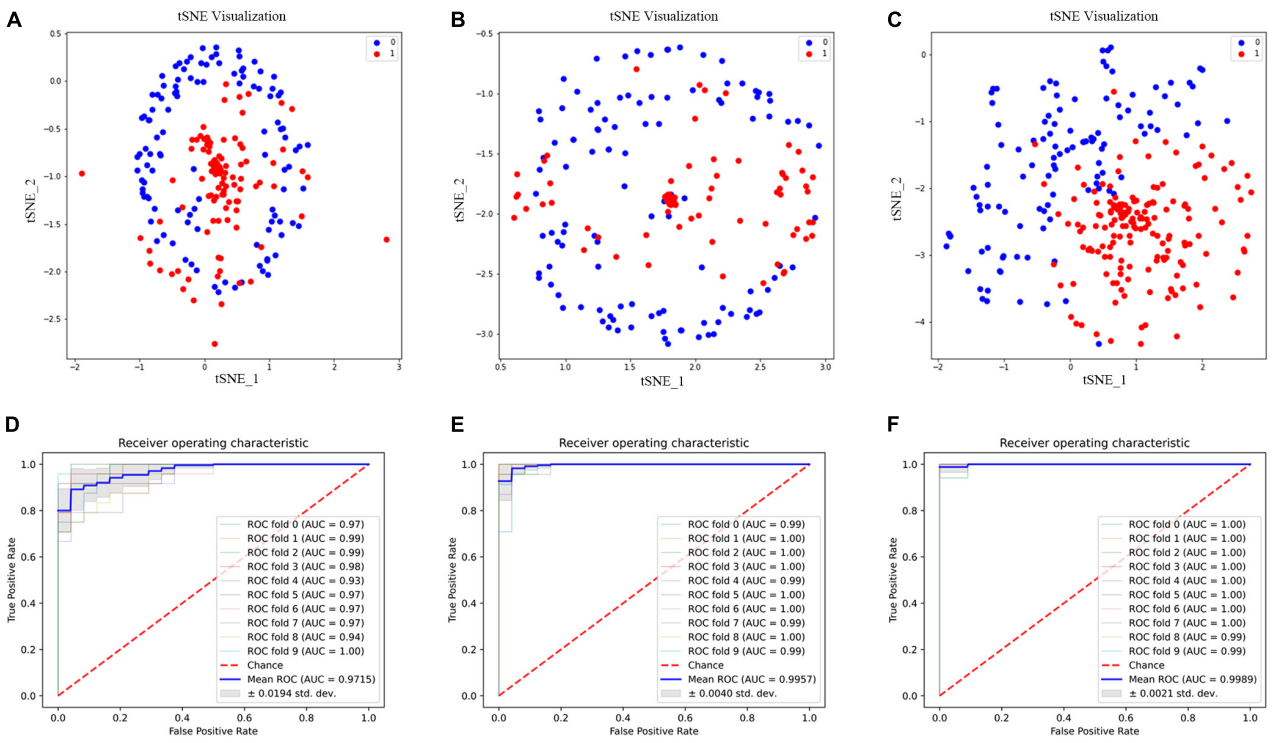
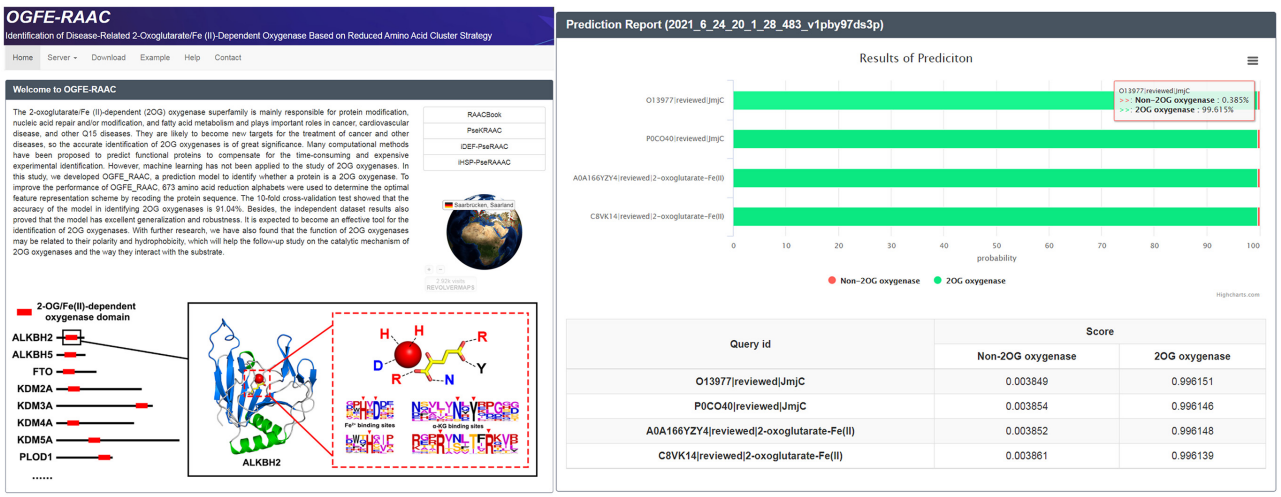


FIGURE 5 | Feature set t-SNE clustering scatter diagram and receiver operating characteristic (ROC) curve diagram. **(A–C)** are the t-SNE clustering analysis diagrams of the feature set after unreduced, reduced, and feature screening, respectively. 0 and 1 represent positive samples and negative samples, respectively. **(D–F)** are the ROC curves of the three models 2OG-SwissProt, 2OG-Fe, and 2OG-domain, respectively.



Query id	Score	
	Non-2OG oxygenase	2OG oxygenase
O13977|reviewed|JmjC	0.003849	0.996151
P0CO40|reviewed|JmjC	0.003854	0.996146
A0A166YZY4|reviewed|2-oxoglutarate-Fe(II)	0.003852	0.996148
C8VK14|reviewed|2-oxoglutarate-Fe(II)	0.003861	0.996139

FIGURE 6 | Home page and results page of OGFE-RAAC web server.

model also achieves an optimal performance of 96.07% with 350 feature combinations (**Supplementary Figure 1C**). For more clearly showing that the filtered features can better reflect the nature of 2OG oxygenases, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the feature sets after unreduced, reduced, and feature screening in a 2D feature space (**Figures 5A–C**). Obviously, the results show that the feature set clustering effect after feature screening is better, and it can effectively separate 2OG oxygenases from non-2OG oxygenases.

Performance Evaluation

We evaluated our model by 10-fold cross-validation to verify that our model is effective (**Table 3**). At the same time, we drew the receiver operating characteristic (ROC) curve through the 10-fold cross-validation (**Figures 5D–F**).

In order to further evaluate our predictor, we used an independent test set to test 2OG-SwissProt, 2OG-Fe, and 2OG-domain models. The 2OG-SwissProt model accurately predicts 143 samples out of 150 test set samples, and the accuracy rate is 95.33%. The 2OG-Fe model accurately predicts 149 samples out of 159 test set samples, with an accuracy rate of 93.71%. The 2OG-domain model accurately predicts 963 samples out of 1,036 test set samples, with an accuracy rate of 92.95%. These show that our predictor is effective and robust.

Web Server Guidance

For the purpose of other researchers to use our model more conveniently, an easy-to-use web server was established to implement our predictor, which can be freely accessed at http://bioinfor.imu.edu.cn/ogferaac. When you want to use our tool, you need to click the "Service" module and then import the FASTA protein sequence into the input box or upload the button to upload your protein data. Meanwhile, according to the different sequences you provide, you can also choose different modules (2OG-SwissProt, 2OG-Fe, and 2OG-domain) for prediction. After submitting the task, the website will provide the corresponding forecast report, which will display the forecast results and probability of each sequence in the form of tables and flowcharts (**Figure 6**).

DISCUSSION

At present, the research on 2OG oxygenases is more in-depth, and its many functions (such as demethylation) occupy an important position in the research of diseases (Liu et al., 2019; Ao et al., 2021). Based on RAAC strategy and SVM, the prediction model of 2OG oxygenases is constructed. t-SNE results show that RAAC can effectively reduce protein complexity, extract conservative features hidden in noise information, and improve prediction accuracy. OGFE_RAAC has strong robustness and generalization to accurately predict 2OG oxygenases. We anticipate that OGFE_RAAC can accurately and rapidly identify 2OG oxygenases based on peptide sequence and promote the development of related drug research. Not only that, we also found that the function of 2OG oxygenases may be related to its hydrophobicity and polarity during the prediction process, which also provides a new research idea for the future study of 2OG oxygenases.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://bioinfor.imu.edu.cn/ogferaac/public/Download.

AUTHOR CONTRIBUTIONS

YZ conceived and designed the study. JZ and PL organized and collected the data and carried out the computation.

LZ designed and developed the web server. JZ and HW wrote the manuscript. SB participated in all subsequent revisions of the manuscript. YZ planned overall and revised the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Scientific Foundation of China (Nos. 62061034 and 61861036), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT-18-B01), the Fund for Excellent Young Scholars of Inner Mongolia (2017JQ04), and the Science and Technology Major Project of Inner Mongolia Autonomous Region of China to the State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock (2019ZD031).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.707938/full#supplementary-material

REFERENCES

Alemu, E. A., He, C., and Klungland, A. (2016). ALKBHs-facilitated RNA modifications and de-modifications. *DNA Repair* 44, 87–91. doi: 10.1016/j.dnarep.2016.05.026

Ao, C., Yu, L., and Zou, Q. (2021). Prediction of bio-sequence modifications and the associations with diseases. *Brief. Funct. Genomics* 20, 1–18. doi: 10.1093/bfgp/elaa023

Arifuzzaman, S., Khatun, M. R., and Khatun, R. (2020). Emerging of lysine demethylases (KDMs): from pathophysiological insights to novel therapeutic opportunities. *Biomed. Pharmacother.* 129:110392. doi: 10.1016/j.biopha.2020.110392

Beer, M. A. (2017). Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.* 38, 1251–1258. doi: 10.1002/humu.23185

Bjornstad, L. G., Zoppellaro, G., Tomter, A. B., Falnes, P. O., and Andersson, K. K. (2011). Spectroscopic and magnetic studies of wild-type and mutant forms of the Fe(II)- and 2-oxoglutarate-dependent decarboxylase ALKBH4. *Biochem. J.* 434, 391–398. doi: 10.1042/bj20101667

Bleijlevens, B., Shivarattan, T., Flashman, E., Yang, Y., Simpson, P. J., Koivisto, P., et al. (2008). Dynamic states of the DNA repair enzyme AlkB regulate product release. *EMBO Rep.* 9, 872–877. doi: 10.1038/embor.2008.120

Brylinski, M., and Skolnick, J. (2011). FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. *Proteins* 79, 735–751. doi: 10.1002/prot.22913

Chang, B. S., Chen, Y., Zhao, Y. M., and Bruick, R. K. (2007). JMJD6 is a histone arginine demethylase. *Science* 318, 444–447. doi: 10.1126/science.1145801

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Chen, Z., Wang, Y., Zhai, Y. F., Song, J., and Zhang, Z. (2013). ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences. *Mol. Biosyst.* 9, 2213–2222. doi: 10.1039/c3mb70100j

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273, 236–247. doi: 10.1016/j.jtbi.2010.12.024

Chou, K. C., and Shen, H. B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. doi: 10.1038/nprot.2007.494

Chou, K. C., Wu, Z. C., and Xiao, X. (2012). iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* 8, 629–641. doi: 10.1039/c1mb05420a

Farrow, S. C., and Facchini, P. J. (2014). Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. *Front. Plant Sci.* 5:524. doi: 10.3389/fpls.2014.00524

Fedeles, B. I., Singh, V., Delaney, J. C., Li, D. Y., and Essigmann, J. M. (2015). The AlkB Family of Fe(II)/alpha-ketoglutarate-dependent dioxygenases: repairing nucleic acid alkylation damage and beyond. *J. Biol. Chem.* 290, 20734–20742. doi: 10.1074/jbc.r115.656462

Fu, L. M., Niu, B. F., Zhu, Z. W., Wu, S. T., and Li, W. Z. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Georgiou, D. N., Karakasidis, T. E., Nieto, J. J., and Torres, A. (2009). Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.* 257, 17–26. doi: 10.1016/j.jtbi.2008.11.003

Hausinger, R. P. (2004). FeII/alpha-ketoglutarate-dependent hydroxylases and related enzymes. *Crit. Rev. Biochem. Mol. Biol.* 39, 21–68. doi: 10.1080/10409230490440541

He, S., Guo, F., Zou, Q., and Ding, H. (2020). MRMD2.0: a python tool for machine learning with feature ranking and reduction. *Curr. Bioinform.* 15, 1213–1221. doi: 10.2174/1574893615999200503030350

Hewitson, K. S., Granatino, N., Welford, R. W., Mcdonough, M. A., and Schofield, C. J. (2005). Oxidation by 2-oxoglutarate oxygenases: non-haem iron systems in catalysis and signalling. *Philos. Trans. A Math. Phys. Eng. Sci.* 363, 807–828. discussion 1035-1040., doi: 10.1098/rsta.2004.1540

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., and Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics* 15, 41–51.

Huang, Y., Niu, B. F., Gao, Y., Fu, L. M., and Li, W. Z. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Islam, M. S., Leissing, T. M., Chowdhury, R., Hopkinson, R. J., and Schofield, C. J. (2018). 2-oxoglutarate-dependent oxygenases. *Annu. Rev. Biochem.* 87, 585–620.

Jaillard, M., Palmieri, M., Van Belkum, A., and Mahe, P. (2020). Interpreting k-mer-based signatures for antibiotic resistance prediction. *Gigascience* 9:giaa110. doi: 10.1093/gigascience/giaa110

Jia, B., Tang, K., Chun, B. H., and Jeon, C. O. (2017). Large-scale examination of functional and sequence diversity of 2-oxoglutarate/Fe(II)-dependent oxygenases in Metazoa. *Biochim. Biophys. Acta Gen. Sub.* 1861, 2922–2933. doi: 10.1016/j.bbagen.2017.08.019

Kawai, Y., Ono, E., and Mizutani, M. (2014). Evolution and diversity of the 2-oxoglutarate-dependent dioxygenase superfamily in plants. *Plant J.* 78, 328–343. doi: 10.1111/tpj.12479

Lee, Y. F., Miller, L. D., Chan, X. B., Black, M. A., Pang, B., Ong, C. W., et al. (2012). JMJD6 is a driver of cellular proliferation and motility and a marker of poor prognosis in breast cancer. *Breast Cancer Res.* 14:R85.

Levy, R., Edelman, M., and Sobolev, V. (2009). Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins* 76, 365–374. doi: 10.1002/prot.22352

Li, F. Y., Leier, A., Liu, Q. Z., Wang, Y. A., Xiang, D. X., Akutsu, T., et al. (2020). Procleave: predicting protease-specific substrate cleavage sites by combining sequence and structural information. *Genomics Proteomics Bioinformatics* 18, 52–64. doi: 10.1016/j.gpb.2019.08.002

Li, J., and Wang, W. (2007). Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids. *Sci. China Series C Life Sci.* 50, 392–402. doi: 10.1007/s11427-007-0023-3

Liao, C. H., and Zhang, Q. (2020). ASIP COTRAN EARLY CAREER INVESTIGATOR AWARD LECTURE Understanding the oxygen-sensing pathway and its therapeutic implications in diseases. *Am. J. Pathol.* 190, 1584–1595. doi: 10.1016/j.ajpath.2020.04.003

Lin, H. H., Han, L. Y., Zhang, H. L., Zheng, C. J., Xie, B., Cao, Z. W., et al. (2006). Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. *BMC Bioinform.* 7(Suppl. 5):S13. doi: 10.1186/1471-2105-7-S5-S13

Liou, Y. F., Charoenkwan, P., Srinivasulu, Y., Vasylenko, T., Lai, S. C., Lee, H. C., et al. (2014). SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides. *BMC Bioinform.* 15(Suppl. 16):S4. doi: 10.1186/1471-2105-15-S16-S4

Liu, D. Y., Li, G. P., and Zuo, Y. C. (2019). Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief. Bioinform.* 20, 1826–1835. doi: 10.1093/bib/bby053

Liu, R., and Hu, J. (2011). HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinform.* 12:207. doi: 10.1186/1471-2105-12-207

Liu, W., Ma, Q., Wong, K., Li, W. B., Ohgi, K., Zhang, J., et al. (2013). Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* 155, 1581–1595. doi: 10.1016/j.cell.2013.10.056

Manavalan, B., Shin, T. H., and Lee, G. (2018). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Meng, C., Guo, F., and Zou, Q. (2020). CWLy-SVM: a support vector machine-based tool for identifying cell wall lytic enzymes. *Comput. Biol. Chem.* 87:107304. doi: 10.1016/j.compbiolchem.2020.107304

Mohan, A., Anishetty, S., and Gautam, P. (2010). Global metal-ion binding protein fingerprint: a method to identify motif-less metal-ion binding proteins. *J. Bioinform. Comput. Biol.* 8, 717–726. doi: 10.1142/s0219720010004884

Myllyharju, J. (2008). Prolyl 4-hydroxylases, key enzymes in the synthesis of collagens and regulation of the response to hypoxia, and their roles as treatment targets. *Ann. Med.* 40, 402–417. doi: 10.1080/07853890801986594

Rose, N. R., Mcdonough, M. A., King, O. N. F., Kawamura, A., and Schofield, C. J. (2011). Inhibition of 2-oxoglutarate dependent oxygenases. *Chem. Soc. Rev.* 40, 4364–4397.

Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. doi: 10.1093/bioinformatics/btm344

Shen, Z. J., and Zou, Q. (2020). Basic polar and hydrophobic properties are the main characteristics that affect the binding of transcription factors to methylation sites. *Bioinformatics* 36, 4263–4268. doi: 10.1093/bioinformatics/btaa492

Shu, N., Zhou, T., and Hovmöller, S. (2008). Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* 24, 775–782. doi: 10.1093/bioinformatics/btm618

Sun, Z. J., Huang, S. H., Zheng, L., Liang, P. F., Yang, W. R. T., and Zuo, Y. C. (2020). ICTC-RAAC: an improved web predictor for identifying the types of ion channel-targeted conotoxins by using reduced amino acid cluster descriptors. *Comput. Biol. Chem.* 89:107371. doi: 10.1016/j.compbiolchem.2020.107371

Tahir, M., and Idris, A. (2020). MD-LBP: an efficient computational model for protein subcellular localization from HeLa cell lines using SVM. *Curr. Bioinform.* 15, 204–211. doi: 10.2174/1574893614666190723120716

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Valasatava, Y., Rosato, A., Banci, L., and Andreini, C. (2016). MetalPredator: a web server to predict iron-sulfur cluster binding proteomes. *Bioinformatics* 32, 2850–2852. doi: 10.1093/bioinformatics/btw238

Vasta, J. D., and Raines, R. T. (2018). Collagen Prolyl 4-Hydroxylase as a therapeutic target. *J. Med. Chem.* 61, 10403–10411. doi: 10.1021/acs.jmedchem.8b00822

Wang, H., Xi, Q. L. M. G., Liang, P. F., Zheng, L., Hong, Y., and Zuo, Y. C. (2021). IHEC_RAAC: a online platform for identifying human enzyme classes via reduced amino acid cluster strategy. *Amino Acids* 53, 239–251. doi: 10.1007/s00726-021-02941-9

Wang, Z., Liu, D., Xu, B., Tian, R., and Zuo, Y. (2021). Modular arrangements of sequence motifs determine the functional diversity of KDM proteins. *Brief. Bioinform.* 22:bbaa215. doi: 10.1093/bib/bbaa215

Xiao, M. Z., Liu, J. M., Xian, C. L., Chen, K. Y., Liu, Z. Q., and Cheng, Y. Y. (2020). Therapeutic potential of ALKB homologs for cardiovascular disease. *Biomed. Pharmacother.* 131:110645. doi: 10.1016/j.biopha.2020.110645

Xu, B., Liu, D., Wang, Z., Tian, R., and Zuo, Y. (2021). Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family. *Cell. Mol. Life Sci.* 78, 129–141. doi: 10.1007/s00018-020-03594-9

Yan, R., Wang, X., Tian, Y., Xu, J., Xu, X., and Lin, J. (2019). Prediction of zinc-binding sites using multiple sequence profiles and machine learning methods. *Mol. Omics* 15, 205–215. doi: 10.1039/c9mo00043g

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 75, 140–149. doi: 10.1016/j.inffus.2021.02.015

Zhang, D., Chen, H.-D., Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Zhang, Z.-Y., et al. (2021). iBLP: an XGBoost-based predictor for identifying bioluminescent proteins. *Comput. Math. Methods Med.* 2021:6664362.

Zhang, H. Y., Xi, Q., Huang, S. H., Zheng, L., Yang, W., and Zuo, Y. C. (2020). iSP-RAAC: identify secretory proteins of malaria parasite using reduced amino acid composition. *Comb. Chem. High Throughput Screen.* 23, 536–545. doi: 10.2174/1386207323666200402084518

Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., et al. (2019). RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. *Database (Oxford)* 2019:baz131.

Zheng, L., Liu, D., Yang, W., Yang, L., and Zuo, Y. (2020). RaacLogo: a new sequence logo generator by using reduced amino acid clusters. *Brief. Bioinform.* 22:bbaa096. doi: 10.1093/bib/bbaa096

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* 21, 1–10.

Zou, Q., Zeng, J. C., Cao, L. J., and Ji, R. R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Zuo, Y., Lv, Y., Wei, Z., Yang, L., Li, G., and Fan, G. (2015). iDPF-PseRAAAC: a web-Server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition. *PLoS One* 10:e0145541. doi: 10.1371/journal.pone.0145541

Zuo, Y. C., Chang, Y., Huang, S. H., Zheng, L., Yang, L., and Cao, G. F. (2019). iDEF-PseRAAC: identifying the defensin peptide by using reduced amino acid composition descriptor. *Evol. Bioinform.* 15:1176934319867088.

Zuo, Y. C., and Li, Q. Z. (2009). Using reduced amino acid composition to predict defensin family and subfamily: integrating similarity measure and structural alphabet. *Peptides* 30, 1788–1793. doi: 10.1016/j.peptides.2009.06.032

Zuo, Y. C., Li, Y., Chen, Y. L., Li, G. P., Yan, Z. H., and Yang, L. (2017). PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* 33, 122–124. doi: 10.1093/bioinformatics/btw564

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Zhou, Bo, Wang, Zheng, Liang and Zuo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 20 July 2021
doi: 10.3389/fcell.2021.712020

Identification of Novel Susceptible Genes of Gastric Cancer Based on Integrated Omics Data

Huang Yaoxing, Yu Danchun, Sun Xiaojuan, Jiang Shuman, Yan Qingqing and Jia Lin**

Department of Gastroenterology, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

Gastric cancer (GC) is one of the most common causes of cancer-related deaths in the world. This cancer has been regarded as a biological and genetically heterogeneous disease with a poorly understood carcinogenesis at the molecular level. Thousands of biomarkers and susceptible loci have been explored via experimental and computational methods, but their effects on disease outcome are still unknown. Genome-wide association studies (GWAS) have identified multiple susceptible loci for GC, but due to the linkage disequilibrium (LD), single-nucleotide polymorphisms (SNPs) may fall within the non-coding region and exert their biological function by modulating the gene expression level. In this study, we collected 1,091 cases and 410,350 controls from the GWAS catalog database. Integrating with gene expression level data obtained from stomach tissue, we conducted a machine learning-based method to predict GC-susceptible genes. As a result, we identified 787 novel susceptible genes related to GC, which will provide new insight into the genetic and biological basis for the mechanism and pathology of GC development.

Keywords: gastric cancer, GWAS, integrated omics data, a machine learning based method, biomarkers

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Shihua Zhang,
Wuhan University of Science and Technology, China
Haixiu Yang,
Harbin Medical University, China

****Correspondence:***
Huang Yaoxing
huangyaoxing@sina.com
Jia Lin
sqt661841@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *19 May 2021*
Accepted: *23 June 2021*
Published: *20 July 2021*

Citation:
Yaoxing H, Danchun Y, Xiaojuan S, Shuman J, Qingqing Y and Lin J (2021) Identification of Novel Susceptible Genes of Gastric Cancer Based on Integrated Omics Data. Front. Cell Dev. Biol. 9:712020. doi: 10.3389/fcell.2021.712020

INTRODUCTION

Gastric cancer (GC) is one of the most common malignant dangerous neoplasms and fatal diseases in the world. It has been reported that there were approximately one million newly diagnosed gastric carcinoma cases, which caused 780,000 deaths in 2018 (Bray et al., 2018). It is notable that nearly half of the GC incidences occurred in the Asian region, which partly resulted from the diverse hereditary background, behavioral factors, and the spread of and infection by *Helicobacter pylori* (Chen et al., 2016; Zheng et al., 2017). A lot of work has been done to improve the diagnosis and therapy of GC. However, the survival rate of GC patients remains poor at approximately 30% in the recent 5 years (DeSantis et al., 2014). Therefore, many efforts have been made to discover new biomarkers to help in staging and in prognosis of the tumor diagnosis, which could help in improving early diagnosis and prognostic prediction of GC (Ludwig and Weinstein, 2005).

Disease gene prediction is a task of identifying the significant susceptible genes related to diseases. There have been a variety of approaches proposed, such as annotation-based approaches, network-based approaches, and machine learning-based approaches. Annotation-based approaches, such as prioritization of candidate genes using statistics (POCUS) (Turner et al., 2003), SUSPECTS (Adie et al., 2006), Endeavor (Aerts et al., 2006), and Transcriptome Ontology Pathway PubMed based prioritization of Genes (ToppGene) (Chen et al., 2007), are proposed based on annotating the genes with respect to biological structures or functions then comparing the

annotations with known disease causal genes. However, these methods have a limitation of failing to capture the indirect relationships between the genes that may have common features or functions but are still not annotated. Network-based methods are proposed to overcome this by utilizing the large scale of interactome data between cellular molecules covering most of the genome and proteome (Wang et al., 2011).

Machine learning techniques have been applied to solve various biomedical problems (Lu and Zhao, 2019; Zhao et al., 2020a,b), such as pattern recognition (Barral et al., 2012; Cui et al., 2020), classification (Basford et al., 2013; Zhao et al., 2021a), prediction of drug target (Ding et al., 2014; Tianyi et al., 2020), and genome annotation (Yip et al., 2013). Thus, there is no doubt that machine learning methods have been applied in disease-associated genes prediction (Calvo et al., 2006; Xu and Li, 2006). In recent years, emerging evidences have illustrated the essential role of single-nucleotide polymorphisms (SNPs) in GC development and progression. Since genome-wide association studies (GWAS) are a widely known powerful approach to explore complex susceptible variants of diseases, many studies have reported a number of susceptibility loci associated with GC through GWAS analysis; however, they can only explain a small fraction of GC heritability (Wang et al., 2017; Park et al., 2019). Moreover, most disease-related SNPs identified by GWAS fall into intergenic or non-coding regions, which may influence the process of pathogenesis by modulating the expression level of target genes (Maurano et al., 2012). However, genetic variants are still powerful and high-quality biomarkers for screening GC susceptibility (Mocellin et al., 2015).

In addition, it has been proven that gene expression is significantly related to diseases (Zhao et al., 2021b). Expression quantitative trait locus (eQTL) analysis has been regarded as a powerful approach to provide prior weights for the statistical analysis of new causal SNP identification and prioritize SNPs or genes for further validation (Li et al., 2013b). Due to the theory of linkage disequilibrium (LD), which is reflected by the non-random association of alleles of different loci, it can be inferred that SNPs can regulate the pathologies of diseases by modulating the expression level of target genes. However, most studies select the representative SNPs by their closest located gene, which may inevitably obscure the genetic effect between that candidate gene and the trait. Thus, integrating GWAS data and eQTL data can help us to detect the genetic mechanism of complex disease. The Genotype-Tissue Expression (GTEx) project has provided the largest comprehensive public database of whole-genome and transcriptome sequencing data to help better understand the effects and molecular mechanism of function variations.

Considering the fact that regulatory causal SNPs may exert their function by affecting their target gene expression, we collected 1,091 GC cases and 410,350 controls from GWAS catalog database and eQTL summary data from stomach tissues in the GTEx database (Rashkin et al., 2020). We then extracted gene features from both GWAS summary data and eQTL data and integrated them as a 10D vector to represent gene feature, then we performed several machine learning methods to assess the classification performance of the models and selected random forest (RF) classifier based on its excellent performance. We identified 787 novel susceptible genes related to GC which may help provide new insights into the mechanism of GC.

MATERIALS AND METHODS

GC GWAS Datasets

In this study, we obtained a GC GWAS dataset consisting of 1,091 cases and 410,350 controls from the UK Biobank (Rashkin et al., 2020). All the subjects were genotyped with Affymetrix Genome-Wide Human SNP Array. Eleven significant susceptible SNPs were identified related to GC and esophageal carcinoma with p-value $< 1 \times 10^{-6}$. However, they identified 9,986,610 susceptible loci associated with stomach and esophageal cancer in total, which is utilized in our study to be further identified.

Tissue eQTL Dataset

Expression level-associated SNPs in stomach tissues were obtained from the GTEx v8 database. Genotyping was performed utilizing Illumina HumanOmni 5 and 2.5 M. And transcriptome dataset was generated by using Affymetrix Expression Array or Illumina TruSeq RNA sequencing. As a result, 24,291 susceptible loci were identified based on gene expression level.

GC-Related Genes and Candidate Genes

After obtaining both omics summary data, we obtained 5,632 gastric-related genes from DisGeNET database, which are considered as positive genes for machine learning methods (Bauer-Mehren et al., 2010). Then we downloaded the human gene networks from HumanNet v2.0, which was a database of human gene networks illustrating gene–gene interactions (Hwang et al., 2018). After filtering the genes related to the GC causal genes obtained from DisGeNET, we obtained 3,227 genes whose correlation scores were <1, which indicate that they are negative genes, and the rest of the genes are regarded candidate genes which may have association with GC.

Feature Extraction

We first obtained GWAS dataset and eQTL dataset associated with GC from the GWAS catalog database and the GTEx database, which represent the gene feature from the aspect of phenotype and transcription, respectively. Since the LD correction structure means that the majority of the identified variants associated with the traits frequently point to the regions where many genes are located, it is extremely difficult to prioritize among these susceptible genes to identify the most functionally relevant causal genes merely based on GWAS data. It is widely known that SNPs can exert their regulation function by modulating the expression level of target genes, which may further have a significant influence on the phenotype. Many studies have applied analytical approaches to integrate eQTL and GWAS data to detect the causal genes associations with complex traits (Giambartolomei et al., 2014; Gusev et al., 2016). However, to our knowledge, there is no method utilizing machine learning classifiers to prioritize susceptible genes of GC based on integrated GWAS and eQTL summary data.

After obtaining both omics summary data, we obtained gastric-related genes from DisGeNET database (Piñero et al., 2016). We first performed a data preprocessing process to transform the names of genes downloaded from different databases. We used an R package to obtain the detailed information of the genes, such as chromosome, start position, and end position information of these genes. After mapping these genes to the GWAS and eQTL data based on the location information, we finally got 5,633 genes regarded as a training set. According to the mapping information, we kept the genes with at least one susceptible loci identified by GWAS analysis. After prioritizing the SNPs by their p-value obtained by GWAS analysis, we used the p-value of the top five SNPs related to each gene as a 5D phenotype-based feature vector. For those genes with less than five associated SNPs, the feature vector is filtered with 1, which means the gene has no correlation with GC. Thus, the phenotype-based feature vector of G_i can be denoted as follows:

$$G_i^p = \left[P_p^1,\ P_p^2,\ P_p^3,\ P_p^4,\ P_p^5 \right] \quad (1)$$

After obtaining the top five SNPs associated with these genes, these SNPs can be mapped to the stomach tissue eQTL data. Based on the same method, we use the p-value of those SNPs successfully mapped in eQTL data to represent a 5D transcriptome-based feature vector of each gene. For those SNPs not mapped, we all use 1 to fill up the feature vector. Thus, the transcriptome-based feature vector can be denoted as follows:

$$G_i^T = \left[P_T^1,\ P_T^2,\ P_T^3,\ P_T^4,\ P_T^5 \right] \quad (2)$$

Thus, each gene feature can be represented as a 10D feature vector based on the integrated omics data.

Gene Prediction Using Random Forest Classifier

To date, the binary classification methods have been widely applied in disease causal gene prediction problems, such as naïve Bayesian classifier (NB), support vector machine (SVM), RF, and some deep learning methods such as convolutional neural network (CNN), graph neural network (GCN), and deep neural network (DNN). In this work, we used a RF classifier to predict GC causal genes. In order to evaluate the RF classifier, we performed a 10-fold cross-validation on the training set. The main workflow is shown in **Figure 1**. First, the 5,633 genes are randomly divided into 10 groups; 9 of them were chosen to be training samples, and the last one is the testing set for a total of 10 training iterations. Grid searches were performed to obtain the best performance of the parameters of the RF classifier. The final statistical results are averaged after 10 iterations. The receiver operating characteristic (ROC) curve and the area under the curve (AUC) are utilized to assess the performance of the classifier. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings to illustrate the diagnostic ability of a binary classifier system. This measure is related to the evaluation criteria TPR, which can be denoted as follows:

$$\mathrm{TPR} = \frac{TP}{TP + FN} \quad (3)$$

where TP means true positive conditions, and FN means the false negative conditions. TPR is also known as sensitivity or recall. Another measurement is true negative rate (TNR), which is also known as specificity or selectivity; TNR can be denoted as follows:

$$\mathrm{TNR} = \frac{TN}{TN + FP} \quad (4)$$

Another measurement is FPR, which can be denoted as follows:

$$\mathrm{FPR} = \frac{FP}{FP + TN} = 1 - \mathrm{TNR} \quad (5)$$

where FP indicates the false positive results and TN indicates the true negative results. AUC means the AUC. The AUC of each iteration is shown in **Table 1**. After the model is trained, we use the model to predict the candidate genes.

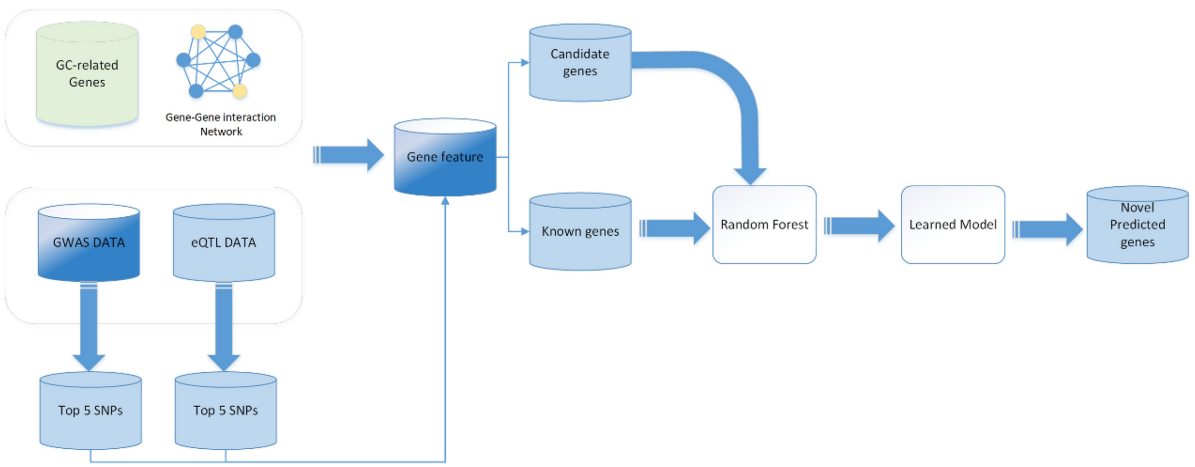


FIGURE 1 | Workflow of gastric cancer-related genes prediction.

TABLE 1 | Area under the curve of 10-fold cross-validation.

	AUC
1	0.881
2	0.91
3	0.894
4	0.886
5	0.878
6	0.872
7	0.896
8	0.892
9	0.902
10	0.909
Average	0.892

TABLE 2 | Predicted susceptible genes by random forest.

Symbol	Ensembl ID	Symbol	Ensembl ID
ANLN	ENSG00000011426	CD38	ENSG00000004468
TYROBP	ENSG00000011600	PDK4	ENSG00000004799
MATR3	ENSG00000015479	RARB	ENSG00000077092
NCDN	ENSG00000020129	ITGA2B	ENSG00000005961
TYMP	ENSG00000025708	CRLF1	ENSG00000006016
ALG1	ENSG00000033011	NRAS	ENSG00000213281
KRAS	ENSG00000133703	GTF2IRD1	ENSG00000006704
CDH1	ENSG00000039068	CACNA2D2	ENSG00000007402
BEST2	ENSG00000039987	DNAJC11	ENSG00000007923
TNFRSF17	ENSG00000048462	RPS20	ENSG00000008988
ADAMTS6	ENSG00000049192	GIPR	ENSG00000010310
PLAUR	ENSG00000011422	SLC6A7	ENSG00000011083

RESULTS

Performance Comparison Over All Known Disease Genes

In this study, we use all known GC-related genes obtained from DisGeNET as positive training samples, and genes with correlation scores under 1 were obtained from HumanNet v2.0 as negative training samples. We compared the predictive performance of RF, SVM, NB, and DNN. After 10-fold cross-validation of each method, the ROC curve and average AUC value is shown in **Figure 2**. The AUC of RF is 0.892, followed by the AUC of 0.811 of DNN, an AUC of 0.753 of SVM, and an AUC of 0.593 of NB, which means RF did the best performance in disease gene classification. For details, the AUC value of each iteration in RF classifier is shown in **Table 1**.

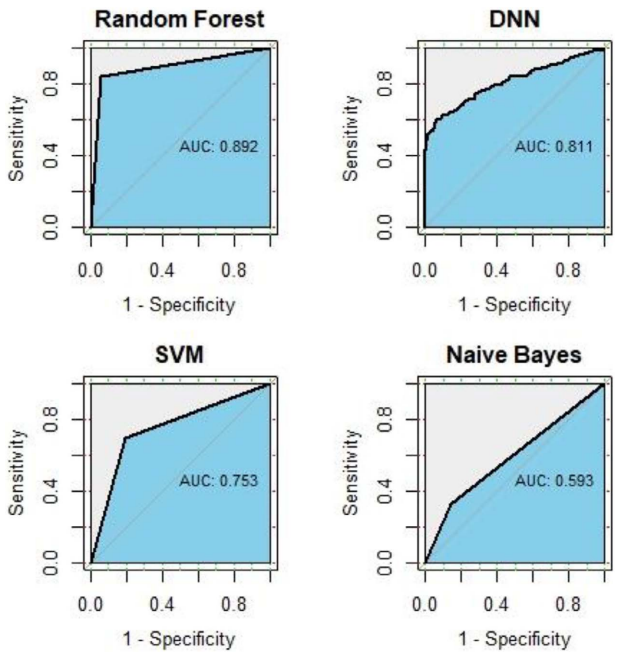


FIGURE 2 | Comparison of prediction performance.

Case Study

We obtained 7,406 candidate genes from HumanNet v2.0, and then extracted the feature representation of each gene. Then we used the trained RF classifier to predict the prioritizing genes associated with GC. The top 24 susceptible genes are shown in **Table 2**. From **Table 2**, 11 of the 24 predicted genes are reported to have direct or indirect association with GC. For example, CD38 has been determined to be expressed at higher levels in the IL-10-producing Breg cells of GC patients (Wang et al., 2015). The study of Cheng et al. found that compared to control tissues, RARB messenger RNAs were significantly reduced in human gastric tumor samples (Cheng et al., 2013). As an indirect evidence, Wen et al. found that NRAS can be a target gene of miR-26a to improve the sensitivity of GC cells to cisplatin-based chemotherapies, which can be an evidence for the potential function of NRAS in chemotherapy for GC (Wen et al., 2015). GIPR, short for gastric inhibitory polypeptide receptor, has been regarded as a promising target for imaging and therapy in gastric and neuroendocrine tumors, and it has also been reported that GIPR is significantly overexpressed in stomach tissue compared with normal tissue (Sherman et al., 2013, 2014). In recent years, many studies have shown that PLAUR can be an effective prognostic biomarker and potential therapeutic target for GC due to the fact that the suppression of PLAUR could sensitize cancer cell death by inducing DNA damages (Li et al., 2013a; Ai et al., 2020). ANLN is a conserved actin-binding protein that exerts its functions in cytoskeletal dynamics during cell division and may affect cancer progression through Wnt/B-catenin pathway in GC.

CONCLUSION

Gastric cancer is one of the most malignant neoplasms in human health around the world causing approximately 10% of all cancer deaths. The main therapeutic strategies of GC include two ways: surgery and chemotherapeutic regimens. Thus, it is important to identify the susceptible genes in order to better understand the pathologies of the disease, which can further help in drug designing. Machine learning methods have been used in predicting the functions of unclassified or unannotated

genes by utilizing genomic features (Schläpfer et al., 2017). In this study, we combined machine learning methods with genetic association data (GWAS analysis) and gene expression data (eQTL) to advance our understanding of GC etiology and pathology. Considering the LD, genes located close to susceptible loci identified by GWAS analysis may not be the causal genes of the disease. Since SNPs may also influence the expression level of the gene, the genes with different genotypes of the genetic variant will show differences in phenotype, which means that SNPs can also show effects on the diseases or traits. Therefore, we performed a RF classier on the collected 1,091 cases and 410,350 controls from GWAS dataset, integrated with stomach tissue eQTL data, to identify genes whose expression levels were associated with GC due to its causality. Compared to three other widely used binary classifiers, SVM, NB, and DNN, RF has the best performance in classifying GC-related genes. Since the overall importance score from the RF classifier is a sum of multiple individual importance scores, and each individual importance score is obtained from an average over multiple trees and cross-validations, the gene importance score can be regarded as not being affected by LD.

It is widely known that the accuracy of the prediction of genetic risk of complex diseases varies greatly between different diseases due to the heritability, phenotype, and the power and the amount of reported variants. Though GC is the most common cancer and causes a high mortality rate around the world, most studies only focused on the prognosis and treatment of GC. In this study, we identified 787 novel susceptible genes related to GC and focused on the top 24 of the susceptible genes. We found that CD38, RARB, NRAS, GIPR, PLAUR, ANLN, etc. have strong association with GC and have been reported to be related with GC indirectly; for example, they impact other pathways or exert their function by cooperating with other genes. However, we identified 787 novel susceptible genes related to GC, which is helpful in further studies to understand the etiology and pathology of GC and may be a strong theoretical basis for drug designing.

We used a machine learning method to identify GC-related genes. Although we have proven the effectiveness of our method by cross-validation and accuracy of our results by case study, biological experiments are still necessary to further reveal pathogenic mechanisms.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: GWAS dataset: downloaded from GWAS Catalog, accession: GCST90011807.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

HY and JL designed the study. HY, YD, and SX interpreted the data. HY, JS, and YQ analyzed the results. HY, YD, and JL were major contributors in writing and revising the manuscript. All authors read and approved the final manuscript.

REFERENCES

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2006). Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 22, 773–774. doi: 10.1093/bioinformatics/btk031

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., et al. (2006). Gene prioritization through genomic data fusion. *Nat. Biotechnol.* 24, 537–544. doi: 10.1038/nbt1203

Ai, C., Zhang, J., Lian, S., Ma, J., Győrffy, B., Qian, Z., et al. (2020). FOXM1 functions collaboratively with PLAU to promote gastric cancer progression. *J. Cancer* 11, 788–794. doi: 10.7150/jca.37323

Barral, S., Bird, T., Goate, A., Farlow, M., Diaz-Arrastia, R., Bennett, D., et al. (2012). Genotype patterns at PICALM, CR1, BIN1, CLU, and APOE genes are associated with episodic memory. *Neurology* 78, 1464–1471. doi: 10.1212/wnl.0b013e3182553c48

Basford, K. E., McLachlan, G. J., and Rathnayake, S. I. (2013). On the classification of microarray gene-expression data. *Brief. Bioinform.* 14, 402–410. doi: 10.1093/bib/bbs056

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics* 26, 2924–2926. doi: 10.1093/bioinformatics/btq538

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Calvo, S., Jain, M., Xie, X., Sheth, S. A., Chang, B., Goldberger, O. A., et al. (2006). Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.* 38, 576–582. doi: 10.1038/ng1776

Chen, J., Xu, H., Aronow, B. J., and Jegga, A. G. (2007). Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* 8:392. doi: 10.1186/1471-2105-8-392

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer statistics in China, 2015. *CA Cancer J. Clin.* 66, 115–132. doi: 10.3322/caac.21338

Cheng, A. S., Li, M. S., Kang, W., Cheng, V. Y., Chou, J. L., Lau, S. S., et al. (2013). *Helicobacter* pylori causes epigenetic dysregulation of FOXD3 to promote gastric carcinogenesis. *Gastroenterology* 144, 122–133. doi: 10.1053/j.gastro.2012.10.002

Cui, S. J., Ganjawala, T. H., Abrams, G. W., and Pan, Z. H. (2020). Effect of proteasome inhibitors on the AAV-mediated transduction efficiency in retinal bipolar cells. *Curr. Gene Ther.* 19, 404–412. doi: 10.2174/1566523220666200211111326

DeSantis, C. E., Lin, C. C., Mariotto, A. B., Siegel, R. L., Stein, K. D., Kramer, J. L., et al. (2014). Cancer treatment and survivorship statistics, 2014. *CA Cancer J. Clin.* 64, 252–271. doi: 10.3322/caac.21235

Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. (2014). Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Brief. Bioinform.* 15, 734–747. doi: 10.1093/bib/bbt056

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., et al. (2014). Bayesian test for colocalisation between pairs of

genetic association studies using summary statistics. *PLoS Genet.* 10:e1004383. doi: 10.1371/journal.pgen.1004383

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252. doi: 10.1038/ng.3506

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2018). HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* 47, D573–D580. doi: 10.1093/nar/gky1126

Li, D., Wei, P., Peng, Z., Huang, C., Tang, H., Jia, Z., et al. (2013a). The critical role of dysregulated FOXM1–PLAUR signaling in human colon cancer progression and metastasis. *Clin. Cancer Res.* 19, 62–72. doi: 10.1158/1078-0432.ccr-12-1588

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., et al. (2013b). Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front. Genet.* 4:103. doi: 10.3389/fgene.2013.00103

Lu, X. X., and Zhao, S. Z. (2019). Gene-based therapeutic tools in the treatment of cornea disease. *Curr. Gene Ther.* 19, 7–19. doi: 10.2174/1566523219666181213120634

Ludwig, J. A., and Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* 5, 845–856. doi: 10.1038/nrc1739

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

Mocellin, S., Verdi, D., Pooley, K. A., and Nitti, D. (2015). Genetic variation and gastric cancer risk: a field synopsis and meta-analysis. *Gut* 64, 1209–1219. doi: 10.1136/gutjnl-2015-309168

Park, B., Yang, S., Lee, J., Woo, H. D., Choi, I. J., Kim, Y. W., et al. (2019). Genome-wide association of genetic variation in the PSCA gene with gastric cancer susceptibility in a Korean population. *Cancer Res. Treat.* 51, 748–757. doi: 10.4143/crt.2018.162

Piñero, J., Bravo, À, Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Rashkin, S. R., Graff, R. E., Kachuri, L., Thai, K. K., Alexeeff, S. E., Blatchins, M. A., et al. (2020). Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat. Commun.* 11:4423. doi: 10.1038/s41467-020-18246-6

Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* 173, 2041–2059. doi: 10.1104/pp.16.01942

Sherman, S. K., Carr, J. C., Wang, D., O'Dorisio, M. S., O'Dorisio, T. M., and Howe, J. R. (2013). Gastric inhibitory polypeptide receptor (GIPR) is a promising target for imaging and therapy in neuroendocrine tumors. *Surgery* 154, 1206–1214. doi: 10.1016/j.surg.2013.04.052

Sherman, S. K., Maxwell, J. E., Carr, J. C., Wang, D., O'Dorisio, M. S., O'Dorisio, T. M., et al. (2014). GIPR expression in gastric and duodenal neuroendocrine tumors. *J. Surg. Res.* 190, 587–593. doi: 10.1016/j.jss.2014.01.044

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22, 2141–2150. doi: 10.1093/bib/bbaa044

Turner, F. S., Clutterbuck, D. R., and Semple, C. A. (2003). POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.* 4:R75. doi: 10.1186/gb-2003-4-11-r75

Wang, W., Yuan, X., Chen, H., Xie, G., Ma, Y., Zheng, Y., et al. (2015). CD19+ CD24hiCD38hiBregs involved in downregulate helper T cells and upregulate regulatory T cells in gastric cancer. *Oncotarget* 6, 33486–33499. doi: 10.18632/oncotarget.5588

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. *Brief. Funct. Genomics* 10, 280–293. doi: 10.1093/bfgp/elr024

Wang, Z., Dai, J., Hu, N., Miao, X., Abnet, C. C., Yang, M., et al. (2017). Identification of new susceptibility loci for gastric non-cardia adenocarcinoma: pooled results from two Chinese genome-wide association studies. *Gut* 66, 581–587. doi: 10.1136/gutjnl-2015-310612

Wen, L., Cheng, F., Zhou, Y., and Yin, C. (2015). MiR-26a enhances the sensitivity of gastric cancer cells to cisplatin by targeting NRAS and E2F2. *Saudi J. Gastroenterol.* 21, 313–319. doi: 10.4103/1319-3767.166206

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22, 2800–2805. doi: 10.1093/bioinformatics/btl467

Yip, K. Y., Cheng, C., and Gerstein, M. (2013). Machine learning and genome annotation: a match meant to be? *Genome Biol.* 14:205. doi: 10.1186/gb-2013-14-5-205

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 13:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021a). Prediction and collection of protein–metabolite interactions. *Brief. Bioinform.* 12:bbab014. doi: 10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Zheng, R., Zeng, H., Zhang, S., and Chen, W. (2017). Estimates of cancer incidence and mortality in China, 2013. *Chin. J. Cancer* 36:66. doi: 10.1186/s40880-017-0234-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Yaoxing, Danchun, Xiaojuan, Shuman, Qingqing and Lin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 21 July 2021
doi: 10.3389/fcell.2021.715275

Identification of Somatic Mutation-Driven Immune Cells by Integrating Genomic and Transcriptome Data

Ying Jiang[1†], Baotong Zheng[2†], Yang Yang[2†], Xiangmei Li[2†] and Junwei Han[2]*

[1] College of Basic Medical Science, Heilongjiang University of Chinese Medicine, Harbin, China, [2] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Tumor somatic mutations in protein-coding regions may generate neoantigens which may trigger antitumor immune cell response. Increasing evidence supports that immune cell response may profoundly influence tumor progression. However, there are no calculated tools to systematically identify immune cells driven by specific somatic mutations. It is urgent to develop a calculated method to comprehensively detect tumor-infiltrating immune cells driven by the specific somatic mutations in cancer. We developed a novel software package (SMDIC) that enables the automated identification of somatic mutation-driven immune cell. SMDIC provides a novel pipeline to discover mutation-specific immune cells by integrating genomic and transcriptome data. The operation modes include inference of the relative abundance matrix of tumor-infiltrating immune cells, detection of differential abundance immune cells with respect to the gene mutation status, conversion of the abundance matrix of significantly dysregulated cells into two binary matrices (one for upregulated and one for downregulated cells), identification of somatic mutation-driven immune cells by comparing the gene mutation status with each immune cell in the binary matrices across all samples, and visualization of immune cell abundance of samples in different mutation status for each gene. SMDIC provides a user-friendly tool to identify somatic mutation-specific immune cell response. SMDIC may contribute to understand the mechanisms underlying anticancer immune response and find targets for cancer immunotherapy. The SMDIC was implemented as an R-based tool which was freely available from the CRAN website https://CRAN.R-project.org/package=SMDIC.

Keywords: multi-omics data, somatic mutations, tumor-infiltrating immune cells, tumor microenvironment, gene expression

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Jing Li,
Fujian Medical University, China
Juan Wang,
Inner Mongolia University, China

****Correspondence:***
Junwei Han
hanjunwei1981@163.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *26 May 2021*
Accepted: *25 June 2021*
Published: *21 July 2021*

Citation:
Jiang Y, Zheng B, Yang Y, Li X and Han J (2021) Identification of Somatic Mutation-Driven Immune Cells by Integrating Genomic and Transcriptome Data. Front. Cell Dev. Biol. 9:715275. doi: 10.3389/fcell.2021.715275

INTRODUCTION

Tumor generally harbors somatic mutations, some of which in protein-coding regions may generate neoantigens which may trigger antitumor immune cell response in the tumor microenvironment (TME) (Yang et al., 2019). The mutation-reactive immune cells are usually found infiltrating into solid tumors (Robbins et al., 2013), and they profoundly influence

tumor initiation, progression, metastasis, and treatment response (Chen and Mellman, 2017). For example, Tran et al. (2014) identify erbb2 interacting protein (ERBB2IP) mutation-specific $CD4^+$ T (TH1) cells in a patient with metastatic cholangiocarcinoma and demonstrate that the ERBB2IP mutated-specific TH1 cell response can be used to mediate regression of metastatic epithelial cancer. Therefore, identifying mutation-specific immune cell response will illustrate the mechanisms underlying the anticancer immune response and might help to find targets for cancer immunotherapy (Yang et al., 2019). However, there are very few calculated tools that could comprehensively identify which immune cells are triggered by which mutations. Thus, it is urgent to develop a calculated method to comprehensively detect tumor-infiltrating immune cells driven by the specific somatic mutations in cancer. Recently, a number of computational approaches have been developed for quantifying tumor-infiltrating immune cells using tumor bulk gene expression data (Newman et al., 2015; Senbabaoglu et al., 2016; Aran et al., 2017). This will help to examine the correlation among somatic mutations and immune cells by using high-throughput sequencing data, for example whole-exome sequencing and RNA-seq.

Here, we developed a novel software package, SMDIC, to identify immune cells driven by specific somatic mutations by integrating genomic and transcriptome data, which is available at https://CRAN.R-project.org/package=SMDIC. SMDIC also provides visualization of the relative abundance of identified immune cells in the gene mutation status using heat maps, as well as the waterfall plot of mutation genes correlated with immune cells and mutually exclusive and co-occurring plot. We used the Genomic Data Commons (GDC) TCGA breast cancer somatic mutation and expression data to show the applications of the package and the visualization of the results. The results showed that SMDIC could effectively identify immune cells which may be driven by the specific somatic mutation genes, with strong robustness.

MATERIALS AND METHODS

SMDIC has three main functions (**Figure 1**): (a) inferring the relative abundance matrix of tumor-infiltrating immune cells, (b) detecting differential abundance immune cells with respect to a particular gene mutation status and converting the abundance matrix of significantly dysregulated immune cells into two binary matrices (one for upregulated and one for downregulated cells), and (c) identifying somatic mutation-driven immune cells by comparing the gene mutation status with each immune cell in the binary matrices across all samples.

Inference of the Relative Abundance Matrix of Immune Cells

Recently, various computational approaches have been developed for quantifying tumor-infiltrating immune cells from gene expression data of human bulk tumors. The SMDIC package provides three methods for estimating the relative infiltration abundance of different cell types in the tumor microenvironment (TME), namely, xCell (Aran et al., 2017), ssGSEA estimated method using the signatures of (Bindea et al., 2013; Senbabaoglu et al., 2016), and CIBERSORT (Newman et al., 2015). Through these methods, the relative abundance matrix of immune cells is inferred with cells as rows and samples as columns (**Figure 1A**). In the package, the default method is xCell, which estimates abundances of 64 cell types including adaptive and innate immunity cells, hematopoietic progenitors, epithelial cells, and extracellular matrix cells (Aran et al., 2017). The users can choose the method they are interested in.

Detection of Differential Abundance Immune Cells With Respect to the Gene Mutation Status

The somatic mutation data used in the package are the mutation annotation file (MAF) format. We extract the non-silent somatic mutations (non-sense mutation, missense mutation, frame-shift indels, splice site, non-stop mutation, translation start site, inframe indels) in protein-coding regions and built a binary mutations matrix, in which 1 represents any mutation occurs in a particular gene in a particular sample; otherwise, the element is 0. The genes with a particular mutation frequency greater than a given threshold value are retained for the following analysis. Then, for a given mutation gene, the significance analysis of microarrays (SAM) method (Tusher et al., 2001) is used to detect differential abundance immune cells with respect to the mutation status (defined by binary mutation vector) across all the samples. With a false discovery rate (FDR) < 0.05 in SAM, the immune cells were deemed as significantly dysregulated between mutation and non-mutation samples and were considered to be correlated with the gene mutation status. Then, the abundance matrix elements of significantly dysregulated immune cells are converted to z-scores by row. For the upregulated (or downregulated) immune cells, any element with a z-score > 2.0 (or ≤ 2.0) is assigned with 1; otherwise, the element is 0 (**Figure 1B**). Thus, we obtain two binary abundance matrices for upregulated and downregulated immune cells, respectively.

Identification of Somatic Mutation-Driven Immune Cells

For a particular mutation gene, we compare the binary mutation vector with each binary cell abundance vector in the upregulated or downregulated immune cell–matrix, respectively, and a 2×2 contingency table is constructed. The Fisher's exact test is applied to recover the cells that had drastic mutation-correlated upregulated or downregulated response (**Figure 1C**). This process is repeated for each immune cell in the binary abundance matrices. To correct for multiple comparisons, we adjust the exact test p-values by using the FDR method proposed by Benjamini and Hochberg (1995). The immune cells with the default FDR < 0.05 are deemed as statistically significant mutation-correlated and may be driven by the somatic mutation. We repeat the above process for each mutated gene.

To further test if the somatic mutation-specific immune cells are associated with the prognosis of tumor patients, we constructed a signature by using immune cells driven by a

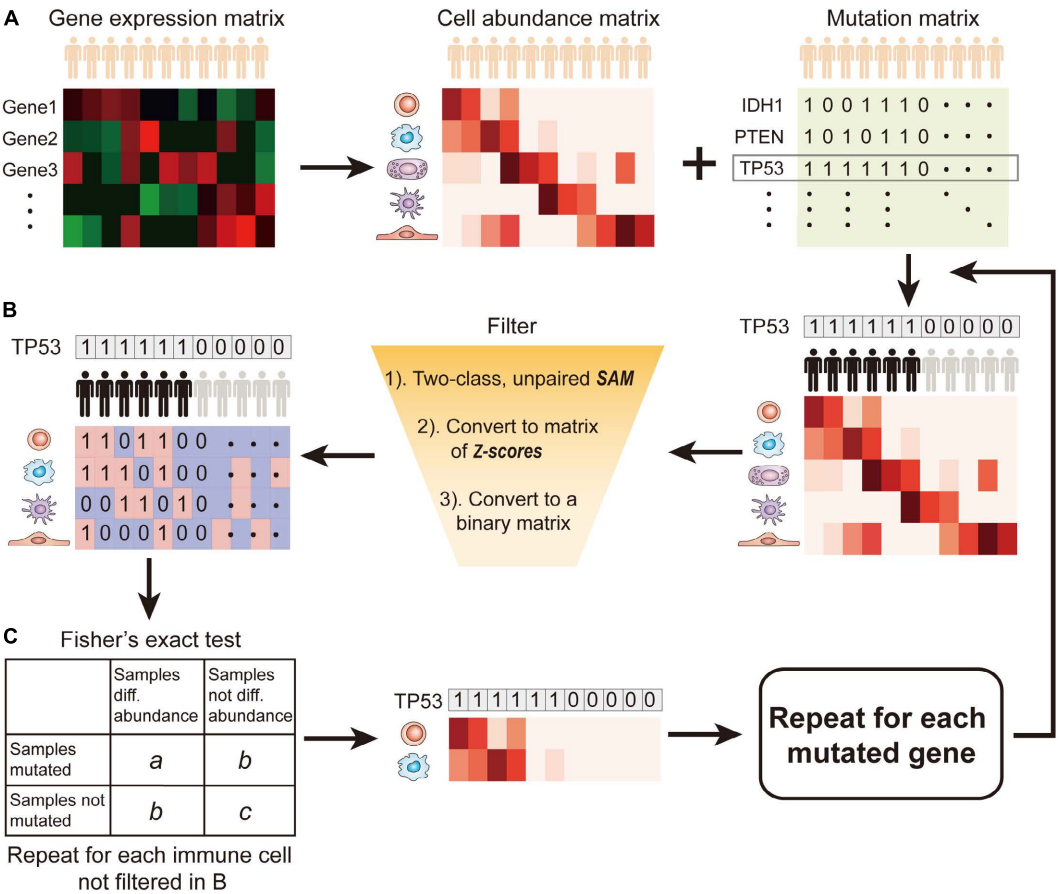


FIGURE 1 | Flow diagram of SMDIC. **(A)** Inference of the relative abundance matrix of immune cells; **(B)** Detection of differential abundance immune cells with respect to the gene mutation status; **(C)** Identification of significant immune cells associated with somatic mutation.

particular gene mutation. The immune cell risk score was calculated with a prognostic score model for each patient based on the abundances of cells. The model is as follows:

$$Risk\ score = \sum_{k \in S} \beta_k a_k \quad (1)$$

where S is the set of cells correlated with a particular gene mutation, a_k is the abundance of cell k, and β_k is the regression coefficient of a multivariate Cox proportional hazard regression model estimated on a_k and the overall survival data. According to the median of risk scores of the immune cell signature, patients are classified into high-risk and low-risk groups.

RESULTS

We selected the GDC TCGA Breast Cancer (BRCA) data cohorts[1] as examples to explain the application of the SMDIC package. We downloaded the MAF file (derived from VarScan 2) from the GDC data portal, which is obtained from TCGA whole-exome sequencing (WES) data. Meantime, the RNA-seq FPKM gene expression data and the associated clinical data were also downloaded. These data have been deposited in **Supplementary Data**.

Identification of Immune Cells Driven by Somatic Mutations in Breast Cancer

For the expression data, log-transformed FPKM expression values were used and were inputted into the "exp2cell" function for inferring the relative abundance matrix of infiltrated immune cells. The "exp2cell" function provides three methods for inferring the cell abundance matrix, namely, xCell (Aran et al., 2017), ssGSEA estimated method using the immune signatures of Bindea et al. (2013; Senbabaoglu et al., 2016), and CIBERSORT (Newman et al., 2015). The users can use the argument "method" to select the method they are interested in. The default argument is "xCell" method. Thus, a cell abundance matrix with 64 cells and 974 samples was obtained.

We then extracted the non-silent somatic mutations (non-sense mutation, missense mutation, frame-shift indels, splice site, non-stop mutation, translation start site, inframe indels) in protein-coding regions from the MAF file and built a binary mutation matrix, in which 1 represents any mutation occurs in a particular gene in a particular sample; otherwise, the element is 0. The "maf2matrix" function was used to implement this process.

[1]https://portal.gdc.cancer.gov/

TABLE 1 | The summary results of immune cells driven by somatic mutation genes.

Gene	Cells	Cell count	Mutation rate
TP53	Astrocytes, $CD8^+$ naive T cells, $CD8^+$ Tem, DC, epithelial cells, keratinocytes, macrophages, macrophages M1, melanocytes, NK cells, pDC, pericytes, plasma cells, pro-B cells, sebocytes, Tgd cells, Th1 cells, Th2 cells, Tregs	19	0.34
CDH1	Adipocytes, $CD4^+$ Tcm, chondrocytes, CMP, endothelial cells, fibroblasts, HSC, ly endothelial cells, megakaryocytes, mesangial cells, mv endothelial cells, NKT	12	0.13
NBPF14	$CD8^+$ Tem, DC, iDC, keratinocytes, pro-B cells, sebocytes	6	0.013
OBSCN	Epithelial cells, keratinocytes, pro-B cells, sebocytes, Th1 cells	5	0.030
NF1	aDC, B cells, memory B cells, pDC, plasma cells	5	0.037
ANKRD30A	Keratinocytes, sebocytes, Th1 cells, Th2 cells	4	0.015
ARAP3	$CD8^+$ naive T cells, NK cells, preadipocytes, Th1 cells	4	0.010
NOS1	$CD4^+$ Tem, $CD8^+$ T cells, $CD8^+$ Tcm, pDC	4	0.012
PCDH19	$CD4^+$ memory T cells, $CD4^+$ T cells, $CD8^+$ Tem, NK cells	4	0.018
KCNA4	$CD8^+$ T cells, $CD8^+$ Tcm, NK cells, pDC	4	0.011
SCN5A	$CD8^+$ T cells, $CD8^+$ Tcm, plasma cells	3	0.013
ACTN2	NK cells, pDC, Th2 cells	3	0.011
FBXW7	MEP, Th1 cells, Th2 cells	3	0.016
TRPS1	$CD4^+$ Tem, MSC, skeletal muscle	3	0.011

The mutation genes correlated at least three significant cells are shown.

Alternatively, the users can directly input a binary mutation matrix which may not be derived from the MAF file. This will increase the usability of our package. With a given mutation frequency threshold (the default value is 1%), a binary matrix with 821 mutations and 974 samples was obtained.

In the SMDIC package, the "mutcorcell" function is implemented to identify somatic mutation-specific immune cell response by inputting the abundance matrix of immune cells and binary mutations matrix. This function firstly detects differential abundance immune cells with respect to a particular gene mutation status with the SAM method (Tusher et al., 2001). The immune cells with the default FDR < 0.05 in SAM are deemed as significantly dysregulated between mutation and non-mutation samples and are considered to be correlated with the gene mutation status. The significantly dysregulated immune cells are extracted and their abundance matrix elements are converted to z-scores by row. Then, the z-score abundance matrix is converted to two binary abundance matrices for upregulated and downregulated immune cells, respectively. Specifically, for the upregulated (or downregulated) immune cells, any element with a z-score > 2.0 (or ≤ 2.0) is assigned with 1; otherwise, the element is 0 (**Figure 1B**). Finally, for a particular mutation gene, the binary mutation vector is compared with each binary cell abundance vector in the upregulated and downregulated cell matrices, respectively, and a 2×2 contingency table is constructed. Moreover, the Fisher's exact test is applied to recover the cells that had drastic mutation-correlated upregulated or downregulated response. This process is repeated for each immune cell in the binary abundance matrices. To correct for multiple comparisons, the exact test p-values are adjusted by using the FDR method (Benjamini and Hochberg, 1995), and the immune cells with the default FDR < 0.05 are deemed as statistically significant mutation-correlated and may be driven by the somatic mutation. The above process is repeated for each mutated gene.

The "mutcorcell" function will output the summary results of somatic mutation-driven immune cells and the detailed information of immune cells for each mutation, which are used as input data for the visualization functions. The detailed commands are listed in the **Supplementary Material**, and the GDC TCGA breast cancer datasets (mutation data, gene expression data, and survival data of patients) are stored in **Supplementary Data**. The summary results of immune cells driven by somatic mutation genes are listed in **Table 1**, and many of them have been reported in the recent literature.

TABLE 2 | The detailed information of immune cells driven by the TP53 gene mutation.

Cell	Full name	*p*-value	FDR
Th2 cells	Type 2 T-helper cells	8.86E-13	4.16E-11
Sebocytes	Sebocytes	1.51E-09	3.55E-08
Macrophages M1	Macrophages M1	6.33E-09	9.91E-08
Astrocytes	Astrocytes	1.01E-07	1.19E-06
Keratinocytes	Keratinocytes	2.33E-07	2.02E-06
Th1 cells	Type 1 T-helper cells	2.58E-07	2.02E-06
Macrophages	Macrophages	0.00018	0.0012
Pericytes	Pericytes	0.00035	0.0020
Tregs	Regulatory T cells	0.00048	0.0025
NK cells	Natural killer cells	0.00054	0.0025
Epithelial cells	Epithelial cells	0.00086	0.0037
Pdc	Plasmacytoid dendritic cells	0.0020	0.0077
Tgd cells	Gamma delta T cells	0.0021	0.0077
Pro-B cells	Pro-B cells	0.0026	0.0089
$CD8^+$ naive T cells	$CD8^+$ naive T cells	0.0040	0.013
Melanocytes	Melanocytes	0.0052	0.015
DC	Dendritic cells	0.016	0.042
Plasma cells	Plasma cells	0.016	0.042
$CD8^+$ Tem	$CD8^+$ effector memory T cells	0.017	0.043

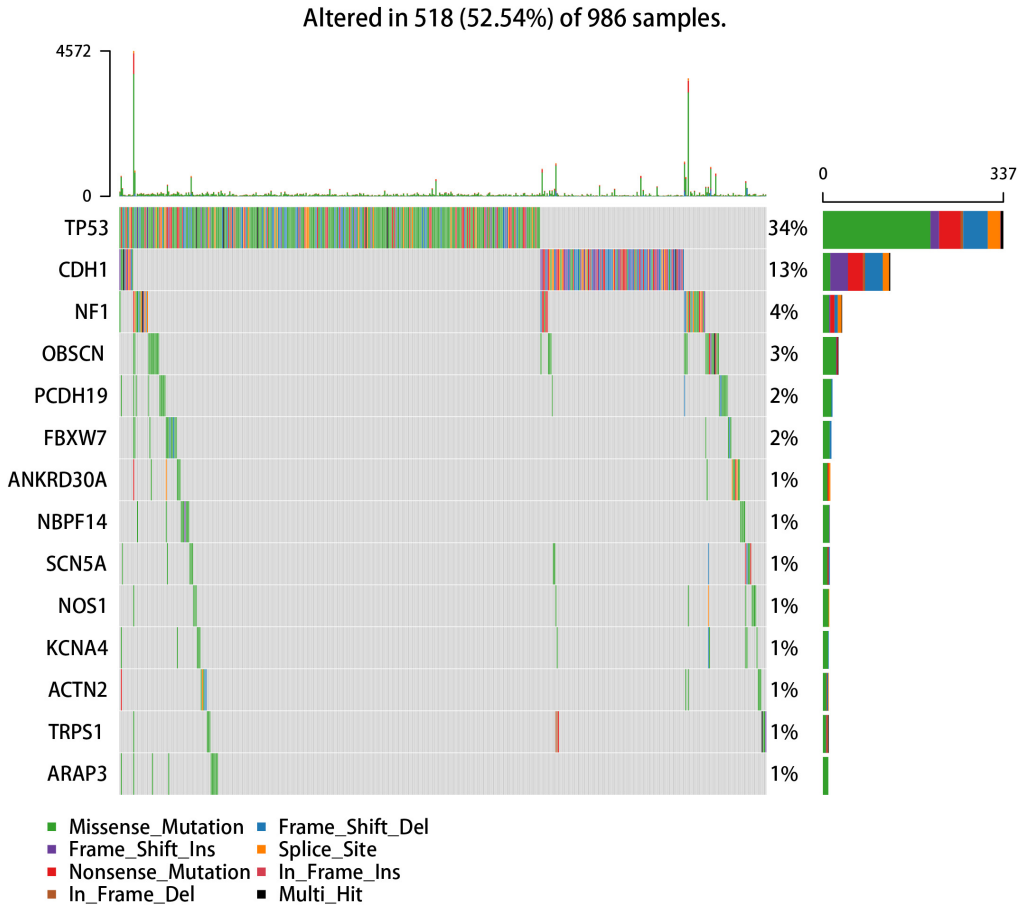


FIGURE 2 | Waterfall plot of mutation genes that drive immune cells in the GDC TCGA breast cancer dataset.

For example, regulatory T cells (Treg) have been found to be associated with a high mutation rate of TP53 genes in breast cancer (Oshi et al., 2020). The TP53 mutant has been identified to be associated with a higher expression of cytotoxic T-cell lymphocytes, natural killer (NK) cells, and Th1 genes characteristic of a proinflammatory immune cell signature (Agupitan et al., 2020). E-cadherin (CDH1) has been reported to target platelet endothelial cell adhesion molecule-1 (PECAM-1) for ubiquitination and degradation in endothelial cells (Liu et al., 2020), whose mutation may regulate endothelial cell homeostasis. Our results are consistent with these findings, and SMDIC could also identify some additional immune cells driven by these mutations, which may provide some new biological insights.

Moreover, most of these cells driven by somatic mutations have been proposed to be associated with the progress of breast cancer. For example, in the TP53 mutation, 18 cells were identified (**Table 2**). It was proposed that type 1 T-helper (Th1) cells eradicated tumor mass by inducing cellular immunity, and type 2 T-helper (Th2) cells destroyed the tumor by inducing tumor necrosis (Nishimura et al., 1999). Macrophages M1 cells generate interleukin (IL)-12 and tumor necrosis factor with antitumor effects in breast cancer cells (Jeong et al., 2019). NK cells were reported to be critical immune components in controlling breast tumor growth and dissemination (Liu et al., 2019). For the cells driven by CDH1, $CD4^+$ central memory T cell (Tcm) was proposed to be negatively associated with recurrence-free survival in breast cancer (Deng et al., 2019). The above results indicate that the SMDIC method could effectively identify immune-related cells driven by somatic mutations and may help to facilitate the development of cancer immunotherapy.

Presentation of the Results

We used the visualization function of the SMDIC package to demonstrate the analysis results of breast cancer. The detailed commands for these visualization functions are listed in **Supplementary Material**. The "plotwaterfall" function is applied to show the waterfall plot of mutation genes that drive immune cells (**Figure 2**). To further analyze the correlation between these mutation genes, we used the "plotCoocMutex" function to plot the co-occurrence and mutual exclusivity plots (**Figure 3**). TP53 and CDH1 show significant mutual exclusions ($p < 0.05$). For a better display of the immune cell response triggered by somatic mutation, we used the "heatmapcell" function to plot a heat map to show the significant difference of cell abundance between gene mutation and non-mutation status. We selected

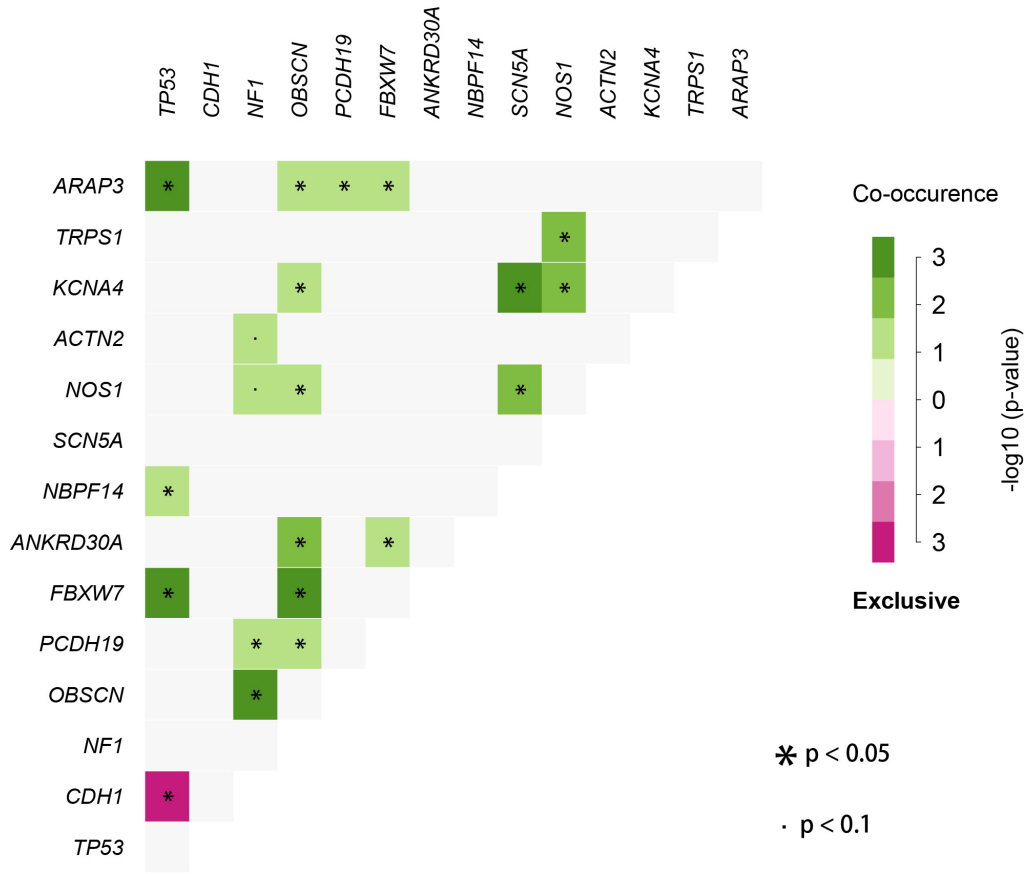


FIGURE 3 | Co-occurrence and mutual exclusivity plots between mutation genes correlated with immune cells.

the immune cells driven by TP53 as an example to show their heat map (**Figure 4**). We can see that the heat maps of immune cells display two obviously different blocks between TP53 mutation and non-mutation samples. From the heat maps of immune cells driven by CDH1, a similar result was obtained (**Supplementary Figure 1**). We then used the "survcell" function to plot the Kaplan–Meier survival curves of patients classified into high- and low-risk groups using the median of the immune cell risk score (see Implementation section). The immune cells driven by TP53, CDH1, NBPF14, and OBSCN mutations were, respectively, selected to calculate the immune risk score for plotting Kaplan–Meier survival curves (**Figures 5A–D**). It can be seen that immune cells driven by these mutations could be used as prognostic signatures. These results indicate that these mutations may be used as neoantigens for immunotherapy.

Robustness Analysis of the SMDIC Method

We first tested if the immune cells driven by the specific somatic mutation had the same or similar prognostic value in different independent datasets. As the public cancer datasets with both genome and transcriptome are seldom, we used the mutation-specific immune cell signatures (see *Implementation* section) obtained from the TCGA breast cancer dataset to test their prognosis performance in independent gene expression datasets. We enrolled four independent datasets of breast cancer, namely, GSE1992 (Hu et al., 2006), GSE3143 (Bild et al., 2006), GSE1456 (Pawitan et al., 2005), and GSE7390 (Desmedt et al., 2007) datasets. The cell signature derived from TP53 mutation was firstly used to classify patients from four independent breast cancer expression datasets, respectively. For each independent dataset, the "exp2cell" was applied to convert gene expression profiles to cell abundance profiles, and the patients were divided into high-risk and low-risk groups based on the cell signature. Interestingly, the *p*-values of log-rank tests in these four datasets were all < 0.05 (**Supplementary Figures 2A–D**). For the cell signature driven by CDH1, it was shown that the CDH1-specific cell signature could classify patients into high- and low-risk groups in three of the above four datasets (log-rank tests *p*-value < 0.05, **Supplementary Figures 3A–D**). We also found that the cell signatures derived from NBPF14 and OBSCN mutations respectively obtained similar results (**Supplementary Figures 4, 5**). These results confirmed that the mutation-driven immune cells could be used as prognostic signatures.

We then tested if the SMDIC method is robust to missing data in the gene expression profiles. To do this, we performed data removal tests using the GDC TCGA breast cancer expression

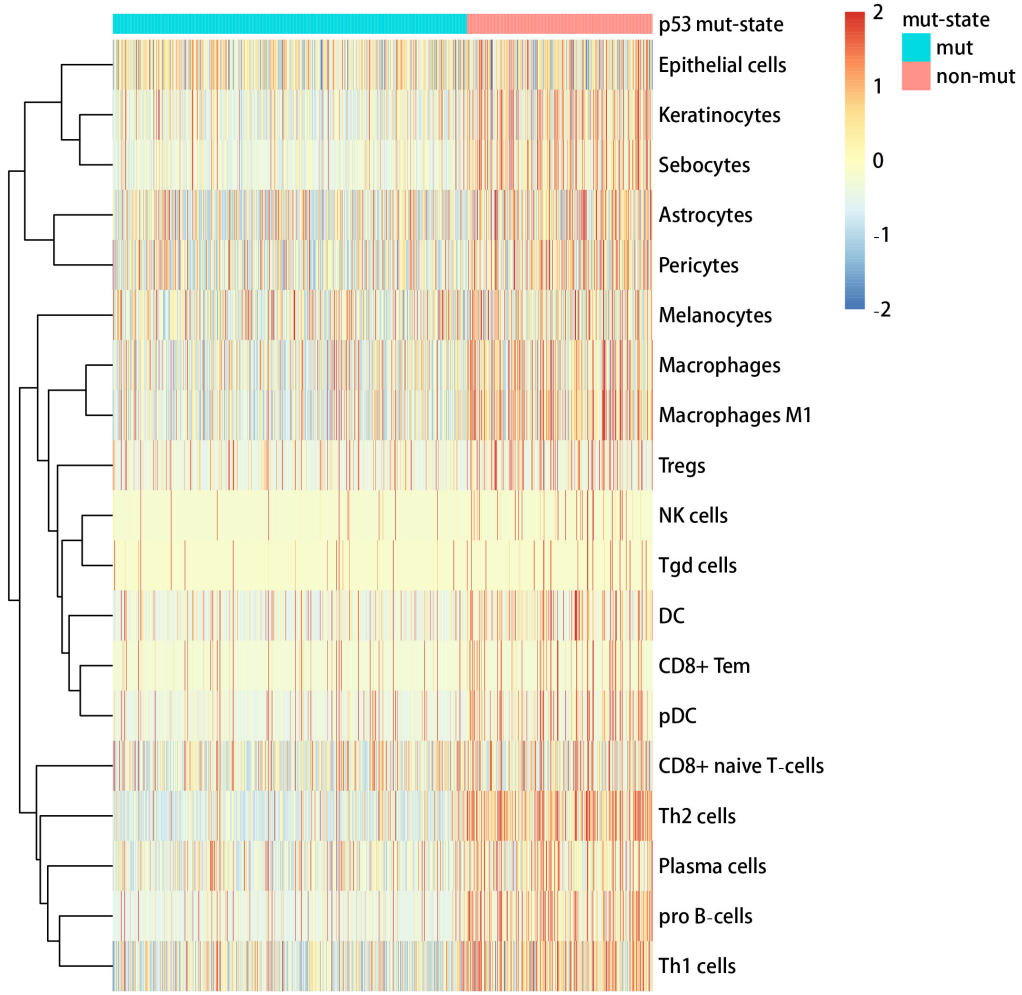


FIGURE 4 | Heat map of cell abundance between TP53 mutation and non-mutation status.

profiles. For the gene expression data, we randomly removed the genes in the original expression profiles from 5 to 20% at 5% intervals. For each data removal, we repeated the SMDIC method to identify somatic mutation-specific immune cells. We then compared these results with our original results. The mutation genes which trigger at least five immune cells, namely, TP53, CDH1, OBSCN, NF1 and NBPF14, were used to test how many significant immune cells are overlapped. With the data removal increasing, we found that the number of overlapped cells driven by each mutation gene fell slowly compared with the original data (**Figure 6**). Moreover, 89% (17/19), 66% (8/12), 60% (3/5), and 50% (3/6) cells were shared with our original results for the TP53, CDH1, NBPF14, and OBSCN mutations, respectively, even after removal of up to 15% of the gene expression data. These results indicate that the SMDIC method is robust to missing data in the gene expression profiles.

We further tested if the SMDIC method is robust to different methods for estimating abundance of immune cell infiltration. The default method for estimating cell abundance is xCell (Aran et al., 2017) in the package. We used an alternative ssGSEA estimated method (Senbabaoglu et al., 2016) to test if the SMDIC method could obtain consistent results. In the GDC TCGA breast cancer datasets, the SMDIC method identified 19 and 15 TP53 mutation-specific immune cell response in the xCell and ssGSEA cell abundance estimated methods (**Table 2** and **Supplementary Table 1**). There are only five overlapped immune cells, namely, type 1 T-helper (Th1) cell, type 2 T-helper (Th2) cell, gamma delta T cells (Tgd), regulatory T cells (Tregs), and NK cell. This is because xCell and ssGSEA use different cell estimated strategies and quantify different cell types. However, through calculating the Pearson correlation between xCell and ssGSEA values for relevant cell types driven by TP53 mutation, we observed relatively stronger levels of concordance (e.g., $CD8^+$ naive T cells vs. CD8 T cells, Pearson $r = 0.31$, p-value $<$ 1e-20; dendritic cells vs. activated dendritic cells, Pearson $r = 0.53$, p-value $<$ 1e-20) (**Supplementary Table 2**). These validation results show that our somatic mutation-driven immune cell method is robust to different cell abundance estimated methods.

To further illustrate the applicability of our SMDIC method, we applied it to TCGA skin cutaneous melanoma (SKCM) data,

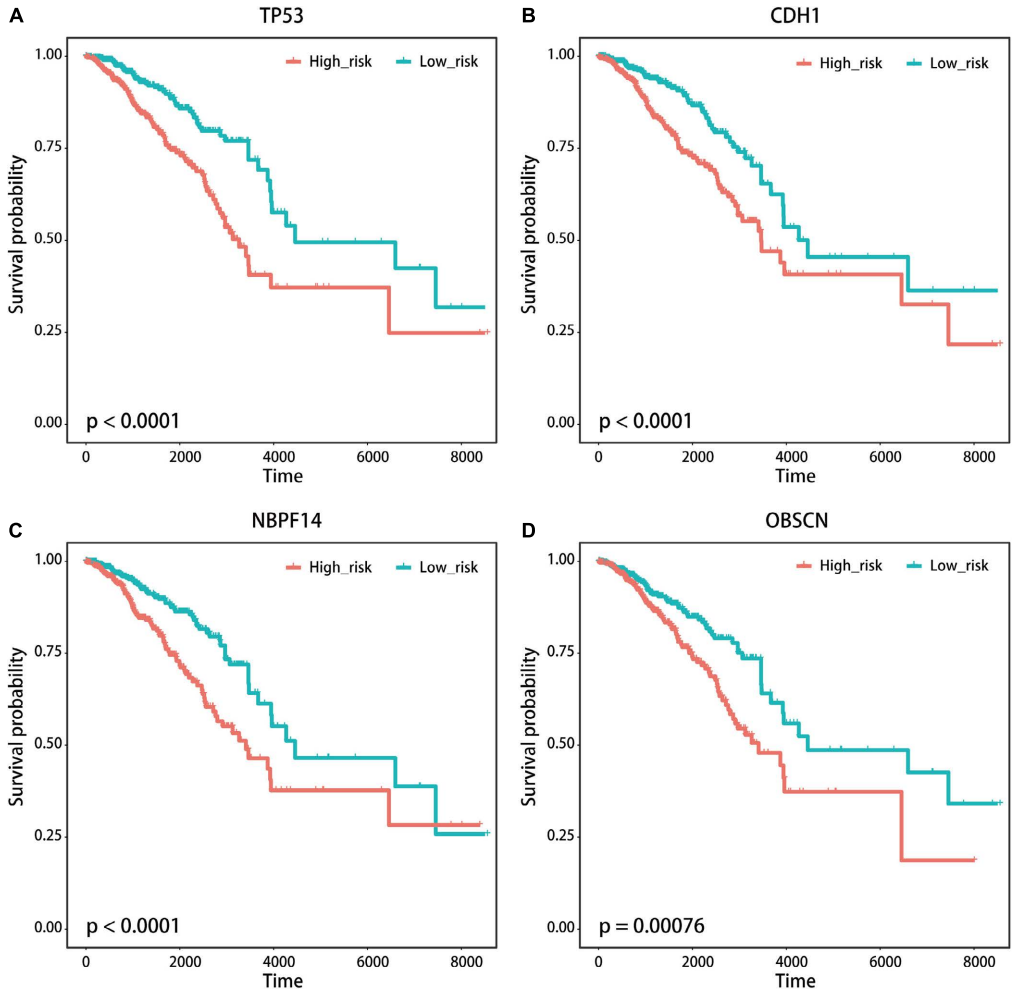


FIGURE 5 | Kaplan–Meier survival curves of patients with breast cancer classified into high- and low-risk groups based on the immune cells driven by **(A)** TP53, **(B)** CDH1, **(C)** NBPF14, and **(D)** OBSCN mutations, respectively.

which were downloaded from the GDC data portal (see text footnote 2). We inputted all the related data into the SMDIC package and performed the same operations like in the breast cancer data. The detailed results are listed in **Supplementary Table 3**. We found that some mutation cell associations identified by the method have been reported in the literature. For example, it has been proposed that deficiency of E3 ubiquitin ligase (ITCH) altered CD4 T-cell and B-cell responses in mice and humans (Field et al., 2020); HIVEP2 was found to play a crucial role in the control of Th2 cell differentiation by regulating NF-κB function, whose ortholog gene HIVEP3 may substitute for the function of HIVEP2 (Kimura et al., 2005). This indicates that the SMDIC method could identify mutation-specific cells in different cancers.

DISCUSSION

Recently, a number of calculated methods have been developed to identify dysregulated genes and pathways in cancer using molecular omics data (genome or transcriptome) (Cheng et al., 2019; Han et al., 2020; Sheng et al., 2021). Moreover, some methods were proposed to prioritize cancer candidate drugs and prognostic markers (Han et al., 2018, 2021; Di et al., 2019). However, the methods for identifying immune cells driven by specific somatic mutations are mainly based on biological experiments, and as far as we have known, no calculated tools could systematically identify mutation-specific immune cells.

In the study, SMDIC R package is an automated, computationally fast, and efficient tool for exploring somatic mutation-driven immune cell response by integrating genome and transcriptome data. The users input the somatic mutation data and gene expression data with the same samples; the package will return the immune cells triggered by each mutation and provide visualization of results. Moreover, our package is not specifically designed for breast cancers and can be generally applicable for various cancer datasets with genomic and transcriptome data. The SMDIC tool was designed to identify somatic mutation-specific immune cells by integrating high-throughput genomic and transcriptome data, which may

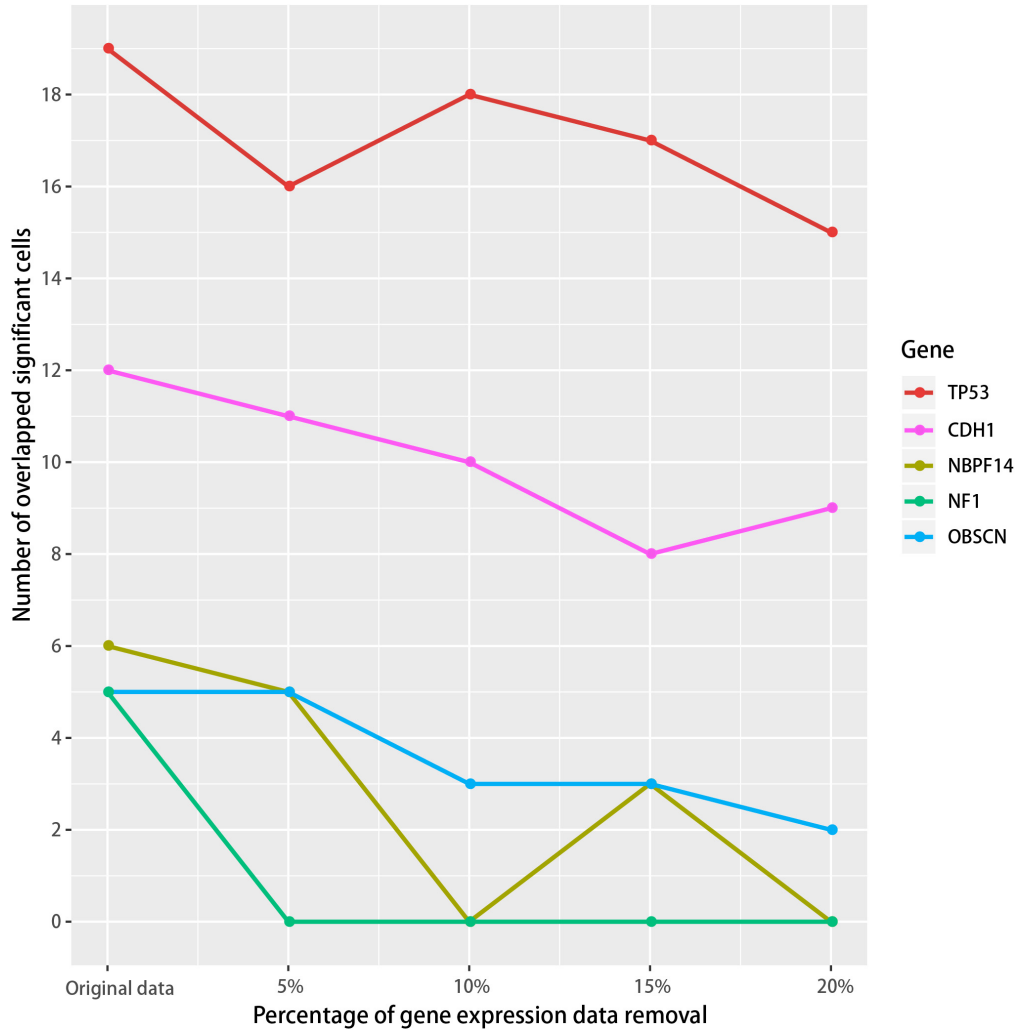


FIGURE 6 | Robustness analysis of the SMDIC method. The gene expression values were randomly removed from 5 to 20% at 5% intervals. For each removal, the overlapped number of the immune cells with the original results was calculated for the mutation genes which activate at least five immune cells.

help to find neoantigens and facilitate the development of personalized immunotherapy to patients with cancers.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JH and YJ conceived and designed the study. BZ developed software. YY and XL analyzed the data and implemented the methodology. YJ revised the manuscript. JH drafted the manuscript. All authors read and agreed to the final version of the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (grant nos. 62072145 and 81804158), the Natural Science Foundation of Heilongjiang Province (grant no. LH2019C042), and the University Student Innovation and Entrepreneurship Training Project of Heilongjiang Province (grant no. 201810226010).

ACKNOWLEDGMENTS

We thank the Training Center for Students Innovation and Entrepreneurship Education in Harbin Medical University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.715275/full#supplementary-material

REFERENCES

Agupitan, A. D., Neeson, P., Williams, S., Howitt, J., Haupt, S., and Haupt, Y. (2020). P53: a guardian of immunity becomes its saboteur through mutation. *Int. J. Mol. Sci.* 21:3452. doi: 10.3390/ijms21103452

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 18:220.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353–357. doi: 10.1038/nature04296

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 39, 782–795. doi: 10.1016/j.immuni.2013.10.003

Chen, D. S., and Mellman, I. (2017). Elements of cancer immunity and the cancer-immune set point. *Nature* 541, 321–330. doi: 10.1038/nature21349

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for Identifying Similar Diseases. *Mol. Ther. Nucleic Acids* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Deng, L., Lu, D., Bai, Y., Wang, Y., Bu, H., and Zheng, H. (2019). Immune Profiles of Tumor Microenvironment and Clinical Prognosis among Women with Triple-Negative Breast Cancer. *Cancer Epidemiol. Biomarkers Prev.* 28, 1977–1985. doi: 10.1158/1055-9965.epi-19-0469

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., et al. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* 13, 3207–3214. doi: 10.1158/1078-0432.ccr-06-2765

Di, J., Zheng, B., Kong, Q., Jiang, Y., Liu, S., Yang, Y., et al. (2019). Prioritization of candidate cancer drugs based on a drug functional similarity network constructed by integrating pathway activities and drug activities. *Mol. Oncol.* 13, 2259–2277. doi: 10.1002/1878-0261.12564

Field, N. S., Moser, E. K., and Oliver, P. M. (2020). Itch regulation of innate and adaptive immune responses in mice and humans. *J. Leukoc. Biol.* 108, 353–362. doi: 10.1002/jlb.3mir0320-272r

Han, J., Han, X., Kong, Q., and Cheng, L. (2020). psSubpathway: a software package for flexible identification of phenotype-specific subpathways in cancer progression. *Bioinformatics* 36, 2303–2305. doi: 10.1093/bioinformatics/btz894

Han, J., Liu, S., Jiang, Y., Xu, C., Zheng, B., Jiang, M., et al. (2018). Inference of patient-specific subpathway activities reveals a functional signature associated with the prognosis of patients with breast cancer. *J. Cell Mol. Med.* 22, 4304–4316. doi: 10.1111/jcmm.13720

Han, X., Kong, Q., Liu, C., Cheng, L., and Han, J. (2021). SubtypeDrug: a software package for prioritization of candidate cancer subtype-specific drugs. *Bioinformatics* doi: 10.1093/bioinformatics/btab011 [Epub Online ahead of print].

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 7:96. doi: 10.1186/1471-2164-7-96

Jeong, H., Hwang, I., Kang, S. H., Shin, H. C., and Kwon, S. Y. (2019). Tumor-Associated Macrophages as Potential Prognostic Biomarkers of Invasive Breast Cancer. *J. Breast Cancer* 22, 38–51. doi: 10.4048/jbc.2019.22.e5

Kimura, M. Y., Hosokawa, H., Yamashita, M., Hasegawa, A., Iwamura, C., Watarai, H., et al. (2005). Regulation of T helper type 2 cell differentiation by murine Schnurri-2. *J. Exp. Med.* 201, 397–408. doi: 10.1084/jem.20040733

Liu, J., Yao, Q., Xiao, L., Li, F., Ma, W., Zhang, Z., et al. (2020). APC/Cdh1 targets PECAM-1 for ubiquitination and degradation in endothelial cells. *J. Cell Physiol.* 235, 2521–2531. doi: 10.1002/jcp.29156

Liu, S., Dhar, P., and Wu, J. D. (2019). NK Cell Plasticity in Cancer. *J. Clin. Med.* 8:1492. doi: 10.3390/jcm8091492

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Nishimura, T., Iwakabe, K., Sekimoto, M., Ohmi, Y., Yahata, T., Nakui, M., et al. (1999). Distinct role of antigen-specific T helper type 1 (Th1) and Th2 cells in tumor eradication in vivo. *J. Exp. Med.* 190, 617–627. doi: 10.1084/jem.190.5.617

Oshi, M., Asaoka, M., Tokumaru, Y., Angarita, F. A., Yan, L., Matsuyama, R., et al. (2020). Abundance of Regulatory T Cell (Treg) as a Predictive Biomarker for Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer. *Cancers* 12:3038. doi: 10.3390/cancers12103038

Pawitan, Y., Bjohle, J., Amler, L., Borg, A. L., Egyhazi, S., Hall, P., et al. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.* 7, R953–64.

Robbins, P. F., Lu, Y. C., El-Gamil, M., Li, Y. F., Gross, C., Gartner, J., et al. (2013). Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* 19, 747–752. doi: 10.1038/nm.3161

Senbabaoglu, Y., Gejman, R. S., Winer, A. G., Liu, M., Van Allen, E. M., De Velasco, G., et al. (2016). Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* 17:231.

Sheng, Y., Jiang, Y., Yang, Y., Li, X., Qiu, J., Wu, J., et al. (2021). CNA2Subpathway: identification of dysregulated subpathway driven by copy number alterations in cancer. *Brief Bioinform* 2021:bbaa413. doi: 10.1093/bib/bbaa413

Tran, E., Turcotte, S., Gros, A., Robbins, P. F., Lu, Y. C., Dudley, M. E., et al. (2014). Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* 344, 641–645. doi: 10.1126/science.1251102

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121. doi: 10.1073/pnas.091062498

Yang, F., Kim, D. K., Nakagawa, H., Hayashi, S., Imoto, S., Stein, L., et al. (2019). Quantifying immune-based counterselection of somatic mutations. *PLoS Genet.* 15:e1008227. doi: 10.1371/journal.pgen.1008227

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Jiang, Zheng, Yang, Li and Han. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/fcell.2021.722410

Identify Inflammatory Bowel Disease-Related Genes Based on Machine Learning

Lili Ye[1†], Yongwei Lin[2†], Xing-di Fan[2†], Yaoming Chen[2], Zengli Deng[2], Qian Yang[2], Xiaotian Lei[2], Jizong Mao[2] and Chunhui Cui[2]*

[1] Daycare Chemotherapy Center, Zhujiang Hospital, Southern Medical University, Guangzhou, China, [2] Department of General Surgery, Zhujiang Hospital, Southern Medical University, Guangzhou, China

The patients of Inflammatory bowel disease (IBD) are increasing worldwide. IBD has the characteristics of recurring and difficult to cure, and it is also one of the high-risk factors for colorectal cancer (CRC). The occurrence of IBD is closely related to genetic factors, which prompted us to identify IBD-related genes. Based on the hypothesis that similar diseases are related to similar genes, we purposed a SVM-based method to identify IBD-related genes by disease similarities and gene interactions. One hundred thirty-five diseases which have similarities with IBD and their related genes were obtained. These genes are considered as the candidates of IBD-related genes. We extracted features of each gene and implemented SVM to identify the probability that it is related to IBD. Ten-cross validation was applied to verify the effectiveness of our method. The AUC is 0.93 and AUPR is 0.97, which are the best among four methods. We prioritized the candidate genes and did case studies on top five genes.

Keywords: index term-inflammatory bowel disease, IBD-related genes, SVM, machine learning, disease ontology

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Shihua Zhang,
Wuhan University of Science and Technology, China
Hong Ju,
Heilongjiang Vocational College of Biology Science and Technology, China

****Correspondence:***
Chunhui Cui
drcuich@163.com

[†]These authors share first authorship

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *08 June 2021*
Accepted: *06 July 2021*
Published: *26 July 2021*

Citation:
Ye L, Lin Y, Fan X-d, Chen Y, Deng Z, Yang Q, Lei X, Mao J and Cui C (2021) Identify Inflammatory Bowel Disease-Related Genes Based on Machine Learning. Front. Cell Dev. Biol. 9:722410. doi: 10.3389/fcell.2021.722410

INTRODUCTION

Inflammatory bowel disease (IBD) (Graham and Xavier, 2020) is a worldwide high incidence of intestinal inflammation, which is divided into Crohn's disease (CD) (Roda et al., 2020) and ulcerative colitis (UC) (Danese et al., 2020). It can lead to diarrhea, rectal bleeding, abdominal pain, and malnutrition. In addition, IBD is a major risk factor for colorectal cancer (CRC) (Cheng et al., 2020; Olén et al., 2020). The incidence rate of colon cancer in IBD patients is 18 times than that of the general population. In view of the increasing incidence rate of IBD worldwide and the huge medical burden, IBD has become a public health problem which needs to be solved urgently. Therefore, it is of great significance to carry out the research on the pathogenesis of IBD (Mayer, 2010), which enables the development of IBD drug targets for the prevention and treatment.

At present, it has been found that the main causes of IBD are the destruction of intestinal microorganism homeostasis, the lack of intestinal epithelial barrier function and the disorder of innate/adaptive immune system, which are caused by the external environmental factors such as smoking, obesity, eating habits (Hovde et al., 2014). Studies also reported that genetic factors such as MUC2 (Van der Sluis et al., 2006) and IL10 (Franke et al., 2008), also play an important role in the development of IBD. In view of the important role of gene mutation in the development of IBD, the genome wide association studies (GWAS) have been applied to the risk prediction and mechanism of IBD (Franke et al., 2008; Zhang et al., 2021). In addition to the discovered NOD2,

IL-23R, ATG16L1, IL-10R, IL-10, XIAP, etc. of IBD-related genes in European and American populations, recent studies have reported more new involvement in cell autophagy, immune regulation, intestinal mucosal barrier, etc. The total number of functional IBD-related sites has reached 200. However, GWAS can only identify the relationship between one single mutation and IBD. The biological functions of mutation and the significance of each gene cannot be explored by GWAS. Although many research have combined expression quantitative trait loci (eQTL) with GWAS to explore the biological functions, this method cannot perform large-scale disease-related gene prediction (Zhao et al., 2019, 2020c).

With the development of computational methods, the applications of algorithms have widely changed the way of doing biological research (Tianyi et al., 2020; Zhao et al., 2020a). Computational methods such as machine learning and deep learning have achieved great achievement in discovering biological mechanism (Zhao et al., 2020b). Predictions by algorithms have shown power in identifying multiple aspects of diseases, such as genes, RNAs, proteins, metabolites, drugs, drug targets, etc. (Zhao et al., 2021). The most common hypothesis of these computational methods is similar diseases are related to similar genes or similar genes cause similar diseases. Research have reported that many disease susceptibility genes do not exist independently, but are associated with a variety of diseases, including IBD. The latest research found that the n2081d allele of LRRK2 gene, which is closely related to the risk of CD, is located in the kinase domain of G2019S gene (Poulopoulos et al., 2012), and G2019S is one of the major mutations involved in familial or sporadic Parkinson's disease. It is also found that three variants of SIAP (s123n, r233q, p257a) of the primary immune deficiency disease type 2 X-linked lymphoproliferative disease related mutation genes (s123n, r233q, p257a) are related to the activation of CDsusceptible gene NOD2 signaling pathway, which is closely related to CD pathogenesis. Therefore, it is theoretically supported to identify IBD related genes from IBD similar diseases.

In this paper, we purposed a Support Vector Machine (SVM)-based on method to identify IBD-related genes. Based on all known genes related to IBD and known related genes of diseases similar to IBD, we constructed a gene network by gene interactions. The features of each gene were extracted from this network and inputted into SVM to identify the pattern that genes are related to IBD.

MATERIALS AND METHODS

Diseases Similar to IBD

Semantic-based methods use the correlation between ontology and the number of disease-related genes to calculate disease similarity. Obviously, not all associations between diseases are represented by ontology, and some of them are reflected by functional associations between disease-related genes. Semfunsim (Cheng et al., 2014) uses disease-related gene sets in a weighted network of human gene functions to calculate disease similarity.

In this paper, we used the results of Semfunsim. One hundred thirty-five diseases have similarities with IBD according to the results of Semfunsim. The distribution of similarities is as **Figure 1**.

The disease most similar to IBD is lower respiratory tract disease with similarity of 0.08. The second one is sensory system disease.

Candidate Genes

We obtained disease-related genes from DisGeNET (Piñero et al., 2016) which collects experimentally verified disease-related genes. Many of the 135 diseases are subtypes of diseases, so their related genes cannot be found in DisGeNET.

We totally obtained 5,928 genes from 88 diseases by DisGeNET. **Figure 2** shows the distribution of these genes.

These 88 diseases have 15,271 entries, which means 15,271 genes are known related to these 88 diseases. However, most of these genes related to more than one disease, so only 5,928 unique genes were obtained. This also reveals that similar diseases share similar genes.

Workflow of SVM-Based Method

We obtained gene interactions from HumanNet v2.0 (Hwang et al., 2019) and constructed a gene interaction network based on candidate gene interactions.

The features of a gene are expressed as the shortest path to other disease-related genes. **Figure 3** shows an example of extracting gene features.

As shown in **Figure 3**, the blue nodes represent the genes related to disease 1, yellow nodes for disease 2 and orange nodes for disease 3. The features of first blue gene could be represented as the shortest paths to disease 2 and 3. The edge of this network is the interaction strength. Therefore, the first dimension of gene 1 is 1 since this gene is related to disease 1 directly. The second

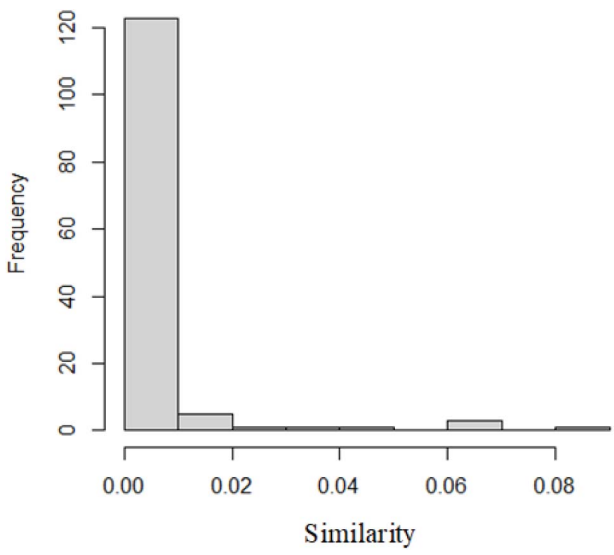


FIGURE 1 | The distribution of similarities.

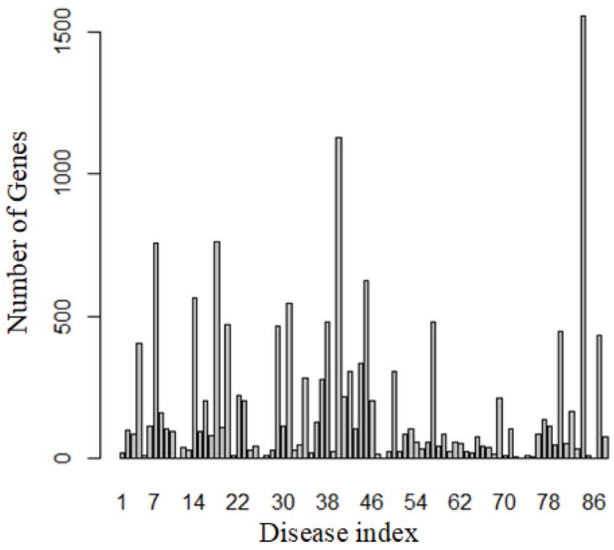


FIGURE 2 | Number of known genes for each disease.

dimension of gene 1 is the interaction strength between gene 1 and the yellow node.

If the network is huge, we cannot extract features of genes manually. Here, we applied Graph search Dijkstra algorithm (Noto and Sato, 2000) to search the shortest path. The algorithm gradually diffuses outwards with the starting point as the center, and the shortest path can be obtained.

Assuming that each node in the network has a label (d_t, p_t),d_t is the shortest path length from the starting point $s$to the point t, and p_t represents the point before the midpoint of the shortest path from$s$to the point t. If we want to find the shortest path, we need to initialize the starting point.

$$\begin{aligned} d_k &= 0 \\ d_i &= \infty \end{aligned} \tag{1}$$

K is the starting point and p_k is NULL.

Then, verify the distance from all marked points k to other directly connected unmarked points j, and set:

$$d_j = \min[d_j, d_k + w(k, j)] \tag{2}$$

$w(k, j)$ represents the length from k to j.

Then, pick the next point. The smallest point i is selected from all unmarked points d_i, and the point i is selected as the point in the shortest path. Find the point directly connected to the point from the set of marked points, and mark it as p_i. Mark the point i. If all the points are marked, the algorithm ends. The flow of this algorithm is as **Table 1**.

We extracted the features of genes in the last step. The features should be inputted into SVM (Friedrichs and Igel, 2004) to predict IBD-related genes. The workflow of SVM is shown as **Figure 4**.

Given a training set, $s = \{(x_1, y_1)(x_2, y_2), \cdots (x_n, y_n)\}$,

$\{x_1 \ldots x_n\}$ represents the gene expression of patients. $\{y_1 \ldots y_n\}$ represents the survival time of patients.

Due to the nonlinear relationship between gene expression and survival time, $\phi(\cdot)$ (kernel function) is needed to map gene expression to high-dimensional feature space.

$$\varphi(x) = (\phi(x_1), \phi(x_2) \cdots \phi(x_n)) \tag{3}$$

The Gaussian Radial Basis Function (RBF) has radial symmetry, which makes it smooth. At the same time, as a kernel function,

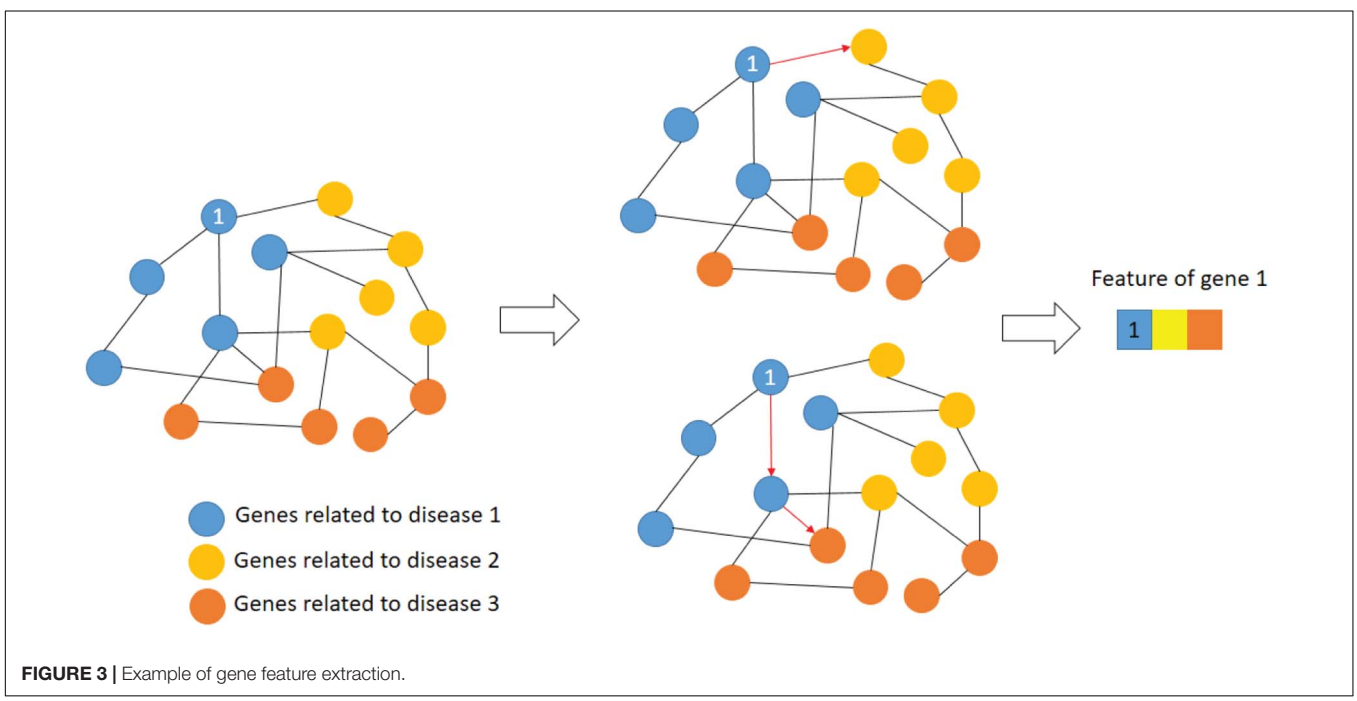


FIGURE 3 | Example of gene feature extraction.

TABLE 1 | Workflow of Dijkstra algorithm.

Input: gene interaction network (V, E, W)	
Output: feature of each gene	
1	**For** i = 1 to N { //N is the number of genes
2	**For** j = i to N {
3	Initialize :$d_i = 0$,other d is infinity;
4	**While** $d_i = d_j$ {
5	**For** m = 1 to M {
6	$d_m = \min[d_m, d_i + W(k, m)]$;
7	}
8	Find the points where d is the smallest, mark it asd_i;
9	}
10	}
11	}
12	Calculate the features of genes in the shortest path
13	**Return** $F' = (F - mean(F))/sd(F)$;

the sample space can be transformed to a high-dimensional space with infinite-dimensional feature space, so that it can better deal with nonlinear relationships. The hyperplane of the feature space can arbitrarily divide the area of the input sample, which can avoid the situation of excessive concentration of training samples.

$$K(x, y) = \exp(-{(x-y)^2}/{2\sigma^2}) \tag{4}$$

σ is kernel width parameter.

Then perform linear regression, the regression function is:

$$f(x) = w^T \cdot \phi(x) + b \tag{5}$$

w^T is weight vector and b is bias. This functional formula could describe relationship between gene expression and survival time. For any given gene features, it can be brought into formula (5) to give a corresponding probability that this is an IBD related gene. Therefore, the next step is to obtain w^T and b.

w^T and b could be obtained based on the principle of structural risk minimization. The structural risk is:

$$R = \gamma \cdot R_{emp} + \frac{1}{2}||w||^2 \tag{6}$$

Where γ is regularization parameter, R_{emp} represents the empirical risk function. In our method, the square of the training error was used as the empirical risk function.

$$R_{emp} = \sum_{i}^{n} \varepsilon_i^2 \tag{7}$$

To minimize R, Lagrange Multiplier Method was introduced.

$$L(w, b, \varepsilon_i, a) = \gamma \cdot \sum_{i}^{n} \varepsilon_i^2 + \frac{1}{2}||w||^2 - \sum_{i=1}^{n}(a_i \cdot (w^T \cdot \phi(x) + b - y_i)) \tag{8}$$

Following the Optimal conditions,

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{n} a_i \cdot \phi(x_i) \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{n} a_i = 0 \\ \dfrac{\partial L}{\partial \varepsilon_i} = 0 \rightarrow a_i = 2\gamma\varepsilon_i \\ \dfrac{\partial L}{\partial a_i} = 0 \rightarrow y_i = w^T \cdot \phi(x_i) + b + \varepsilon_i \end{cases} \tag{9}$$

Therefore, the final SVM model could be represented as:

$$y_i = \sum_{j=1}^{n}(a_j \cdot \langle \phi(x_j)\ ,\ \phi(x_i) \rangle) + b + \frac{1}{2\gamma}a_i \tag{10}$$

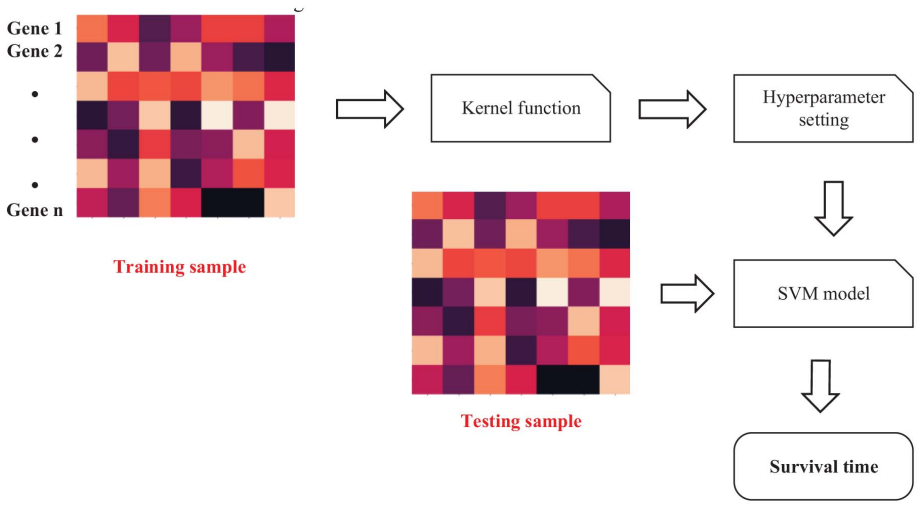


FIGURE 4 | Workflow of Support Vector Machine (SVM)-based prediction method.

RESULTS

Performance of SVM-Based Method in Predicting Survival Time

To verify the effectiveness of our method, 10-fold cross validation was used. We randomly classified 5,928 genes into 10 groups. Then we used 9 of 10 groups to build the SVM model and the rest 1 group was used as the testing set. The process could be repeated 10 times. Therefore, each group could be trained by 9 times and tested by once. We compared our method with several traditional methods, such as Artificial Neural Network (ANN) (Plumb et al., 2005), Random Forest (RF) (Archer and Kimes, 2008), Naïve Bayes (NB) (Archer and Kimes, 2008). The AUC curves were shown as **Figure 5**.

As we can see from **Figure 5**, SVM performed best among these methods. The AUC of SVM is 0.93, which means the IBD-related genes could be precisely predicted.

In addition, we tested the AUPR of SVM. The AUPR is 0.97, which means the false positive rate is low (**Figure 6**).

Identify IBD-Related Genes

Since we have proven the effectiveness of SVM in section "Performance of SVM-Based Method in Predicting Survival Time," we could use this method to identify IBD-related genes in this section.

We totally obtained 1,577 genes related to IBD according to DisGeNET. All these 1,577 genes were used as positive samples to build SVM model. Then, to keep the balance of positive and negative sample sizes, we randomly selected 1,577 genes from 5,928 candidate genes as the negative samples. Finally, the final SVM model could be built. We test whether the rest 4,351 genes are related to IBD by this final SVM model and found that 231 novel genes are related to IBD.

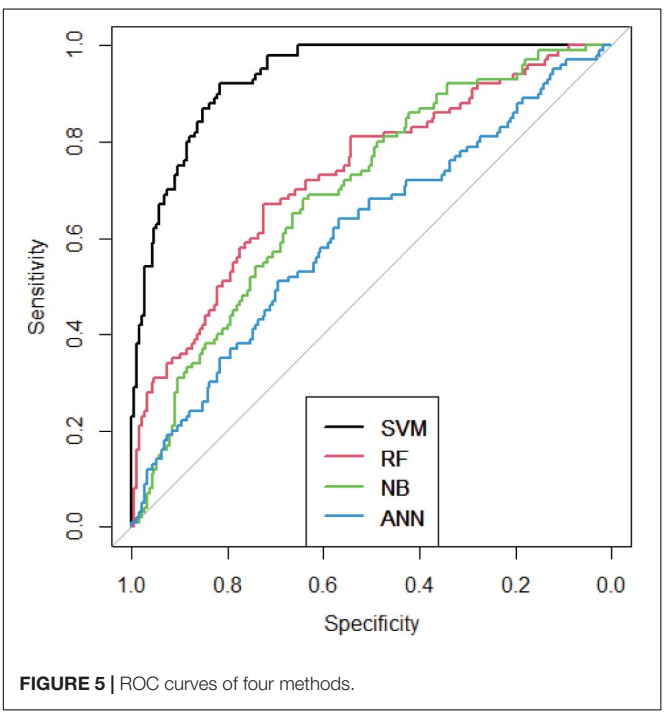


FIGURE 5 | ROC curves of four methods.

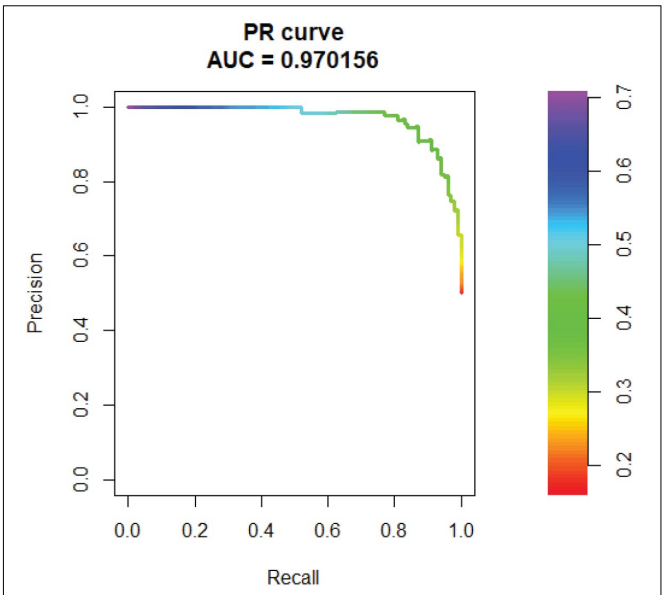


FIGURE 6 | PR curves of Support Vector Machine (SVM).

Case Study

In IBDs, intestinal inflammation has been reported to be accelerated by dysfunction in the epithelial paracellular barrier formed by tight junctions (TJs). Some of the intestinal claudin-family proteins, which form the paracellular barrier, show aberrant expression levels and localizations in IBDs. The intestine-specific Cldn7 deficiency caused colonic inflammation, even though TJ structures were still present due to other claudins according to the study of Tanaka et al. (2015). The paracellular flux (pFlux), determined by measuring the paracellular permeability across the colon epithelium, was enhanced by the Cldn7 deficiency for the small organic solute Lucifer Yellow (457 Da), but not for the larger organic solute FITC-Dextran (4,400 Da). LPA promotes platelet aggregation and induces cellular tension and cell surface fibronectin assembly (Olorundare et al., 2001), which are also important events in wound repair suggesting an important role of LPA in inflammatory disorders. This was confirmed by our group when we demonstrated, that LPA not only promotes epithelial wound healing *in vitro* by a TGF-β-independent pathway, but also ameliorates experimental colitis in an experimental model of colitis in rats (Sturm et al., 2002). LSAMP has been reported to be associated with UC, which is similar to IBD (Brant et al., 2017). A recent study has also found a single-nucleotide polymorphism at rs309 in the *MDM2* gene was associated with UC. They found that people with the GG phenotype of the *MDM2* gene were more prone to UC than those with the TT genotype (Doulabi et al., 2020). Although IBD is not considered as an autoimmune disease, it may trigger autoimmunity caused by the increased antigenic load and mucosal immune activation. Several genetic

alterations are shared between IBD and autoimmune diseases. IBD-related inflammation consists of a variety of abnormalities in humoral and cell-mediated immunity, and a generalized enhanced reactivity against intestinal bacteria. The IFN signature in autoimmune diseases represents a useful tool as a biomarker of disease progression and treatment efficiency. In a recent study, they found that OAS1 is a significant IFN response gene which could serve as an early predictor of disease activity and progression, as well as a supplementary therapeutic target (Andreou et al., 2020).

CONCLUSION

Increasing evidence have shown the relationship between genetics factors and IBD. Identifying IBD-related genes can further reveal the pathogenesis of IBD and provide important support for clinical diagnosis and treatment. A large number of studies have discovered IBD-related genes by biological experiments. However, few genes have been identified.

Since computational methods have shown strong reliability in identifying diseases-related molecular, in this paper, we proposed a SVM-based method for identifying IBD-related genes. A gene network was constructed based on gene interactions and candidate genes. Candidate genes were selected according to the known related genes of diseases similar to IBD. To extract features of genes, we applied Dijkstra algorithm to extract the network topology characteristics as the features of genes. Finally, a SVM-based model was built to identify IBD-related genes.

To verify the effectiveness of our method, we compared SVM with RF, NB, and ANN. SVM showed highest AUC and AUPR among the 10-fold cross validation. After confirming the accuracy of SVM, we built a model using all the data and obtained 231 new genes related to IBD. In order to verify the accuracy of the results, we conducted a case study. Previous studies have shown that five of our newly discovered genes are related to IBD.

Overall, we proposed a novel method of identifying IBD-related genes in large-scale. This method could be also extended into other diseases to help researchers find more diseases-related genes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

CC provided the idea for the study. LY, YL, and X-DF interpreted the results of the data analyses. All authors contributed to designing the algorithm and network construction, interpretation of the results, read, and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 8197032867).

REFERENCES

Andreou, N.-P., Legaki, E., and Gazouli, M. (2020). Inflammatory bowel disease pathobiology: the role of the interferon signature. *Ann. Gastroenterol.* 33:125. doi: 10.20524/aog.2020.0457

Archer, K. J., and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.* 52, 2249–2260. doi: 10.1016/j.csda.2007.08.015

Brant, S. R., Okou, D. T., Simpson, C. L., Cutler, D. J., Haritunians, T., Bradfield, J. P., et al. (2017). Genome-wide association study identifies African-specific susceptibility loci in African Americans with inflammatory bowel disease. *Gastroenterology* 152, 206–217. doi: 10.1053/j.gastro.2016.09.032

Cheng, L., Li, J., Ju, P., Peng, J., and Wang, Y. (2014). SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PloS One* 9:e99415. doi: 10.1371/journal.pone.0099415

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). Gutmdisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48, D554–D560. doi: 10.1093/nar/gkz843

Danese, S., Fiorino, G., and Peyrin-Biroulet, L. (2020). Positioning therapies in ulcerative colitis. *Clin. Gastroenterol. Hepatol.* 18, 1280–1290. doi: 10.1016/j.cgh.2020.01.017

Doulabi, M. S. H., Moghaddam, R. G., and Salehzadeh, A. (2020). Associations between an MDM2 gene polymorphism and ulcerative colitis by ARMS-PCR. *Genomics Inform.* 18:e9. doi: 10.5808/gi.2020.18.1.e9

Franke, A., Balschun, T., Karlsen, T. H., Sventoraityte, J., Nikolaus, S., Mayr, G., et al. (2008). Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet.* 40:1319. doi: 10.1038/ng.221

Friedrichs, F., and Igel, C. (2004). Evolutionary tuning of multiple SVM parameters. *Neurocomputing* 64, 107–117. doi: 10.1016/j.neucom.2004.11.022

Graham, D. B., and Xavier, R. J. (2020). Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* 578, 527–539. doi: 10.1038/s41586-020-2025-2

Hovde, Ø, Kempski-Monstad, I., Småstuen, M. C., Solberg, I. C., Henriksen, M., Jahnsen, J., et al. (2014). Mortality and causes of death in Crohn's disease: results from 20 years of follow-up in the IBSEN study. *Gut* 63, 771–775. doi: 10.1136/gutjnl-2013-304766

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2019). HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* 47, D573–D580. doi: 10.1093/nar/gky1126

Mayer, L. (2010). Evolving paradigms in the pathogenesis of IBD. *J. Gastroenterol.* 45, 9–16. doi: 10.1007/s00535-009-0138-3

Noto, M., and Sato, H. (2000). "A method for the shortest path search by extended Dijkstra algorithm," in *Proceedings of the Smc 2000 conference proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics.'Cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions'(cat. no. 0*, (Nashville, TN: Institute of Electrical and Electronics Engineers), 2316–2320.

Olén, O., Erichsen, R., Sachs, M. C., Pedersen, L., Halfvarson, J., Askling, J., et al. (2020). Colorectal cancer in Crohn's disease: a scandinavian population-based cohort study. *Lancet Gastroenterol. Heptol.* 5, 475–484. doi: 10.1016/S2468-1253(20)30005-4

Olorundare, O. E., Peyruchaud, O., Albrecht, R. M., and Mosher, D. F. (2001). Assembly of a fibronectin matrix by adherent platelets stimulated by lysophosphatidic acid and other agonists. blood. *J. Am. Soc. Hematol.* 98, 117–124. doi: 10.1182/blood.v98.1.117

Piñero, J., Bravo, À, Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: doi:10.1093/nar/gkw943

Plumb, A. P., Rowe, R. C., York, P., and Brown, M. (2005). Optimisation of the predictive ability of artificial neural network (ANN) models: a comparison of three ANN programs and four classes of training algorithm. *Eur. J. Pharm. Sci.* 25, 395–405. doi: 10.1016/j.ejps.2005.04.010

Poulopoulos, M., Cortes, E., Vonsattel, J.-P. G., Fahn, S., Waters, C., Cote, L. J., et al. (2012). Clinical and pathological characteristics of LRRK2 G2019S patients with PD. *J. Mol. Neurosci.* 47, 139–143. doi: 10.1007/s12031-011-9696-y

Roda, G., Ng, S. C., Kotze, P. G., Argollo, M., Panaccione, R., Spinelli, A., et al. (2020). Crohn's disease. *Nat. Rev. Dis. Primers* 6, 1–19. doi: 10.1038/s41572-020-0156-2

Sturm, A., Zeeh, J., Sudermann, T., Rath, H., Gerken, G., and Dignass, A. U. (2002). Lisofylline and lysophospholipids ameliorate experimental colitis in rats. *Digestion* 66, 23–29. doi: 10.1159/000064418

Tanaka, H., Takechi, M., Kiyonari, H., Shioi, G., Tamura, A., and Tsukita, S. (2015). Intestinal deletion of claudin-7 enhances paracellular organic solute flux and initiates colonic inflammation in mice. *Gut* 64, 1529–1538. doi: 10.1136/gutjnl-2014-308419

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22, 2141–2150. doi: 10.1093/bib/bbaa044

Van der Sluis, M., De Koning, B. A., De Bruijn, A. C., Velcich, A., Meijerink, J. P., Van Goudoever, J. B., et al. (2006). Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology* 131, 117–129. doi: 10.1053/j.gastro.2006.04.020

Zhang, S., He, X., Zhang, R., and Deng, W. (2021). LncR2metasta: a manually curated database for experimentally supported lncRNAs during various cancer metastatic events. *Brief. Bioinform.* 22:bbaa178. doi: 10.1093/bib/bbaa178

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 13:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020c). MRTFB regulates the expression of NOMO1 in colon. *Proc. Natl. Acad. Sci. U.S.A.* 117, 7568–7569. doi: 10.1073/pnas.2000499117

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2019). Integrate GWAS, eQTL, and mQTL data to identify alzheimer's disease-related genes. *Front. Genet.* 10:1021. doi: 10.3389/fgene.2019.01021

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Ye, Lin, Fan, Chen, Deng, Yang, Lei, Mao and Cui. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 28 July 2021
doi: 10.3389/fcell.2021.715762

Probing Synergistic Targets by Natural Compounds for Hepatocellular Carcinoma

Jian Gao[1†], Zuojing Yin[1†], Zhuanbin Wu[2†], Zhen Sheng[1], Chao Ma[1], Rui Chen[3], Xiongwen Zhang[3], Kailin Tang[1], Jian Fei[1*] and Zhiwei Cao[1*]*

[1] Department of Gastroenterology, School of Life Sciences and Technology, Shanghai Tenth People's Hospital, Tongji University, Shanghai, China, [2] Shanghai Model Organisms Center, Inc., Shanghai, China, [3] Shanghai Engineering Research Center of Molecular Therapeutics and New Drug Development, School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Xin Chen,
Zhejiang University, China
Xian Zeng,
Fudan University, China

****Correspondence:***
Kailin Tang
kltang@tongji.edu.cn
Jian Fei
jfei@tongji.edu.cn
Zhiwei Cao
zwcao@tongji.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *27 May 2021*
Accepted: *21 June 2021*
Published: *28 July 2021*

Citation:
Gao J, Yin Z, Wu Z, Sheng Z, Ma C, Chen R, Zhang X, Tang K, Fei J and Cao Z (2021) Probing Synergistic Targets by Natural Compounds for Hepatocellular Carcinoma. Front. Cell Dev. Biol. 9:715762. doi: 10.3389/fcell.2021.715762

Background: Designing combination drugs for malignant cancers has been restricted due to the scarcity of synergy-medicated targets, while some natural compounds have demonstrated potential to enhance anticancer effects.

Methods: We here explored the feasibility of probing synergy-mediated targets by Berberine (BER) and Evodiamine (EVO) in hepatocellular carcinoma (HCC). Using the genomics-derived HCC signaling networks of compound treatment, NF-κB and c-JUN were inferred as key responding elements with transcriptional activity coinhibited during the synergistic cytotoxicity induction in BEL-7402 cells. Then, selective coinhibitors of NF-κB and c-JUN were tested demonstrating similar synergistic antiproliferation activity.

Results: Consistent with *in vivo* experiments of zebrafish, coinhibitors were found to significantly reduce tumor growth by 79% and metastasis by 96% compared to blank control, accompanied by anti-angiogenic activity. In an analysis of 365 HCC individuals, the low expression group showed significantly lower malignancies and better prognosis, with the median survival time increased from 67 to 213%, compared to the rest of the groups.

Conclusion: Together, NF-κB and c-JUN were identified as promising synergistic inducers in developing anti-HCC therapies. Also, our method may provide a feasible strategy to explore new targeting space from natural compounds, opening opportunities for the rational design of combinational formulations in combatting malignant cancers.

Keywords: synergistic targets, natural compounds, drug discovery, hepatocellular carcinoma, cancer system biology

INTRODUCTION

Characterized by high heterogeneity, rapid development, and early metastasis, hepatocellular carcinoma (HCC) has been referred to as the leading cause of cancer-related death worldwide with the majority of cases found in the Asia-Pacific region (Yang et al., 2019). Many cellular pathways were reported to be involved in the induction and maintenance of HCC, such as

proliferation, angiogenesis, aberrant metabolism, autocrine growth factors, hypoxia, and so on (Bayat Mokhtari et al., 2017). As of now, only a few inhibitors of sorafenib, lenvatinib (first-line), and regorafenib (second-line) were clinically approved for HCC treatment targeting the VEGFR (vascular endothelial growth factor receptor) (Llovet et al., 2016; Villanueva, 2019). Meanwhile, the PD-1 checkpoint inhibitor nivolumab was reported with conditional approval for sorafenib-treated patients in the United States (Huppert et al., 2020). Being usually diagnosed at an advanced stage, HCC patients clinically receive limited survival benefits from single-drug treatments due to the adaptive therapeutic resistance rapidly developed (Tagliamonte et al., 2020). While conventional chemotherapy was frequently hesitated to recommend, the 5-year survival rate of HCC was reported to be only 18% (Villanueva, 2019).

To seek more effective clinical benefits, combinational agents have been increasingly explored for HCC by simultaneously targeting multiple deregulated targets/pathways, particularly those with potential to produce synergistic therapeutic benefits (Frei et al., 1965). For instance, by blocking both the Ras/Raf/MAPK and PI3K/AKT/mTOR pathways, PI-103, a potent inhibitor of AKT (Ser473) phosphorylation, combined with sorafenib was reported to synergistically inhibit epidermal growth factor (EGF)-stimulated HCC Huh7 cell proliferation (Gedaly et al., 2010). Also, by simultaneously targeting the YAP and EGFR signaling pathways, a strong synergistic effect of simvastatin and gefitinib was detected in BEL-7402 HCC cells (Xia et al., 2018). In this direction, large-scale screening has been initiated by CRISPR-Cas9 technology in searching for additional targets whose knockout can synergize sorafenib efficacy (Wang et al., 2018). Also, computational models have extensively been set up to predict anticancer synergy among candidate drugs with the targeting network and genomic profiling being obtained (Sun et al., 2015).

Notably, an active area of interest is the inclusion of natural therapeutic compounds to sensitize available drugs or formulate new combinations, considering the paucity of FDA-approved agents for HCC. Compound Kushen injection, an extraction from the roots of *Sophora flavescens* and *Smilax glabra* Roxb, was investigated as a sensitizer to sorafenib through relieving immunosuppression in orthotopic, subcutaneous, postsurgical recurrence and tumor rechallenge models (Yang et al., 2020). Moreover, resveratrol, a phytoalexin derived from plants with antioxidant and chemopreventive effects, was reported as a synergistic partner with etoposide in inhibiting proliferation of the HCT-116 and HepG2 cell lines (Amiri et al., 2013). More importantly, a number of natural compounds have been reported with promising anti-HCC effects through *in vitro* or *in vivo* experiments. For instance, Berberine (BER), the main alkaloid of *Coptis chinensis*, was found to inhibit the growth of HepG2 cells by promoting apoptosis through the NF-κB p65 pathway and suppressing glutamine uptake (Li et al., 2017; Zhang et al., 2019). Evodiamine (EVO), an active alkaloid from *Evodia rutaecarpa*, also induced Akt-mediated apoptosis in HepG2 cells (Yang et al., 2017). EVO alone could inhibit the proliferation and promote the apoptosis of HepG2 and BEL-7402 cells *in vitro* and attenuate xenograft tumor formation in nude mice *in vivo* (Zhao et al., 2020). When combined, BER and EVO (CBE) were shown to demonstrate synergistic effects in inducing apoptosis and cell cycle arrest in the HCC cell line of SMMC-7721 (Wang et al., 2008).

Apart from using them for guiding the development of multicomponent therapies, these natural compounds with potential anticancer effects may be explored as valuable probes to detect combinational inducers in targeting space that operate synergistically through multiple signaling pathways. Then further design of selective inhibitors to modulate synergistic targets is therefore becoming an attractive possibility for the treatment of HCC disease. In view of recent synergy predictive efforts (Sun et al., 2015; Yang et al., 2015; Yan et al., 2020), such a task might be accomplished under the analysis of genomic profiling change, the relevant signaling networks, coupled with *in vitro* and *in vivo* validations. Here, this feasibility was evaluated by investigating synergy inducers by the cotreatment of BER and EVO in BEL-7402, a traditional Asian HCC cell line derived from a Chinese male with primary HCC.

MATERIALS AND METHODS

Reagents

BER and EVO (purity: > 98%) were purchased from Shanghai Tauto Biotech Co (Shanghai, China). BER was dissolved in sterile saline, and EVO was dissolved in methyl sulfoxide (DMSO) at 0.01 mol/L to make the stock solution. Other chemicals, except where especially noted, were purchased from Sigma-Aldrich Chemical (St. Louis, MO, United States).

Viability and Synergy Evaluation

Dose-dependent cytotoxicity was measured by the MTT assay as described previously. Briefly, the BEL-7402 cells were seeded in 96-well flat-bottomed plates (Corning, Acton, MA, United States) at a density of 4×10^3 cells/well in a complete medium for 12 h. Then, BER and/or EVO were added alone or in combination at indicated concentrations and left in contact for 48 h. At the end of the incubation, 20 μl MTT (5 mg/ml) was added into each well, and the plates were incubated at 37°C for 3 h. Then 50 μl lysis buffer [20% SDS in 50% N, N-dimethylformamide, containing 0.5% (v/v) 80% acetic acid and 0.4% (v/v) 1 mol/L HCl)] was added into each well and incubated for 12 h. After that, the optical density (proportional to the number of live cells) was assessed with a Genios Microplate Reader (Tecan, Research Triangle Park, NC, United States) at 570 nm.

cDNA Library Preparation and RNA-Seq

For the extraction of total RNA, cells were first rinsed in PBS and then lysed in a Trizol reagent (Invitrogen, United States). Then, mRNA selection, cDNA library preparation, and sequencing were performed by Shanghai Center for Bioinformation Technology (SCBIT) on an Illumina Hiseq 2000 sequencing platform according to the specifications of the manufacturer.

Briefly, mRNA was selected using oligo(dT) probes and then fragmented using divalent cations. cDNA was synthesized using random primers, modified and enriched for attachment to the Illumina flow cell. The raw Illumina sequencing data generated in this study have been deposited as a series with the accession number SRP066753 in the NCBI Sequence Read Archive database (SRA).

Analysis of the RNA-Seq Data

The raw reads were cleaned by removing adaptor sequences, empty reads, and low-quality sequences (reads with unknown sequences "N" or less than 25 bp). Read pairs with at least 36 bp left in each read were kept as clean reads. Then, clean reads were mapped to human reference genome (Ensembl GRCh37.66 = hg19) using aligner Tophat (Trapnell et al., 2012) (version 2.0.6) with parameters (-r/–inner-mate-distance: 70, -g/–max-multihits: 1, –no-coverage-search, –no-novel-junc). Based on the alignment output from TopHat, cuffdiff, an add-on package in Cufflinks, was used to assess the difference in gene expression after those treatments. The differentially expressed genes were screened with the general criteria [an absolute log2 (fold change) ≥ 1].

The Signal Reporter Assay for AP-1/NF-κB

In brief, cells were transfected with the reporter or control constructs by using an Attractene transfection reagent according to the protocol of the manufacturer (Qiagen, Inc., Valencia, CA). After 24-h transfection, cells were exposed to BER (57.36 μM), EVO (4.68 μM), or the combination of BER and EVO (15.78 and 1.58 μM, respectively) for 12 h, and then Luciferase assays were measured using the Dual-Luciferase Reporter assay (Promega, United States) according to the protocol of the manufacturer. Firefly luciferase activity was normalized against Renilla luciferase activity.

Cell Culture

The human hepatocellular carcinoma BEL-7402 cells (Cat. No.: TCHu_10) were obtained from the cell bank of the Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences.[1] Cells were maintained in a humidified 37°C atmosphere containing 5% CO_2 and cultured in an RPMI-1640 medium (GIBCO, Grand Island, NY, United States) supplemented with 10% fetal bovine serum (GIBCO), 2 mmol/L L-glutamine, 50 units/ml penicillin, and 50 mg/ml streptomycin.

Validation of the Synergistic Effects of NF-κB and JNK *in vivo*

In this study, zebrafish embryos containing two strains named of "fli1a-EGFP" and "Casper Casper" were treated with JSH-23 and JNK-IN-8 as well as 0.1% DMSO as vehicle control in 2 days postfertilization (dpf) -5dpf with 20 models in each condition to determine the toxicity and safety. All compounds were dissolved in 0.1% DMSO (in fish water).

[1]http://www.cellbank.org.cn/

After that, DiI-stained human BEL-7402 cells were successfully grafted into the yolk sac of zebrafish embryo at 2 dpf without immunosuppressant treatment. And approximately 200 cells were injected into the yolk sac and assessed by fluorescence microscopy at the 0 days post-injection (dpi). Then, at the 5 dpi, CM-DiI red fluorescence dye-labeled BEL-7402 cancer cells were injected into the embryos.

Proliferation and Survival Analysis of Patients From the TCGA Database

The FPKM values of NF-κB1 (the synonym of NF-κB), MAPK8 (the synonym of JNK), and Ki-67 related with proliferation, as well as the clinical survival status and overall survival (OS), were obtained from the human protein atlas database, which recorded the gene FPKM value and clinical information of patients from the TCGA database.

According to the best cutoff to divide different survival of NF-κB1 (6.18) and MAPK8 (3.06) recorded in this database, the patients were divided into the low or high expression of gene single or combination. Then, the KM survival curve was obtained on two types of samples by using the survival and survminer packages of the R language, and the survival difference was analyzed by the log-rank test. Besides, the boxplot Ki-67 expression among different types of samples was further accomplished by using the ggpubr package of the R language.

Bioinformatics and Statistical Analysis

DAVID (the database for annotation, visualization, and integrated discovery) bioinformatics resources (Huang da et al., 2009) were used to explore the cellular pathways enriched by the differentially expressed genes. Those pathways with a p-value < 0.05 (modified Fisher's exact test) and including at least two differentially expressed genes were considered significant.

Statistical analysis was determined using the two-sided Student's t-test and ANOVA test. The criterion for statistical significance was established as a p-value of < 0.05.

RESULTS

CBE Synergistically Inhibited BEL-7402 Proliferation With Transcriptomics Profiling Changed

Based on previous studies (Yang and Huang, 2013; Zhao et al., 2020), BER and EVO were illustrated as a pair of natural probes to evaluate the feasibility of identifying synergistic targets in BEL-7402. The chemical structures of BER and EVO are shown in **Figure 1A**. The cytotoxicity of BER, EVO, and their combination at a BER/EVO ratio of 10:1 was tested by the MTT method in BEL-7402 (see section "Materials and Methods"). As shown in the dose-response curves in **Figure 1B**, treatment with BER, EVO, or CBE inhibited the growth of BEL-7402 cells in a dose-dependent manner. The derived IC50 values at 48 h of treatment were 57.36 μM for BER and 4.68 μM for

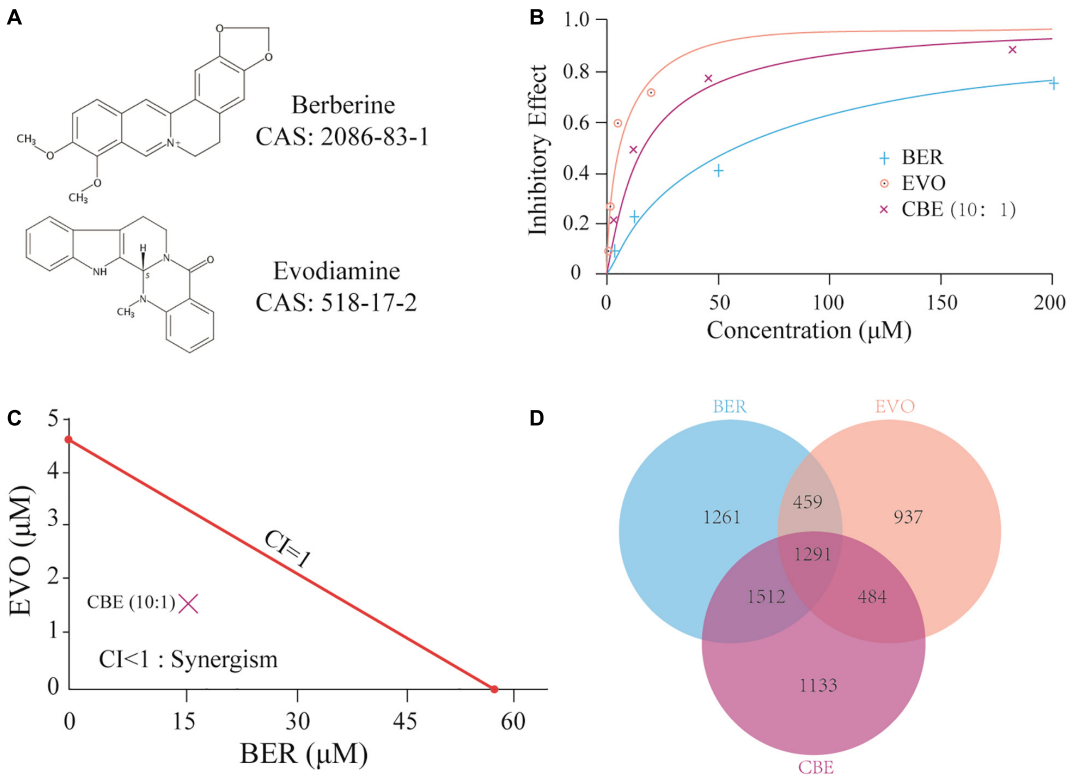


FIGURE 1 | CBE synergisti cally inhibited the growth of BEL-7402 cells. **(A)** The chemical structure of BER and EVO. **(B)** Dose-response curves of BER, EVO, and CBE in BEL-7402 cells. Cells were treated with BER, EVO, or CBE for 48 h. To prepare CBE, a constant ratio (BER/EVO = 10:1) was used, and the curve was plotted against BER concentration. Cell viability was determined by MTT assay. **(C)** Isobologram for CBE at the dose ratio of 10:1. The red line represents the intercept line of the IC50 values for BER or EVO alone, and the "×" indicates the IC50 value for CBE at the dose ratio of 10:1. CI < 1, CI = 1, and CI > 1 indicate synergism, an additive effect, and antagonism, respectively. **(D)** The Venn plot of DEGs under the treatments of BER, EVO, and CBE.

EVO, indicating that EVO might be more cytotoxic to BEL-7402 cells than BER. When BER and EVO were combined, generating CBE, their IC50 values were reduced to 15.78 μM for BER and 1.58 μM for EVO (**Figure 1B**). According to the combination index (CI) determined with the Chou–Talalay method (Ashton, 2015), the synergistic combination effect was attained at a constant molar ratio of BER/EVO = 10:1, which was approximately equal to the ratio of the IC50 values. The drug combination isobologram in **Figure 1C** shows that the CI value at this dose ratio was 0.61, suggesting the synergistic antiproliferative effect of CBE on BEL-7402 cells (CI < 1.0).

To investigate the potential synergy-medicated mechanism of CBE, whole-transcriptome profiling of BEL-7402 cells after 24 h of treatment with BER, EVO, or CBE was performed by RNA-seq to examine rapid changes in gene regulation (see section "Materials and Methods"). BER and EVO at the doses corresponding to their respective IC50 values were used to treat BEL-7402 cells (57.36 μM for BER, 4.68 μM for EVO, and 15.78 and 1.58 μM for BER and EVO, respectively, when combined in CBE). After analyses by TopHat and Cuffdiff (Trapnell et al., 2012), differentially expressed genes (DEGs) following treatment with BER, EVO, or CBE were detected (**Supplementary Tables 1–3**). As shown in **Figure 1D**, 29.21% (1291/4420) of DEGs induced by CBE treatment overlapped with those induced by both BER and EVO treatment. In addition, 25.63% (1133/4420) of the DEGs were uniquely induced by CBE, indicating their potential involvement in drug synergy.

Synergy Responding Elements Suggested by HCC Signaling Transduction Network

Pathway enrichment of the DEGs induced by CBE was obtained by the bioinformatics tool DAVID (**Supplementary Table 4**; Huang da et al., 2009) compared to blank control. When being overlapped with the significant cancer pathways (Sanchez-Vega et al., 2018), three enriched pathways were identified as canonical oncogenic pathways. An extended signal transduction network induced by CBE treatment was then reconstructed based on the cross-talks between canonical carcinogenic pathways (Sanchez-Vega et al., 2018) with DEGs highlighted in different colors compared to blank control (**Figure 2A**). As being suggested, cotreatment might downregulate multiple genes in the same pathway of PI3K, WNT, or MAPK, as well as genes in parallel pathways, such as NF-κB and c-JUN, which directly regulated the

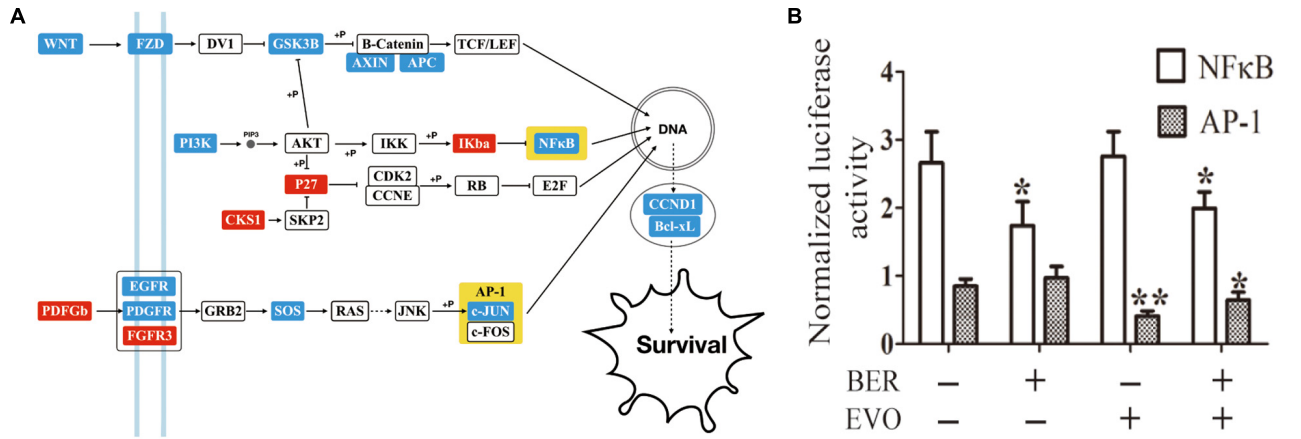


FIGURE 2 | The detection of key players in CBE treatment based on differential signaling network. **(A)** Predicted signaling transduction network inferred for CBE treatment. Red boxes indicate upregulation, and blue boxes indicate downregulation under CBE treatment compared to blank control. Key points are highlighted in yellow. **(B)** Changes in the transcriptional activity of NF-κB and AP-1 by BER, EVO, or CBE treatment. BEL-7402 cells were transfected with the reporter construct for 24 h and then treated with BER, EVO, or CBE for 12 h. For each sample, firefly luciferase activity is presented as normalized to Renilla luciferase activity. The results from each group are presented as the mean ± SD of at least three samples. The drug dose was equal to that used the in the RNA-seq experiment. Columns, mean; bars, SD (*t*-test; $^{*}p < 0.05$; $^{**}p < 0.01$).

downstream biological process and thus inferred as key responding elements.

In addition to mRNA and transcriptional activity, the transcriptional activity of NF-κB1 and c-JUN (reporter gene in AP-1) was also found to be significantly reduced after 6-h incubation upon CBE treatment (**Figure 2B**, see section "Materials and Methods"). All the above results confirmed that CBE simultaneously suppressed the activation of the NF-κB and AP-1 complex. NF-κB and AP-1 are signaling located in the parallel cancer signaling of PI3K and MAPK pathways and are well known with multiple functions of cancer cell proliferation, migration, and invasion (Jiang et al., 2019). Based on previous empirical observation that targeting parallel intracellular signaling or membrane-downstream pathways tend to produce enhanced anticancer effects (Naoum et al., 2018), NF-κB and c-JUN were inferred as key responding elements inducing synergistic anti-HCC effects.

Coinhibitors Synergistically Suppressed HCC Growth and Spreading *in vitro* and *in vivo*

To verify this hypothesis, JSH-23, a highly specific and commonly used NF-κB transcription activity inhibitor (Su et al., 2018), and JNK-IN-8, an irreversible pan-inhibitor of c-JUN N-terminal kinase (JNK) for JNK1, JNK2, and JNK4 (Lalaoui et al., 2016), were used alone or in combination to treat BEL-7402 cells *in vitro* for different durations (**Figure 3** and **Supplementary Figure 1**, see section "Materials and Methods"). After 72 h, cotreatment at different dos ratios indicated that the best ratio of JNK-IN-8/JSH-23 in terms of maximizing their synergistic effect was 1:8 (**Supplementary Figure 1**). At this ratio, strong synergy was observed after 24 h of combination treatment, as illustrated in **Figures 3A,B**. The IC50 values were fitted as 33.36 μM for JNK-IN-8 and 255.95 μM for JSH-23. JSH-23 seemed not to suppress BEL-7402 cell proliferation alone. But when combined with JNK-IN-8, the IC50 values were determined to be 6.22 μM for JNK-IN-8 and 49.77 μM for JSH-23, with a CI index of 0.35. These results support the above hypothesis that coinhibition of NF-κB and c-JUN leads to a synergistic antiproliferative effect in BEL-7402 cells.

To further validate whether the coinhibition of NF-κB and c-JUN could also exert synergistic antitumor effects *in vivo*, we established a tumor metastasis model by using zebrafish embryos xenografted with BEL-7402 HCC cells (**Supplementary Figure 2**). To determine the safe doses of the inhibitors *in vivo*, zebrafish were treated from 2 days post-fertilization (dpf) to 5 dpf, and mortality was recorded every 24 h to derive the maximum non-lethal concentration (MNLC) and LC50 values of lead compounds. The MNLC and LC50 values in zebrafish were 9.3 and 20.1 μM, respectively, for JSH-23, and 24.4 and 56.2 μM, respectively, for JNK-IN-8 (**Supplementary Figure 3**). Detailed information regarding the evaluation of the safety of the combination treatment can be found in **Supplementary Figure 3**. Based on these results, we administered 5 μM JSH-23 and 10 μM JNK-IN-8 alone or in combination with BEL-7402 xenografted embryos at 2 days post-injection (dpi). As shown in **Figures 3C–Q**, inhibition of c-JUN alone had a non-significant anti-HCC effect but enhanced tumor sensitivity to JSH-23 compared to that in the vehicle-treated control, leading to an overall significant synergistic effect through reducing tumor mass by 79% (from 39.29 to 8.13) and disseminated tumor foci by 96% (from 44.90 to 1.80) when combined with JSH-23. In contrast to the poor *in vitro* antiproliferative activity, the NF-κB inhibitor JSH-23 showed striking anti-HCC effects in reducing both tumor mass by 61% (from 39.29 to 15.21) and disseminated tumor foci by 96% (from 44.90 to 2.00) when compared with the vehicle-treated control, which showed nearly no inhibition. In comparison with

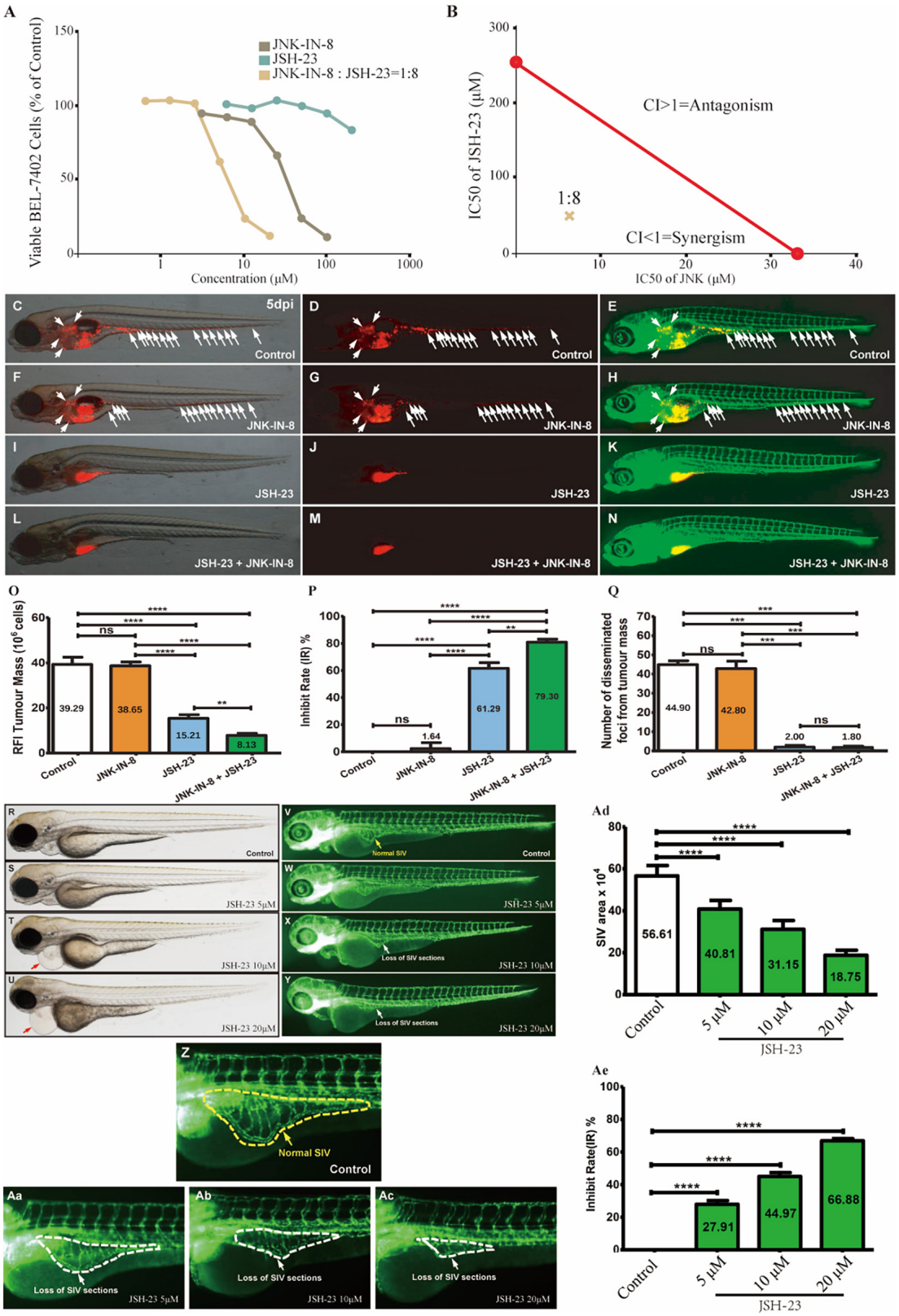


FIGURE 3 | Continued

FIGURE 3 | Coinhibition of NF-κB and c-JUN synergistically suppressed HCC growth and *in vitro* and *in vivo*. **(A)** The dose-response curves for each inhibitor alone or in combination. For the combination, the *x*-axis represents the concentration of JNK-IN-8. **(B)** Isobologram for the combination of JNK-IN-8 and JSH-23 in BEL-7402 cells. The red line represents the intercept line of the IC50 values for JNK-IN-8 and JSH-23 alone, and the "×" indicates the IC50 value for the two inhibitors at a dose ratio of 1:8. CI < 1, = 1, and > 1 indicate synergism, an additive effect, and antagonism, respectively. **(C–E,F–H,I–K,L–N)** Treatments administered to a zebrafish tumor xenograft model were vehicle control, JNK-IN-8 (10 μM) alone, JSH-23 (5 μM) alone, and JNK-IN-8 (10 μM) combined with JSH-23 (5 μM). BEL-7402 cells were labeled with CM-DiI red fluorescence dye, and white arrowheads indicate disseminated tumor foci in 5-dpi zebrafish embryos. **(O–Q)** Quantification of the tumor mass, the inhibition rate, and the number of disseminated tumor foci. **(R–Ac)** Representative bright field and fluorescent images of control embryos or embryos treated with JSH-23 at 4 days post-fertilization (dpf). In control embryos, SIVs developed as a smooth basket-like structure over the yolk at 4 dpf **(V,Z**, yellow dashed lines). In contrast, the treatment of embryos with JSH-23 resulted in specific defects in SIV formation **(W–Y,Aa–Ac**, white dashed lines). Treatment with 10 μM JSH-23 for 48 h caused pericardial edema (red arrow). The complications were caused by vascular defects in the SIV (white arrow). **(Ad,Ae)** Quantification of the SIV area shows a significant decrease in the SIV area in JSH-23-treated embryos compared to control embryos. RFI, relative fluorescence intensity; dpi, days post-injection. Columns, mean; bars, SEM (n = 10; ANOVA; $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$).

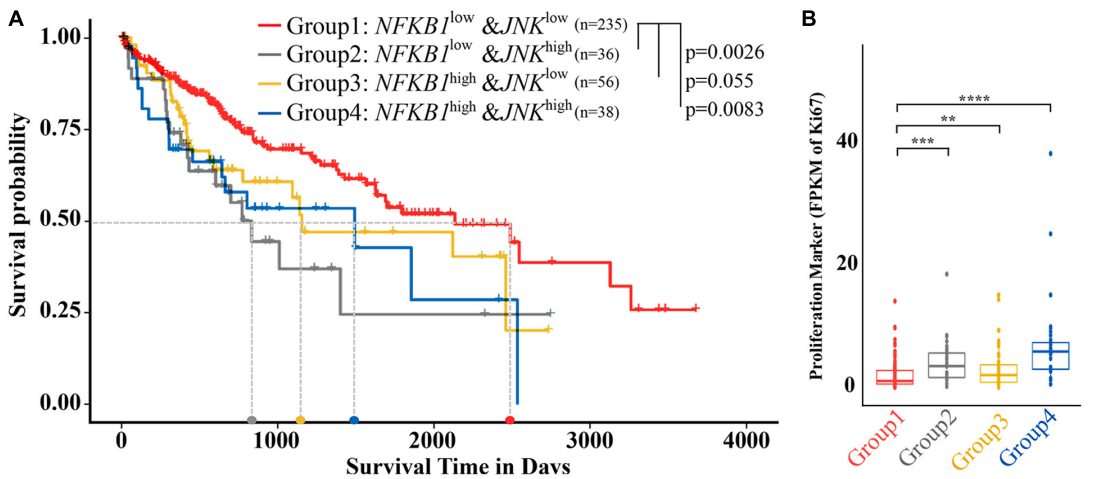


FIGURE 4 | Liver cancer patients with low NF-κB and JNK expression showed a better prognosis and decreased proliferation. **(A)** Survival analysis of samples from patients with low expression of both NF-κB1 and JNK (group 1) and other types of samples. **(B)** Differential expression of Ki-67 in five types of samples. NF-κB^{LOW} and NF-κB^{HIGH} represent samples from patients whose NF-κB expression level is lower or higher, respectively, than the cutoff value of 6.18. JNK^{LOW} and JNK^{HIGH} represent samples from patients whose NF-κB expression level is lower or higher, respectively, than the cutoff value of 3.06. The light gray dotted line represents the median survival time of each group. Columns, median (*t*-test; $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$).

JSH-23 mono-treatment, combination treatment with JSH-23 and JNK-IN-8 reduced the average tumor mass by 46.55% from 15.21 to 8.13, showed an inhibition rate increased by 29.38% from 61.29 to 79.30%, and reduced the average number of disseminated tumor foci by 10.00% from 2 to 1.8. JSH-23 alone almost completely inhibited the dissemination of tumor foci, demonstrating its great antimetastatic potential. In summary, the above results indicate that the coinhibition of NF-κB and c-JUN can produce synergistic anti-HCC effects, with one mechanism (antimetastasis effects) contributing to the inhibition of NF-κ B.

Interestingly, pericardial edema was also observed in zebrafish embryos after JSH-23 treatment (**Supplementary Figure 4**). Since pericardial edema is often caused by vascular defects in the sub-intestinal vein (SIV), the efficacy of JSH-23 was assessed in a zebrafish angiogenesis model by evaluating the SIV. As **Figures 3R–Ae** show, upon treatment with JSH-23 at the selected doses (5, 10, and 20 μM), the inhibition rate of the SIV area was 27.91, 44.97, and 66.88%, respectively, and increased in a dose-dependent manner. JSH-23 impaired the formation of the zebrafish SIV, demonstrating its ability to inhibit angiogenesis *in vivo*, in contrast with vehicle control treatment, which showed nearly no inhibition, or JNK-IN-8 treatment, which had no such effect (data not shown).

Co-downregulation of NF-κB and JNK Correlates With Prolonged Survival and Decreased Malignancy

The above data collected in cell lines and a zebrafish model suggest that the coinhibition of NF-κB and c-JUN results in preferential synergistic anti-HCC effects. Extrapolating these findings to humans, we compared patient survival across 365 individuals with HCC with different expression levels of signature genes based on the recommended parameters from the Protein Atlas database via data from TCGA (Pontén et al., 2008; **Supplementary Table 5**, see section "Materials and Methods"). As shown in **Figure 4A**, low expression of both NF-κB1 and JNK in HCC patients (NF-κB1^{LOW}JNK^{LOW}) was significantly associated with prolonged median survival time (50% survival probability). As determined from the patient data, the median survival time reached approximately 6.85 years (2,500 days) for the NF-κB1^{LOW}JNK^{LOW} patient group and approximately 2 years (800 days), 3.3 years (1,200 days), and 4.1 years

(1,500 days) for the NF-κB1^{LOW}JNK^{HiGH} group, the NF-κB1^{HIGH}JNK^{LOW} group, and the NF-κB1^{HIGH}JNK^{HIGH} group, respectively, which increased the median survival time from 67 to 213%. Noticeably, no patient in the NF-κB1^{HIGH}JNK^{HIGH} group survived longer than 7 years (2,600 days). In contrast, the survival probability of patients in the NF-κB1^{LOW}JNK^{LOW} group who lived longer than 7 years was achieved at 40%. In fact, at the observation endpoint of 11 years (4,000 days), 25% of NF-κB1^{LOW}JNK^{LOW} patients remained alive, demonstrating the potential value of these two targets in combating HCC development.

The Ki-67 protein has been widely used as a proliferation marker for tumor malignancy to assess the growth fraction of a cellular population. A large-scale meta-analysis provided evidence that a high Ki-67 index was closely associated with histological grade, tumor size, the number of tumor nodes, metastasis status, cirrhosis, and vein invasion in HCC patients (Luo et al., 2015). We then investigated the expression of the Ki-67 gene in the different groups (**Figure 4B**). Similar to the results of survival analysis, NF-κB1^{LOW}JNK^{LOW} showed the lowest median Ki-67 expression, indicating the lowest malignancy potential, among the HCC patient groups, while NF-κB1^{HIGH} JNK^{HIGH} had shown the highest Ki-67 expression, suggesting the highest risk of malignancy among the groups. A significant difference in Ki-67 values was observed between the NF-κB 1^{LOW}JNK^{LOW} and other groups. Consequently, compared with the other patients, HCC patients whose NF-κB and JNK are both downregulated are strongly correlated with higher overall survival (OS) and lower malignancies.

So far, the synergy inducers in HCC probed by BER and EVO have been comprehensively supported by selective inhibitors at different levels of cell line, animal model, and clinical patients. Mechanistically, coinhibition of both NF-κB and c-JUN contributes to enhancing not only antiproliferation but also antimetastasis and anti-angiogenesis *in vivo*, which may help to explain the better prognosis in HCC patients.

Strategy of Identifying Synergy-Mediated Targets From Natural Compounds

Identifying appropriate target combinations involving synergistic effects remains challenging in the context of complicated cancer signaling networks, while natural therapeutic compounds with synergistic anti-HCC effects may provide useful clues to pinpoint some candidates. Summarized from above, we here propose a reverse strategy to suggest synergy-responding inducers when potentiating effects are observed by the cotreatment. The main idea is to infer target pairs medicating drug synergy from -omics derived and signaling network, then make subsequent validation via selective inhibitors for suspected targets. The detailed workflow is shown in **Figure 5**: (1) choosing the natural compound pair with synergistic anticancer effects; (2) obtaining differential expressed genes/proteins (DEGs/DEPs) of cotreatment based on -omics data, such as transcriptomics or proteomics profiling; (3) mapping DEGs to cancer signaling transduction networks; (4) inferring key synergy-mediating targets among DEGs based on network topology and empirical experience; (5) deploying selective inhibitors to test the cooperative function between suspected targets through *in vitro* and *in vivo* experiments; (6) conducting survival analysis in clinical samples; and (7) choosing the best candidates.

DISCUSSION

Wide exploration of the targeting space is critically important to design rational combination therapies for aggressive cancers of HCC. Herein, we evaluated the feasibility to probe synergistic targets in cancer signaling networks from natural compounds in the frame of -omics analysis, network pharmacology, and experimental validation via selective inhibitors. Applying this idea, NF-κB and c-JUN were identified by BER and EVO as mediating anti-HCC effects synergistically, which were consistently supported by inhibitor validation *in vitro/vivo*, as well as clinical evidence.

Constitutive NF-κB and JNK activations were found necessitating maintaining the inflammatory and malignant microenvironment of liver cancer (Chang et al., 2009). NF-êB is well known to be widely involved in cell proliferation, apoptosis, invasion, angiogenesis, and metastasis in multiple cancers (Pikarsky et al., 2004). Recently, NF-κB inhibition was found with tumor-suppressive effect in HCC (Tsang et al., 2016) and the NF-κB inhibitor (Bay11-7082) was observed to reduce cell viability, increase ROS formation, release cytochrome c, and induce apoptosis in HepG2 cells (Ahmadian et al., 2018). Meanwhile, c-JUN acts as an oncogene promoting liver tumorigenesis to induce HCC (Trierweiler et al., 2016). Its N-terminal kinase (JNK) is known as an important signaling molecule that can transform external stimuli into a variety of cellular proliferative, invasive, and apoptotic responses, whose upregulation was found to contribute to the development of HCC (Wang and Tai, 2016). While parallel and independent, cross-talks among above pathways could cause interplay between their regulatory mechanisms, controlling numerous cellular events in the process of certain types of cancer (Nakano, 2005; Liu and Lin, 2007). In the complicated HCC networks, the single blocking of one pathway alone may not have sufficient inhibitory effects or even induce inevitable drug resistance (Gedaly et al., 2012; Xia et al., 2018), while the simultaneous targeting of two parallel or cross-talked cancer-related pathways was suggested as a significant mechanism to produce synergistic effects (Jia et al., 2009; Sun et al., 2015). Besides, in invasive HCC tissues, a positive correlation was shown between increased NF-κB, matrix metalloproteinases (MMPs), and metastasis (Zhang et al., 2013). On the other hand, HCC metastasis can be inhibited by norcantharidin through the NF-κB signaling pathway (Yeh et al., 2012), implying the pivotal role of NF-κB in HCC metastasis. In this work, it is also interesting to observe the dose-dependent angiogenesis-suppressive effects of JSH-23 through inhibiting NF-κB activity in the zebrafish model, ideal to test anti-angiogenic compounds (Howe et al., 2013).

Serving as an important source (33%) of cancer drugs developed (Newman and Cragg, 2016), natural therapeutic compounds have been reported by a body of evidence *in vitro*

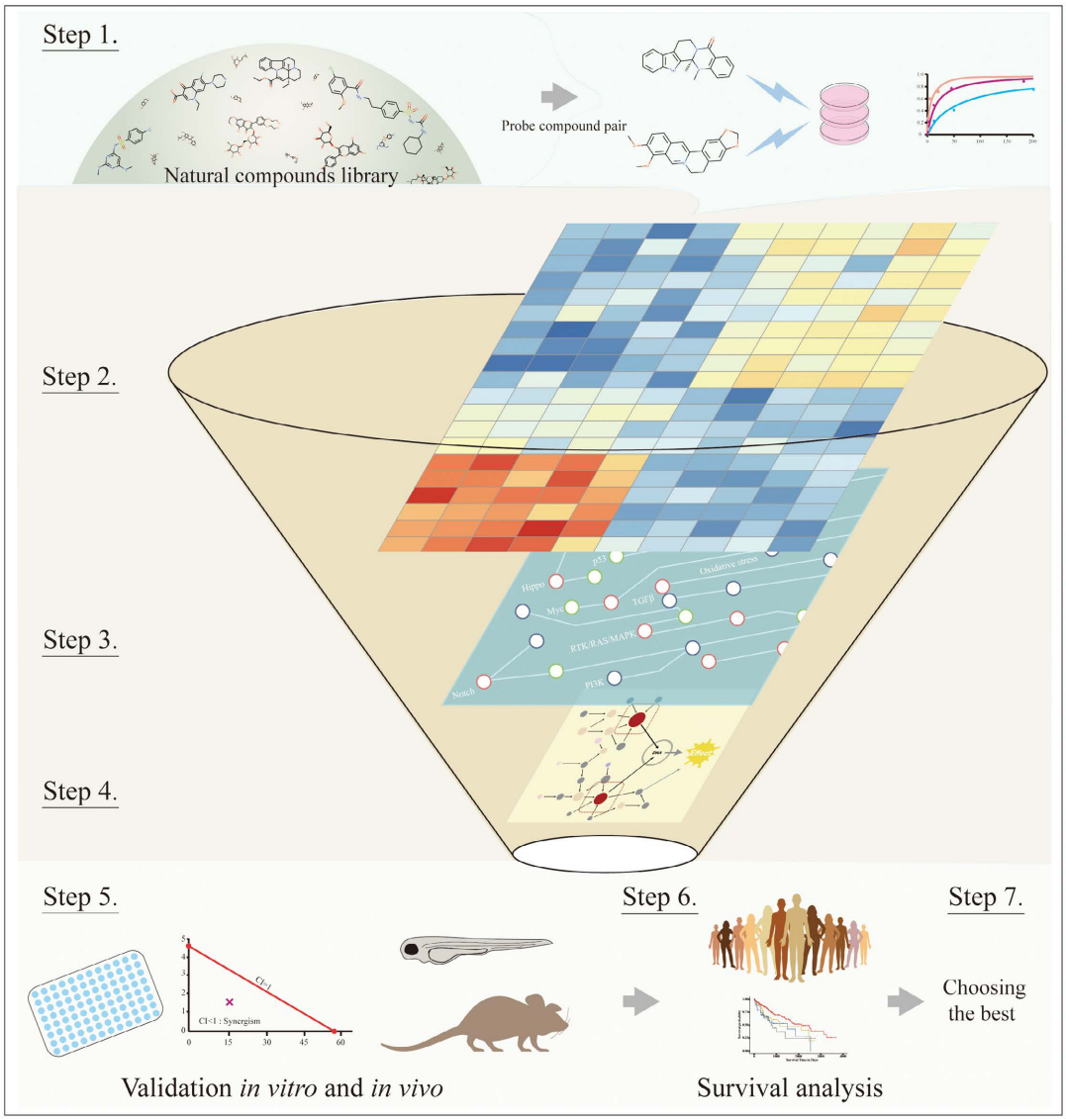


FIGURE 5 | The workflow of systematic processes to identify key synergy-mediated players. Step 1: Choosing the natural compound pair from the natural compounds library. Step 2: Obtaining DEGs from transcriptomics or proteomics profiling. Step 3: Mapping DEGs to cancer signaling transduction networks. Step 4: Inferring key synergy-mediating targets among DEGs based on network topology and empirical experience. Step 5: Validating suspected targets through *in vitro* and *in vivo* experiments. Step 6: Conducting survival analysis in clinical samples. Step 7: Choosing the best candidates.

and *in vivo* on their potential to sensitize cancer drugs (Cote et al., 2015; Cheng et al., 2016). This coupled with their ability to inhibit some targets considered to be conventionally "undruggable," offering a unique advantage in exploring new targeting space for combinational cancer drugs. Previous challenges in efficiently harnessing their information lay in collecting direct targets from the public domain. Strategically, the -omics change to some extent represents a dynamic profiling of compound–target interactions, containing useful information to assess anticancer effects for combinational components as well as other drug activities and responses (Palmer et al., 2019). Advances in cancer system biology enabled the integration of signaling networks, pathway profiles, and -omics change of drug treatment, which can be utilized to model the combinational effects of compound actions (Li et al., 2020) or infer cooperative partners for synergistic actions, as being demonstrated in this work.

In cancer signaling networks, empirical suggestions to infer synergy targets usually cover, but are not limited to, the following points: (1) They are truly DEGs/DEPs in the cotreatment compared to blank controls. (2) They should be consistently regulated in both mono- and cotreatment. (3) They are usually located in key nodes of parallel, or redundant, or cross-talked, or compensatory pathways within certain network topology distances (Jia et al., 2009; Sun et al., 2015; Cheng et al., 2019).

It is cautioned that the synergistic level of different natural combinations may be influenced by genetic variations of the

cancer subtype, the dosage of each ingredient, and even drug scheduling (Bansal et al., 2014). Many synergy-responding elements are much more complicated than those mechanisms inferred so far, and their activities are highly dynamic. Therefore, besides NF-κB and c-JUN signaling, additional combinations might also exist in HCC following the mechanism of updating of drug synergy summarized. So the use of this probing strategy should be more appropriately viewed as a start to a more comprehensive optimized screening of synergy inducers at large-scale and system levels.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

JG and ZY contributed to the overall design. ZW, CM, ZS, and JG performed the wet experiments. RC and XZ contributed to the experiment of metabolism. JG, ZY, CM, and ZS analyzed and interpreted the results. JG, ZY, and ZW wrote the manuscript. XZ, KT, JF, and ZC modified the manuscript. ZC, JF, and KT supervised the project. All authors read, critically reviewed, and approved the final manuscript.

FUNDING

This work was supported in part by the National Key R&D Program of China (2017YFC1700200, 2017YFC0908400, and 2019YFA0905900) and the National Natural Science Foundation of China (32070657 and 81830080).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.715762/full#supplementary-material

REFERENCES

Ahmadian, E., Khosroushahi, A. Y., Eftekhari, A., Farajnia, S., Babaei, H., and Eghbal, M. A. (2018). Novel angiotensin receptor blocker, azilsartan induces oxidative stress and NFkB-mediated apoptosis in hepatocellular carcinoma cell line HepG2. *Biomed. Pharmacother.* 99, 939–946. doi: 10.1016/j.biopha.2018.01.117

Amiri, F., Zarnani, A. H., Zand, H., Koohdani, F., Jeddi-Tehrani, M., and Vafa, M. (2013). Synergistic anti-proliferative effect of resveratrol and etoposide on human hepatocellular and colon cancer cell lines. *Eur. J. Pharmacol.* 718, 34–40. doi: 10.1016/j.ejphar.2013.09.020

Ashton, J. C. (2015). Drug combination studies and their synergy quantification using the Chou-Talalay method–letter. *Cancer Res.* 75:2400. doi: 10.1158/0008-5472.can-14-3763

Bansal, M., Yang, J., Karan, C., Menden, M. P., Costello, J. C., Tang, H., et al. (2014). A community computational challenge to predict the activity of pairs of compounds. *Nat. Biotechnol.* 32, 1213–1222. doi: 10.1038/nbt.3052

Bayat Mokhtari, R., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., et al. (2017). Combination therapy in combating cancer. *Oncotarget* 8, 38022–38043. doi: 10.18632/oncotarget.16723

Chang, Q., Zhang, Y., Beezhold, K. J., Bhatia, D., Zhao, H., Chen, J., et al. (2009). Sustained JNK1 activation is associated with altered histone H3 methylations in human liver cancer. *J. hepatol.* 50, 323–333. doi: 10.1016/j.jhep.2008.07.037

Cheng, F., Kovacs, I. A., and Barabasi, A. L. (2019). Network-based prediction of drug combinations. *Nat. Commun.* 10:1197.

Cheng, Y. T., Yang, C. C., and Shyur, L. F. (2016). Phytomedicine-Modulating oxidative stress and the tumor microenvironment for cancer therapy. *Pharmacol. Res.* 114, 128–143. doi: 10.1016/j.phrs.2016.10.022

Cote, B., Carlson, L. J., Rao, D. A., and Alani, A. W. G. (2015). Combinatorial resveratrol and quercetin polymeric micelles mitigate doxorubicin induced cardiotoxicity in vitro and in vivo. *J. Control Release* 213, 128–133. doi: 10.1016/j.jconrel.2015.06.040

Frei, E. III., Karon, M., Levin, R. H., Freireich, E. J., Taylor, R. J., Hananian, J., et al. (1965). The effectiveness of combinations of antileukemic agents in inducing and maintaining remission in children with acute leukemia. *Blood* 26, 642–656. doi: 10.1182/blood.v26.5.642.642

Gedaly, R., Angulo, P., Hundley, J., Daily, M. F., Chen, C., and Evers, B. M. (2012). PKI-587 and sorafenib targeting PI3K/AKT/mTOR and Ras/Raf/MAPK pathways synergistically inhibit HCC cell proliferation. *J. Surg. Res.* 176, 542–548. doi: 10.1016/j.jss.2011.10.045

Gedaly, R., Angulo, P., Hundley, J., Daily, M. F., Chen, C., Koch, A., et al. (2010). PI-103 and sorafenib inhibit hepatocellular carcinoma cell proliferation by blocking Ras/Raf/MAPK and PI3K/AKT/mTOR pathways. *Anticancer Res.* 30, 4951–4958.

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496, 498–503.

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huppert, L. A., Gordan, J. D., and Kelley, R. K. (2020). Checkpoint Inhibitors for the Treatment of Advanced Hepatocellular Carcinoma. *Clin. Liver Dis.* 15, 53–58. doi: 10.1002/cld.879

Jia, J., Zhu, F., Ma, X., Cao, Z., Cao, Z. W., Li, Y., et al. (2009). Mechanisms of drug combinations: interaction and network perspectives. *Nat. Rev. Drug Discov.* 8, 111–128. doi: 10.1038/nrd2683

Jiang, M., Zhou, L. Y., Xu, N., and An, Q. (2019). Hydroxysafflor yellow A inhibited lipopolysaccharide-induced non-small cell lung cancer cell proliferation, migration, and invasion by suppressing the PI3K/AKT/mTOR and ERK/MAPK signaling pathways. *Thorac. Cancer* 10, 1319–1333. doi: 10.1111/1759-7714.13019

Lalaoui, N., Hanggi, K., Brumatti, G., Chau, D., Nguyen, N. Y., Vasilikos, L., et al. (2016). Targeting p38 or MK2 Enhances the Anti-Leukemic Activity of Smac-Mimetics. *Cancer cell* 29, 145–158. doi: 10.1016/j.ccell.2016.01.006

Li, J., Huo, Y., Wu, X., Liu, E., Zeng, Z., Tian, Z., et al. (2020). Essentiality and transcriptome-enriched pathway scores predict drug-combination synergy. *Biology* 9:278. doi: 10.3390/biology9090278

Li, M., Zhang, M., Zhang, Z. L., Liu, N., Han, X. Y., Liu, Q. C., et al. (2017). Induction of Apoptosis by Berberine in Hepatocellular Carcinoma HepG2 Cells via Downregulation of NF-kappaB. *Oncol. Res.* 25, 233–239. doi: 10.3727/096504016x14742891049073

Liu, J., and Lin, A. (2007). Wiring the cell signaling circuitry by the NF-kappa B and JNK1 crosstalk and its applications in human diseases. *Oncogene* 26, 3267–3278. doi: 10.1038/sj.onc.1210417

Llovet, J. M., Zucman-Rossi, J., Pikarsky, E., Sangro, B., Schwartz, M., Sherman, M., et al. (2016). Hepatocellular carcinoma. *Nat. Rev. Dis. Primers* 2:16018.

Luo, Y., Ren, F., Liu, Y., Shi, Z., Tan, Z., Xiong, H., et al. (2015). Clinicopathological and prognostic significance of high Ki-67 labeling index in hepatocellular carcinoma patients: a meta-analysis. *Int. J. Clin. Exp. Med.* 8, 10235–10247.

Nakano, H. (2005). [Signaling crosstalk between NF-kappaB and JNK]. *Seikagaku* 77, 29–32.

Naoum, G. E., Morkos, M., Kim, B., and Arafat, W. (2018). Novel targeted therapies and immunotherapy for advanced thyroid cancers. *Mol. Cancer* 17:51.

Newman, D. J., and Cragg, G. M. (2016). Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* 79, 629–661. doi: 10.1021/acs.jnatprod.5b01055

Palmer, L. D., Jordan, A. T., Maloney, K. N., Farrow, M. A., Gutierrez, D. B., Gant-Branum, R., et al. (2019). Zinc intoxication induces ferroptosis in A549 human lung cells. *Metallomics* 11, 982–993. doi: 10.1039/c8mt00360b

Pikarsky, E., Porat, R. M., Stein, I., Abramovitch, R., Amit, S., Kasem, S., et al. (2004). NF-kappaB functions as a tumour promoter in inflammation-associated cancer. *Nature* 431, 461–466. doi: 10.1038/nature02924

Pontén, F., Jirström, K., and Uhlen, M. (2008). The Human Protein Atlas—a tool for pathology. *J. Pathol.* 216, 387–393. doi: 10.1002/path.2440

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* 173, 321–337.e10.

Su, S., Chen, J., Yao, H., Liu, J., Yu, S., Lao, L., et al. (2018). CD10(+)GPR77(+) Cancer-Associated Fibroblasts Promote Cancer Formation and Chemoresistance by Sustaining Cancer Stemness. *Cell* 172, 841–856.e16.

Sun, Y., Sheng, Z., Ma, C., Tang, K., Zhu, R., Wu, Z., et al. (2015). Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. *Nat. Commun.* 6:8481.

Tagliamonte, M., Mauriello, A., Cavalluzzo, B., Ragone, C., Manolio, C., Petrizzo, A., et al. (2020). Tackling hepatocellular carcinoma with individual or combinatorial immunotherapy approaches. *Cancer Lett.* 473, 25–32. doi: 10.1016/j.canlet.2019.12.029

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Trierweiler, C., Hockenjos, B., Zatloukal, K., Thimme, R., Blum, H. E., Wagner, E. F., et al. (2016). The transcription factor c-JUN/AP-1 promotes HBV-related liver tumorigenesis in mice. *Cell Death Differ.* 23, 576–582. doi: 10.1038/cdd.2015.121

Tsang, D. P., Wu, W. K., Kang, W., Lee, Y. Y., Wu, F., Yu, Z., et al. (2016). Yin Yang 1-mediated epigenetic silencing of tumour-suppressive microRNAs activates nuclear factor-kappaB in hepatocellular carcinoma. *J. Pathol.* 238, 651–664. doi: 10.1002/path.4688

Villanueva, A. (2019). Hepatocellular Carcinoma. *N. Engl. J. Med.* 380, 1450–1462.

Wang, C., Jin, H., Gao, D., Lieftink, C., Evers, B., Jin, G., et al. (2018). Phospho-ERK is a biomarker of response to a synthetic lethal drug combination of sorafenib and MEK inhibition in liver cancer. *J. Hepatol.* 69, 1057–1065. doi: 10.1016/j.jhep.2018.07.004

Wang, J., and Tai, G. (2016). Role of C-Jun N-terminal Kinase in Hepatocellular Carcinoma Development. *Target. Oncol.* 11, 723–738. doi: 10.1007/s11523-016-0446-5

Wang, X. N., Han, X., Xu, L. N., Yin, L. H., Xu, Y. W., Qi, Y., et al. (2008). Enhancement of apoptosis of human hepatocellular carcinoma SMMC-7721 cells through synergy of berberine and evodiamine. *Phytomedicine* 15, 1062–1068. doi: 10.1016/j.phymed.2008.05.002

Xia, H., Dai, X., Yu, H., Zhou, S., Fan, Z., Wei, G., et al. (2018). EGFR-PI3K-PDK1 pathway regulates YAP signaling in hepatocellular carcinoma: the mechanism and its implications in targeted therapy. *Cell Death Dis.* 9:269.

Yan, X., Yang, Y., Chen, Z., Yin, Z., Deng, Z., Qiu, T., et al. (2020). H-RACS: a handy tool to rank anti-cancer synergistic drugs. *Aging* 12, 21504–21517. doi: 10.18632/aging.103925

Yang, F., Shi, L., Liang, T., Ji, L., Zhang, G., Shen, Y., et al. (2017). Anti-tumor effect of evodiamine by inducing Akt-mediated apoptosis in hepatocellular carcinoma. *Biochem. Biophys. Res. Commun.* 485, 54–61. doi: 10.1016/j.bbrc.2017.02.017

Yang, J., Tang, H., Li, Y., Zhong, R., Wang, T., Wong, S., et al. (2015). DIGRE: drug-induced genomic residual effect model for successful prediction of multidrug effects. *CPT Pharmacometrics Syst. Pharmacol.* 4:e1.

Yang, J. D., Hainaut, P., Gores, G. J., Amadou, A., Plymoth, A., and Roberts, L. R. (2019). A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat. Rev. Gastroenterol. Hepatol.* 16, 589–604. doi: 10.1038/s41575-019-0186-y

Yang, X., and Huang, N. (2013). Berberine induces selective apoptosis through the AMPKmediated mitochondrial/caspase pathway in hepatocellular carcinoma. *Mol. Med. Rep.* 8, 505–510. doi: 10.3892/mmr.2013.1506

Yang, Y., Sun, M., Yao, W., Wang, F., Li, X., Wang, W., et al. (2020). Compound kushen injection relieves tumor-associated macrophage-mediated immunosuppression through TNFR1 and sensitizes hepatocellular carcinoma to sorafenib. *J. Immunother. Cancer* 8:e000317. doi: 10.1136/jitc-2019-000317

Yeh, C. B., Hsieh, M. J., Hsieh, Y. H., Chien, M. H., Chiou, H. L., and Yang, S. F. (2012). Antimetastatic effects of norcantharidin on hepatocellular carcinoma by transcriptional inhibition of MMP-9 through modulation of NF-kB activity. *PLoS One* 7:e31055. doi: 10.1371/journal.pone.0031055

Zhang, J., Zhang, D. L., Jiao, X. L., and Dong, Q. (2013). S100A4 regulates migration and invasion in hepatocellular carcinoma HepG2 cells via NF-kappaB-dependent MMP-9 signal. *Eur. Rev. Med. Pharmacol. Sci.* 17, 2372–2382.

Zhang, P., Wang, Q., Lin, Z., Yang, P., Dou, K., and Zhang, R. (2019). Berberine inhibits growth of liver cancer cells by suppressing glutamine uptake. *Onco Targets Ther.* 12, 11751–11763. doi: 10.2147/ott.s235667

Zhao, S., Xu, K., Jiang, R., Li, D. Y., Guo, X. X., Zhou, P., et al. (2020). Evodiamine inhibits proliferation and promotes apoptosis of hepatocellular carcinoma cells via the Hippo-Yes-Associated Protein signaling pathway. *Life Sci.* 251:117424. doi: 10.1016/j.lfs.2020.117424

Conflict of Interest: ZW was employed by company Shanghai Model Organisms Center, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Gao, Yin, Wu, Sheng, Ma, Chen, Zhang, Tang, Fei and Cao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/fcell.2021.726656

Identification of Claudin-6 as a Molecular Biomarker in Pan-Cancer Through Multiple Omics Integrative Analysis

Chiyuan Zhang[1], Cuishan Guo[1], Yan Li[1], Kuiran Liu[1], Qi Zhao[2] and Ling Ouyang[1]**

[1] Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University, Shenyang, China, [2] School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Fei Guo,
Tianjin University, China
Yang Yang,
Shanghai Jiao Tong University, China

***Correspondence:**
Qi Zhao
zhaoqi@lnu.edu.cn
Ling Ouyang
ouyl@sj-hospital.org

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *17 June 2021*
Accepted: *08 July 2021*
Published: *02 August 2021*

Citation:
Zhang C, Guo C, Li Y, Liu K, Zhao Q and Ouyang L (2021) Identification of Claudin-6 as a Molecular Biomarker in Pan-Cancer Through Multiple Omics Integrative Analysis. Front. Cell Dev. Biol. 9:726656. doi: 10.3389/fcell.2021.726656

Claudin-6 (CLDN6) is one of the 27 family members of claudins and majorly involved in the tight junction and cell-to-cell adhesion of epithelial cell sheets, playing a significant role in cancer initiation and progression. To provide a more systematic and comprehensive dimension of identifying the diverse significance of CLDN6 in a variety of malignant tumors, we explored CLDN6 through multiple omics data integrative analysis, including gene expression level in pan-cancer and comparison of CLDN6 expression in different molecular subtypes and immune subtypes of pan-cancer, targeted protein, biological functions, molecular signatures, diagnostic value, and prognostic value in pan-cancer. Furthermore, we focused on uterine corpus endometrial carcinoma (UCEC) and further investigated CLDN6 from the perspective of the correlations with clinical characteristics, prognosis in different clinical subgroups, co-expression genes, and differentially expressed genes (DEGs), basing on discussing the validation of its established monoclonal antibody by immunohistochemical staining and semi-quantification reported in the previous study. As a result, CLDN6 expression differs significantly not only in most cancers but also in different molecular and immune subtypes of cancers. Besides, high accuracy in predicting cancers and notable correlations with prognosis of certain cancers suggest that CLDN6 might be a potential diagnostic and prognostic biomarker of cancers. Additionally, CLDN6 is identified to be significantly correlated with age, stage, weight, histological type, histologic grade, and menopause status in UCEC. Moreover, CLDN6 high expression can lead to a worse overall survival (OS), disease-specific survival (DSS), and progression-free interval (PFI) in UCEC, especially in different clinical subgroups of UCEC. Taken together, CLDN6 may be a remarkable molecular biomarker for diagnosis and prognosis in pan-cancer and an independent prognostic risk factor of UCEC, presenting to be a promising molecular target for cancer therapy.

Keywords: Cldn-6, pan-cancer analysis, prognostic biomarker, molecular biomarker, omics integrative analysis

INTRODUCTION

Claudins consist of 27 members in mammals that are essential for the regulation of paracellular permeability and the maintenance of cell polarity, due to their significant roles of being key components in tight junctions, involving almost all vital physiological and pathological bioprocesses (Peter and Goodenough, 2004; Lal-Nag and Morin, 2009; Barmeyer et al., 2015; Tabaries and Siegel, 2017). In recent years, claudins have been recognized as crucial regulators in the initiation, progression, and metastasis of cancers, playing distinct roles in a variety of cancers according to their different patterns of tissue-dependent expression (Tabaries and Siegel, 2017).

Claudin-6 (CLDN6) is a member of the claudin family and serves as a tight junction molecule, which plays a vital role in cell-to-cell adhesion in epithelial or endothelial cell sheets. It encodes the tetraspan membrane protein, with the size of 220 amino acids and molecular mass of 23,292 Da. CLDN6 has been identified to be the origination of cell adhesion signaling taking part in the regulation of nuclear receptor activity through targeting molecules of the nuclear receptor superfamily and managing their gene expression (Sugimoto et al., 2019). In terms of its various characters in human cancers, CLDN6 may be a helpful positive marker for further identification of atypical teratoid/rhabdoid tumors (AT/RTs) for diagnosis and therapy (Birks et al., 2010). CLDN6 is also reported to be a possible single prognostic marker and promising therapeutic target for a subgroup of intestinal type gastric cancer (Kohmoto et al., 2020). However, over-expression of CLDN6 may suppress the progression of breast cancer, whereas DNA methylation of CLDN6 can downregulate its gene expression and promote migration and invasion (Liu Y. et al., 2016). Recently, CLDN6 has been identified to be an oncofetal cell surface antigen for chimeric antigen receptor (CAR)-T cell targeting, due to its aberrant activation and high protein expression for solid cancers, while silence for healthy tissues, respectively (Reinhard et al., 2020).

However, as the previous studies reported, CLDN6 may take diverse parts in different cancers, even in different subtypes of the same cancer, either for cancer promotion or for cancer suppression. The discrepancies between the dual roles of CLDN6 in human cancers may be due to the heterogeneity and complexity of tumors. In order to provide a more systematic and comprehensive insight of CLDN6, to the best of our knowledge, we are the first to explore the expression and biofunction of CLDN6 from the perspective of pan-cancer, focusing on its diagnostic and prognostic values, and find that CLDN6 is not only significantly upregulated in 20 types of human cancers but also differently expressed in different molecular subtypes of seven cancer types and different immune subtypes of nine cancer types. Additionally, CLDN6 has a high accuracy in predicting acute myeloid leukemia (LAML), testicular germ cell tumors (TGCT), ovarian serous cystadenocarcinoma (OV), and uterine carcinosarcoma (UCS) while having notable correlations with the overall survival (OS), disease-specific survival (DSS), and progression-free interval (PFI) of uterine corpus endometrial carcinoma (UCEC), adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), and stomach adenocarcinoma (STAD). Next, we put emphasis on UCEC and identify CLDN6 as an independent risk factor for OS, DSS, and PFI in UCEC. Moreover, we further investigate the co-expression genes correlated with CLDN6 and the differentially expressed genes (DEGs) between CLDN6 high expression group and low expression group. Taken together, CLDN6 is a potential biomarker for diagnosis and prognosis in pan-cancer and a promising molecular target for UCEC.

Abbreviations: TCGA, The Cancer Genome Atlas; GTEx, Genotype-Tissue Expression; CCLE, Cancer Cell Line Encyclopedia; DEGs, differentially expressed genes; PPI, protein–protein interaction; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; ROC, receiver operating characteristic; BP, biological process; MF, molecular function; CC, cellular component; AUC, area under the curve; HR, hazard ratio; CI, confidence interval; OS, overall survival; DSS, disease-specific survival; PFI, progression-free interval; TPM, transcripts per million reads; FPKM, fragments per kilobase per million; FC, fold-change; ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

MATERIALS AND METHODS

Gene Expression Analysis

The RNA-seq data and relevant clinical data across 33 tumor types and normal tissues of 15,776 samples were downloaded from The Cancer Genome Atlas (TCGA) database and the Genotype-Tissue Expression (GTEx) database by UCSC XENA[1]. The data of tumor cell line were downloaded from the Cancer Cell Line Encyclopedia (CCLE) database[2]. R software v3.6.3 was used for statistical analysis, and the ggplot2 package was used for visualization. The Wilcoxon rank sum test detected two sets of data, and $p < 0.05$ was considered statistically significant (ns, $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$) (Vivian et al., 2017).

CLDN6 Expression in Molecular Subtypes and Immune Subtypes of Cancers

We explored the correlations between CLDN6 expression and molecular subtypes or immune subtypes in pan-cancer from the TISIDB database (Ru et al., 2019), which integrates multiple data types to assess tumor and immune system interaction. We also explored the correlations between CLDN6 expression and immunomodulators in pan-cancer from the TISIDB database.

[1]https://xenabrowser.net/datapages/

[2]https://portals.broadinstitute.org/ccle/

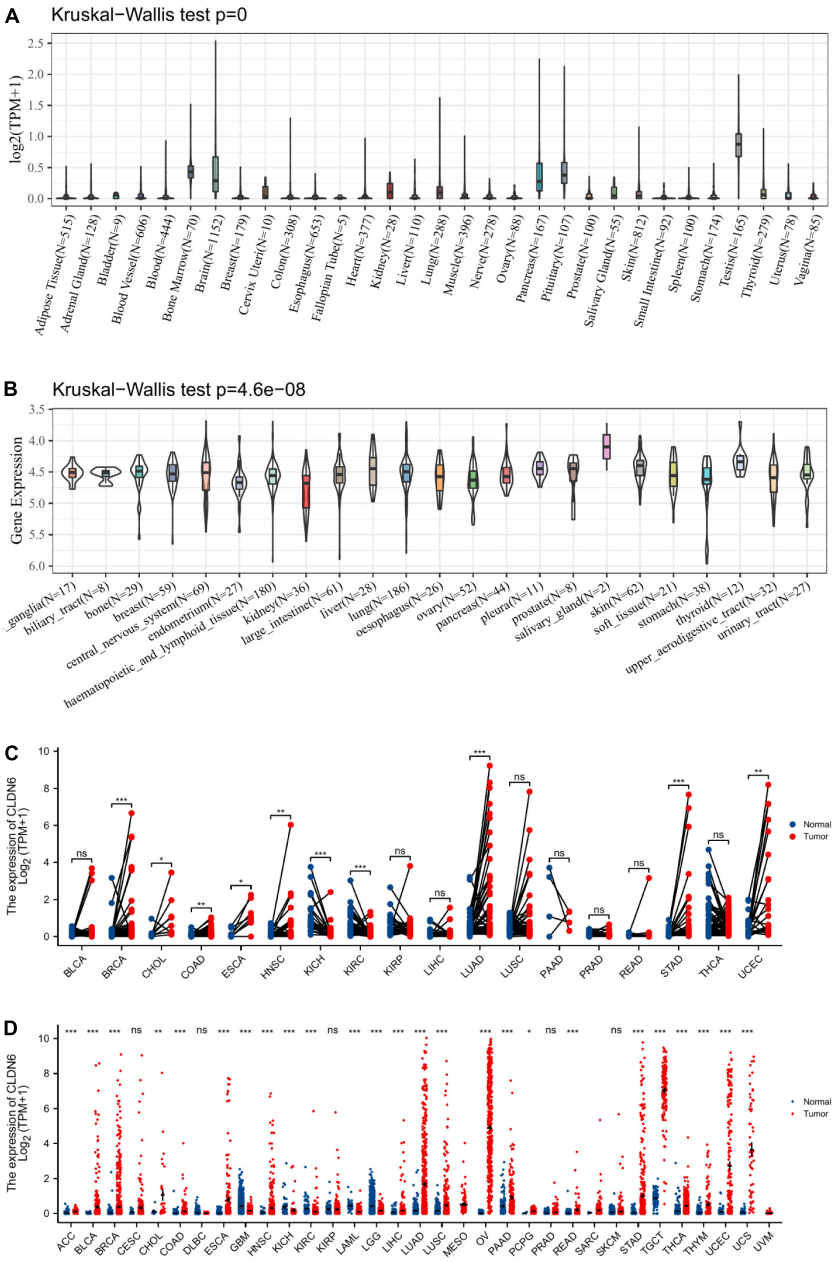


FIGURE 1 | Expression level of CLDN6 gene in tumors and normal tissues. **(A)** CLDN6 expression in normal tissues; **(B)** CLDN6 expression in tumor cell lines; **(C)** CLDN6 expression in TCGA tumors and adjacent normal tissues; **(D)** CLDN6 expression in TCGA tumors and normal tissues with the data of the GTEx database as controls (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

Protein–Protein Interaction Network Building

A total of 50 CLDN6-binding proteins were acquired from the STRING web[3] by setting the following main parameters: minimum required interaction score ["medium confidence (0.400)"] and active interaction sources ("Experiments, Text mining, Databases"). Then, Cytoscape (version 3.7.2) was applied for visualization of protein–protein interaction (PPI) network.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Enrichment Analyses

The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were conducted for 50 CLDN6-binding proteins using the ggplot2

[3]https://string-db.org/

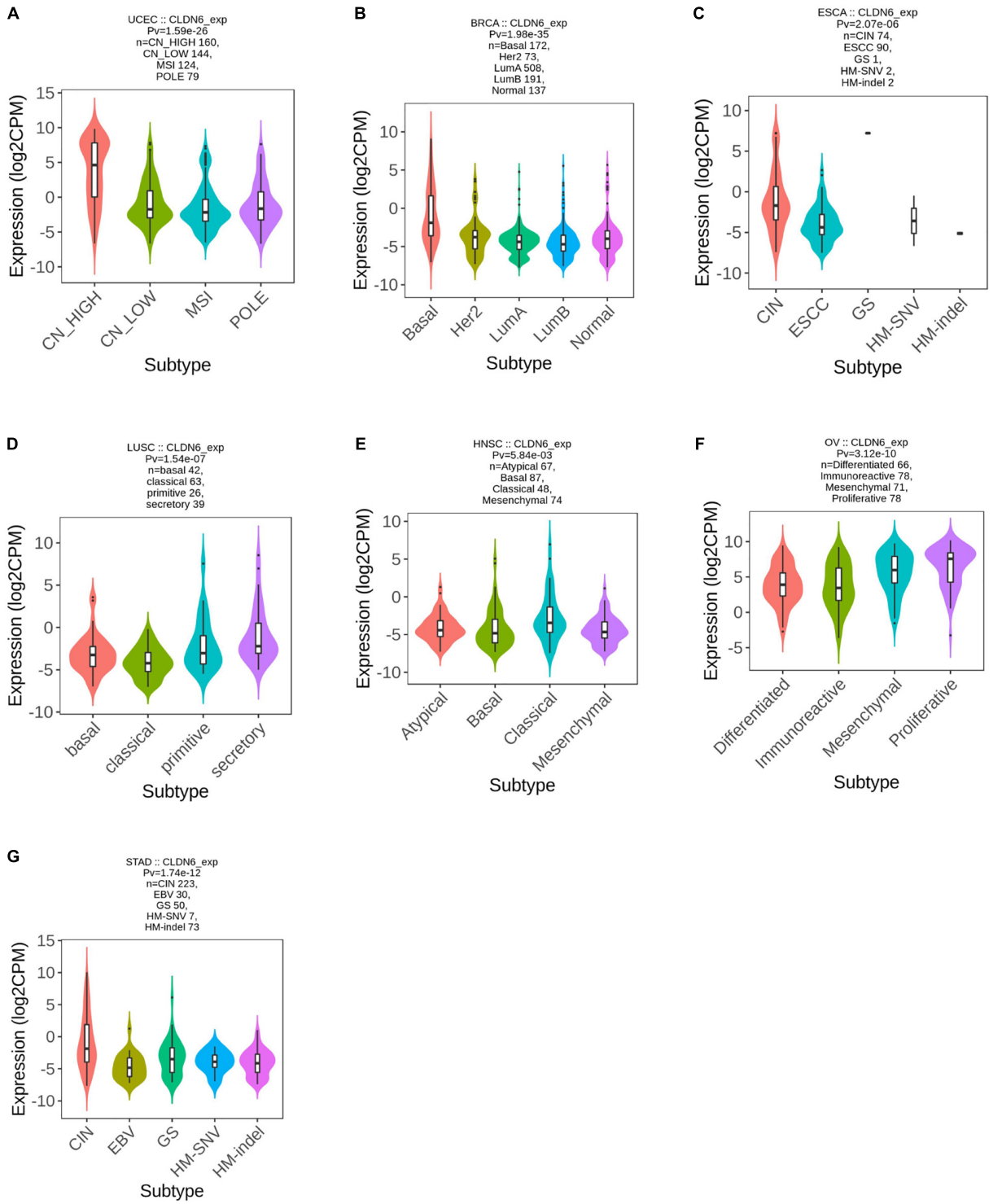


FIGURE 2 | Correlations between CLDN6 expression and molecular subtypes across TCGA tumors. **(A)** UCEC; **(B)** BRCA; **(C)** ESCA; **(D)** LUSC; **(E)** HNSC; **(F)** OV; **(G)** STAD.

package for visualization and the cluster Profiler package for statistical analysis (Subramanian et al., 2005; Yu et al., 2012).

Diagnostic Value Analysis

The receiver operating characteristic (ROC) curve was used to assess the diagnostic value of CLDN6 in pan-cancer. The area

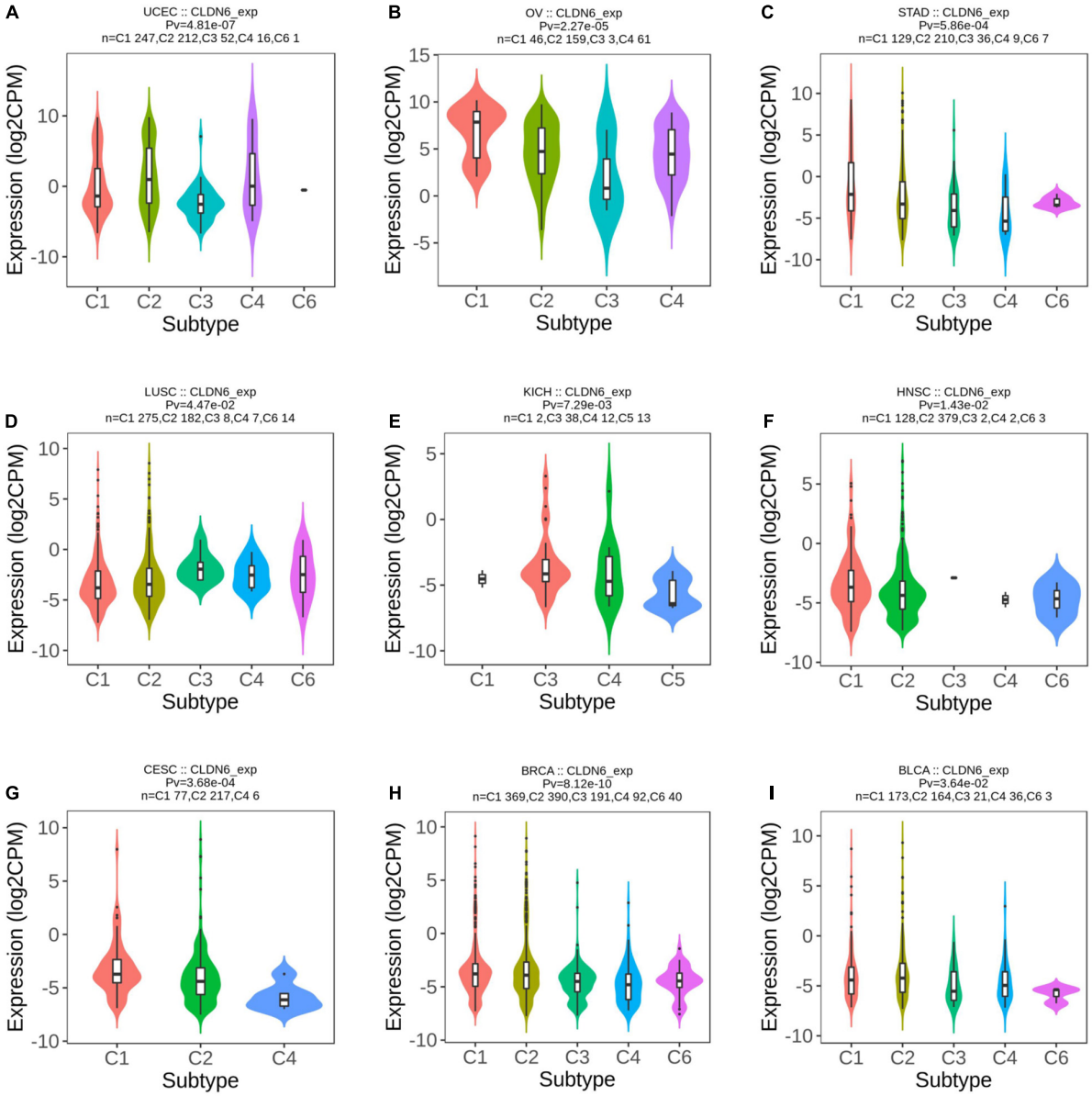


FIGURE 3 | Correlations between CLDN6 expression and immune subtypes across TCGA tumors. **(A)** UCEC; **(B)** OV; **(C)** STAD; **(D)** LUSC; **(E)** KICH; **(F)** HNSC; **(G)** CESC; **(H)** BRCA; **(I)** BLCA.

value under the ROC curve is between 0.5 and 1. The closer the area under the curve (AUC) is (1), the better the diagnostic effect is. AUC in 0.5–0.7 has a low accuracy, AUC in 0.7–0.9 has a certain accuracy, and AUC above 0.9 has a high accuracy.

Survival Prognosis Analysis

Kaplan–Meier plots were used to assess the relationship between CLDN6 expression and prognosis (OS, DSS, and PFI) of cancers. Moreover, we further investigated the associations between CLDN6 expression and prognosis (OS, DSS, and PFI) in different clinical subgroups of UCEC. The survminer package was used for visualization, and the survival package was used for statistical analysis. The Cox regression was used in the hypothesis test, and $p < 0.05$ is considered statistically significant.

Associations Between CLDN6 Expression and Different Clinical Characteristics in UCEC

The box plots and tables were presented for CLDN6 expression levels of patients with different clinical characteristics in UCEC. The RNA-seq data and related clinical data in level 3 HTSeq-fragments per kilobase per million (FPKM) format were downloaded from TCGA database, converted to transcripts per million reads (TPM) format, and then analyzed after log2

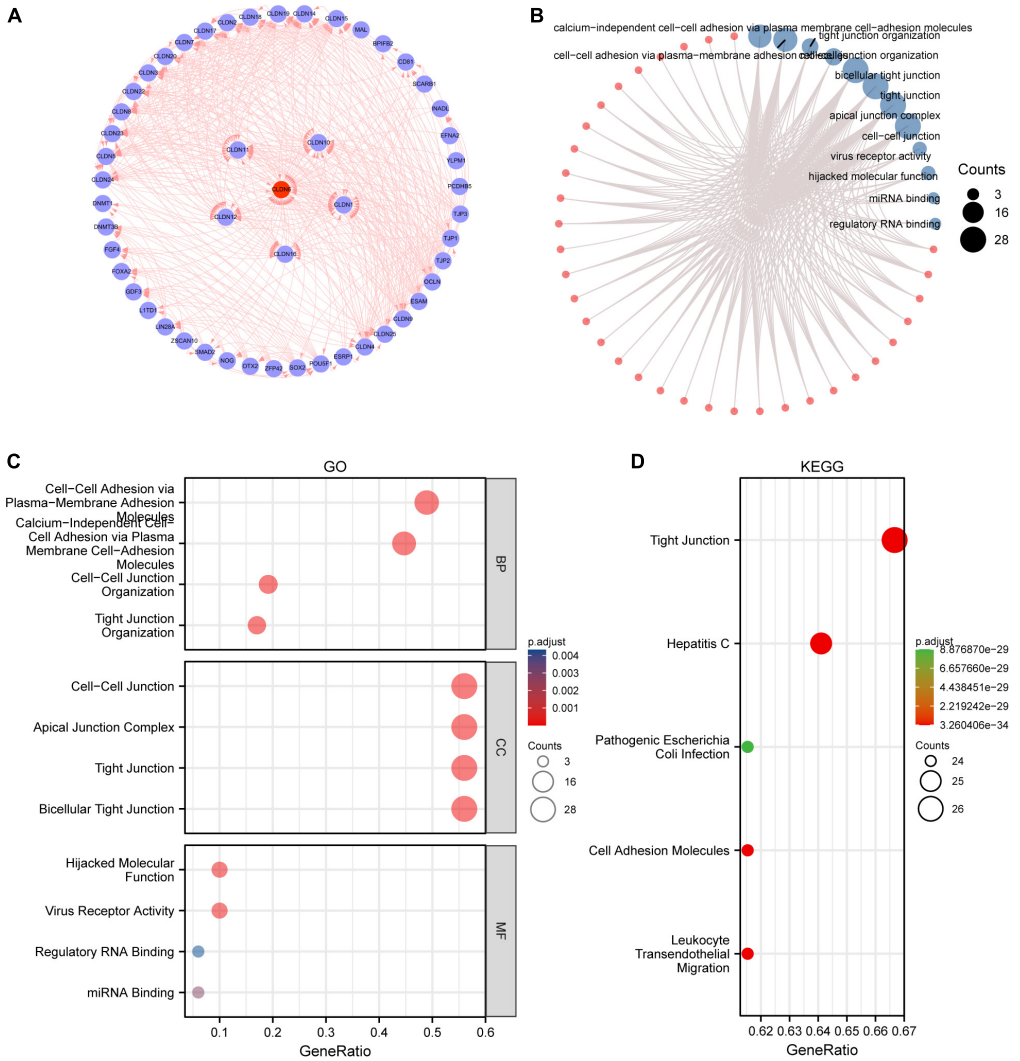


FIGURE 4 | Protein–protein interaction (PPI) network, GO analysis, and KEGG analysis of 50 targeted binding proteins of CLDN6. **(A)** PPI network; **(B)** visual network of GO and KEGG analyses; **(C)** GO analysis; **(D)** KEGG analysis.

conversion. The Wilcoxon rank sum test was used to detect two groups of data, and $p < 0.05$ was considered statistically significant (ns, $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$).

Univariate and Multivariate Cox Regression Analyses in UCEC

Univariate and multivariate Cox regression analyses of CLDN6 and clinical characteristics were performed to identify their prognostic values in OS, DSS, and PFI of UCEC. The survival package was used for statistical analysis.

Co-expression Gene Analysis of CLDN6 in UCEC

We explored the top 50 co-expression genes positively and negatively correlated with CLDN6 expression in UCEC. The gene co-expression heatmaps were displayed using the stat package. We also showed the correlations between CLDN6 expression and top 10 genes expression in the heatmap using Pearson correlation coefficient.

DEGs Between CLDN6 High Expression and Low Expression Groups in UCEC

We explored the DEGs between different CLDN6 expression groups (low expression group: 0–50%; high expression group: 50–100%) in UCEC using the deseq2 package. The volcano map was drawn by the ggplot2 package with the threshold values of $|\log2$ fold-change (FC)$| > 1.0$ and adjusted p-value < 0.05. Then, we performed GO and KEGG enrichment analyses of DEGs using the ggplot2 package for visualization and the cluster Profiler package for statistical analysis. Furthermore, we built a PPI network of DEGs obtained with the threshold values of $|\log2$ fold-change (FC)$| > 2.0$ using STRING web and analyzed the

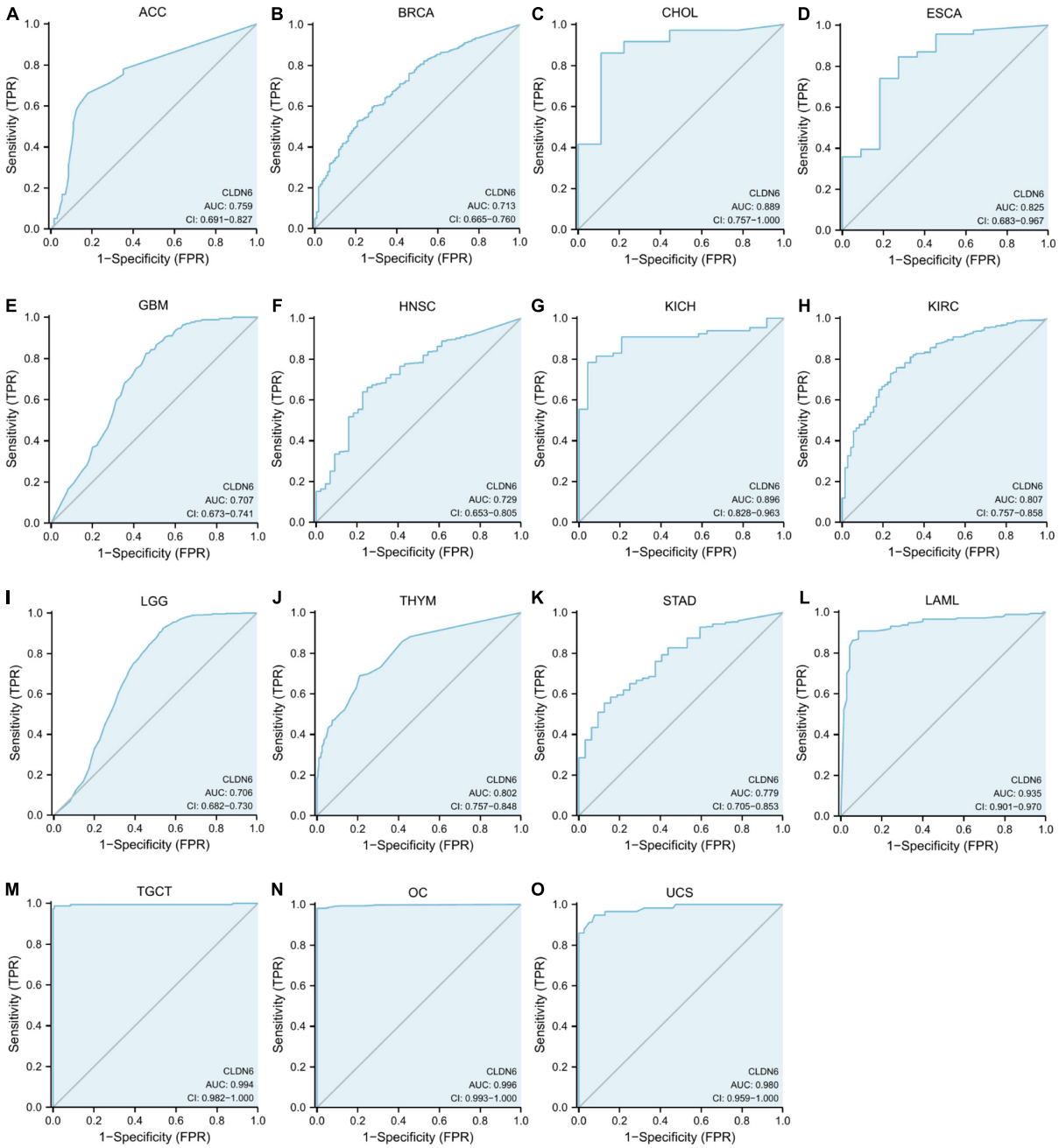


FIGURE 5 | Receiver operating characteristic (ROC) curve for CLDN6 expression in pan-cancer. **(A)** ACC; **(B)** BRCA; **(C)** CHOL; **(D)** ESCA; **(E)** GBM; **(F)** HNSC; **(G)** KICH; **(H)** KIRC; **(I)** LGG; **(J)** THYM; **(K)** STAD; **(L)** LAML; **(M)** TGCT; **(N)** OC; **(O)** UCS.

hub genes by the MCC algorithm of CytoHubba in Cytoscape (version 3.7.2).

RESULTS

CLDN6 Expression in Pan-Cancer

We displayed CLDN6 expression in normal tissues from the GTEx database and found that CLDN6 was less expressed across most normal tissues, and the highest expression tissue was the testis (**Figure 1A**). In contrast, CLDN6 was expressed more in almost all cell lines (**Figure 1B**). For TCGA tumors and adjacent normal tissues, CLDN6 expression was significantly upregulated in eight cancer types, including breast invasive carcinoma (BRCA), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), STAD, and UCEC, while it was downregulated in

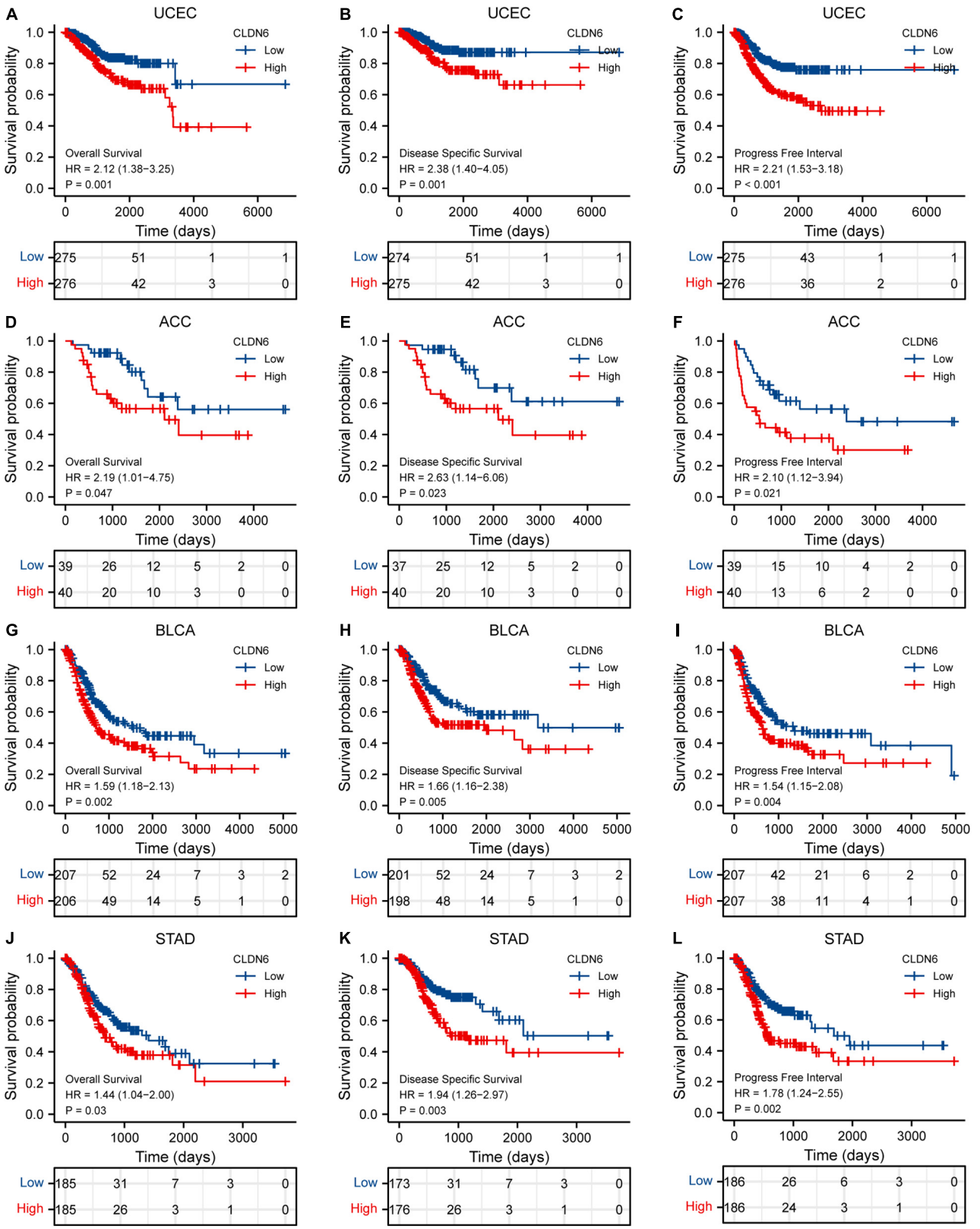


FIGURE 6 | Correlations between CLDN6 expression and the prognosis (OS, DSS, and PFI) of cancers. **(A–C)** UCEC; **(D–F)** ACC; **(G–I)** BLCA; **(J–L)** STAD.

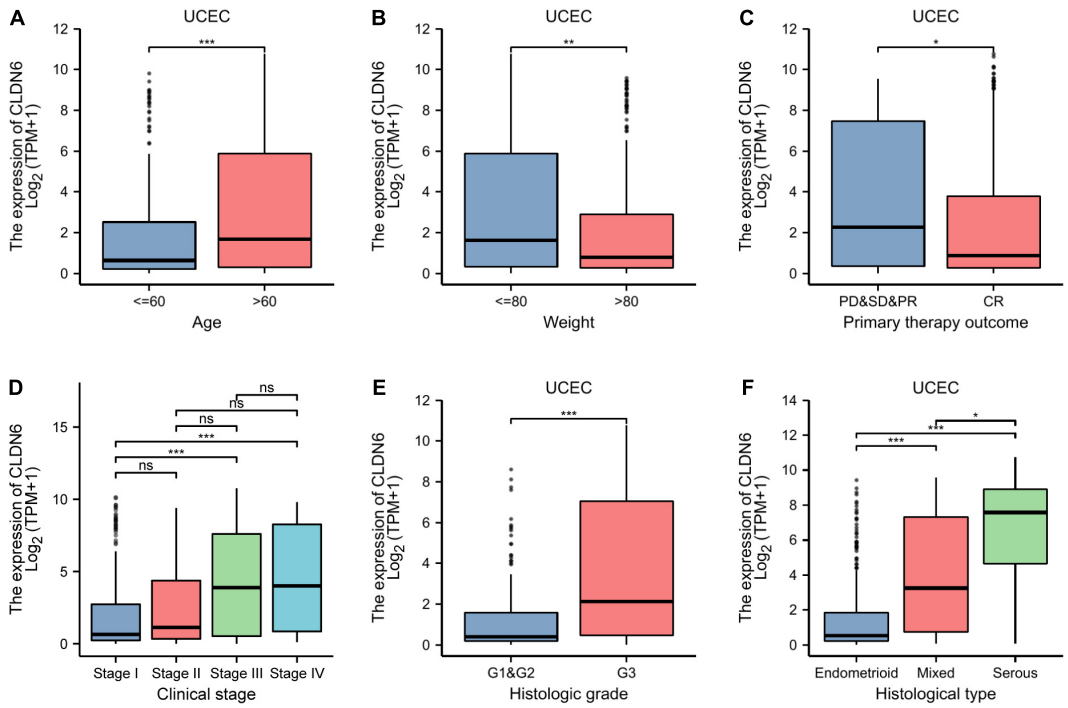


FIGURE 7 | Associations between CLDN6 expression and different clinical characteristics in UCEC. **(A)** Age; **(B)** weight; **(C)** primary therapy outcome; **(D)** clinical stage; **(E)** histologic grade; **(F)** histological type. ns, $p \geq 0.05$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

kidney chromophobe (KICH) and kidney renal clear cell carcinoma (KIRC) (**Figure 1C**). Furthermore, for TCGA tumors with the data of the GTEx database as controls, CLDN6 expression was significantly upregulated in 20 cancer types, including ACC, BLCA, BRCA, CHOL, COAD, ESCA, HNSC, liver hepatocellular carcinoma (LIHC), LUAD, lung squamous cell carcinoma (LUSC), OV, pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), rectum adenocarcinoma (READ), STAD, TGCT, thyroid carcinoma (THCA), thymoma (THYM), UCEC, and UCS, while it was downregulated in glioblastoma multiforme (GBM), KICH, KIRC, LAML, and brain lower grade glioma (LGG) (**Figure 1D**).

Correlations Between CLDN6 and Molecular or Immune Subtypes of Cancers

We explored the correlations between CLDN6 differential expression and molecular subtypes in pan-cancer from the TISIDB database and found that CLDN6 was expressed differently in different molecular subtypes of seven cancer types, including UCEC, BRCA, ESCA, LUSC, HNSC, OV, and STAD. Moreover, for UCEC, CLDN6 was identified to express more in the molecular subtype of CN_HIGH than other molecular subtypes (**Figure 2A**). For BRCA, CLDN6 was expressed the highest in the molecular subtype of basal (**Figure 2B**). For ESCA, CLDN6 was expressed the highest in the molecular subtype of CIN (**Figure 2C**). For LUSC, CLDN6 was expressed the highest in the molecular subtype of secretory (**Figure 2D**). For HNSC, CLDN6 was expressed the highest in the molecular subtype of classical (**Figure 2E**). For OV, CLDN6 was expressed the highest in the molecular subtype of proliferative (**Figure 2F**). For STAD, CLDN6 was expressed the highest in the molecular subtype of CIN (**Figure 2G**).

Meanwhile, we observed that CLDN6 expression was significantly correlated with different immune subtypes (C1: wound healing, C2: IFN-gamma dominant, C3: inflammatory, C4: lymphocyte depleted, C5: immunologically quiet, C6: TGF-b dominant) of nine cancer types, including UCEC (**Figure 3A**), OV (**Figure 3B**), STAD (**Figure 3C**), LUSC (**Figure 3D**), KICH (**Figure 3E**), HNSC (**Figure 3F**), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) (**Figure 3G**), BRCA (**Figure 3H**), and BLCA (**Figure 3I**). Also, we observed that CLDN6 expression was associated with immune stimulators (**Supplementary Figure 1**) and immune inhibitors (**Supplementary Figure 2**) in the majority of malignant tumors except for KIRC, kidney renal papillary cell carcinoma (KIRP), LGG, LIHC, PCPG, prostate adenocarcinoma (PRAD), READ, sarcoma (SARC), and skin cutaneous melanoma (SKCM).

PPI Network and GO and KEGG Enrichment Analyses

We screened out 50 targeted binding proteins of CLDN6 using the STRING database and Cytoscape (**Figure 4A**). Then, we conducted the GO enrichment analysis (**Figure 4B**) of 50

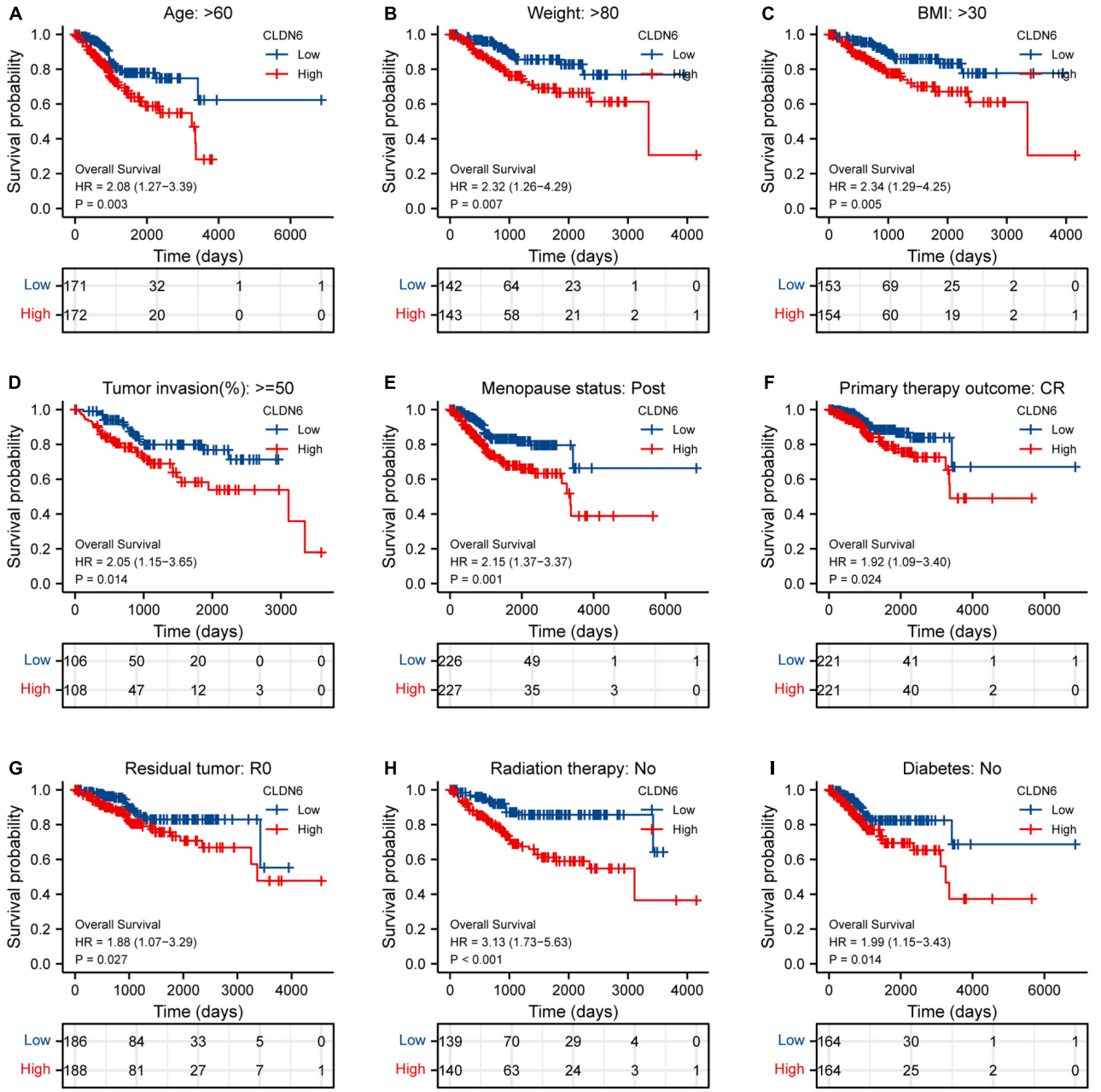


FIGURE 8 | Associations between CLDN6 expression and the OS in different clinical subgroups of UCEC. **(A)** Age > 60; **(B)** weight > 80; **(C)** BMI > 30; **(D)** tumor invasion ≥ 50%; **(E)** postmenopause; **(F)** primary therapy outcome (CR); **(G)** residual tumor (R0); **(H)** radiation therapy (No); **(I)** diabetes (No).

targeted binding proteins, revealing that the primary biological process (BP) contained calcium-independent cell–cell adhesion *via* plasma membrane cell adhesion molecules, cell–cell adhesion *via* plasma membrane adhesion molecules, tight junction organization, and cell–cell junction organization. The cellular component (CC) was mainly enriched in bicellular tight junction, apical junction complex, tight junction, and cell–cell junction. The molecular function (MF) was primarily involved in virus receptor activity, hijacked MF, miRNA binding, and regulatory RNA binding. The KEGG pathway enrichment (**Figures 4C,D**) was mainly related to leukocyte transendothelial migration, tight junction, cell adhesion molecules, hepatitis C, and pathogenic *Escherichia coli* infection.

Diagnostic Value of CLDN6 in Pan-Cancer

The ROC curve was performed to assess the diagnostic value of CLDN6 in pan-cancer. The results showed that CLDN6 had a certain accuracy (AUC > 0.7) in predicting

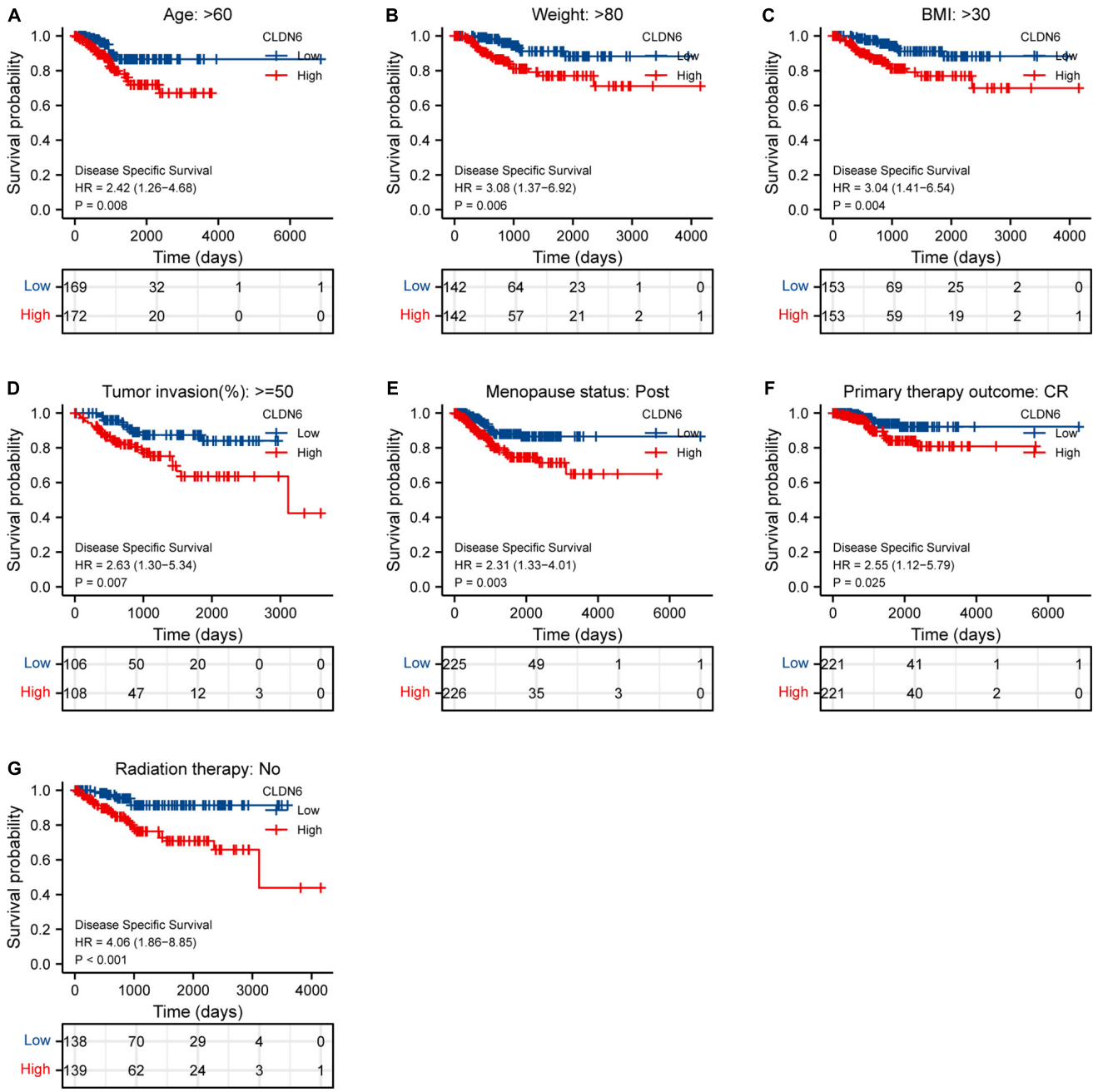


FIGURE 9 | Associations between CLDN6 expression and the DSS in different clinical subgroups of UCEC. **(A)** Age > 60; **(B)** weight > 80; **(C)** BMI > 30; **(D)** tumor invasion ≥ 50%; **(E)** postmenopause; **(F)** primary therapy outcome (CR); **(G)** radiation therapy (No).

15 cancer types, including ACC (AUC = 0.759) (**Figure 5A**), BRCA (AUC = 0.713) (**Figure 5B**), CHOL (AUC = 0.889) (**Figure 5C**), ESCA (AUC = 0.825) (**Figure 5D**), GBM (AUC = 0.707) (**Figure 5E**), HNSC (AUC = 0.729) (**Figure 5F**), KICH (AUC = 0.896) (**Figure 5G**), KIRC (AUC = 0.807) (**Figure 5H**), LGG (AUC = 0.706) (**Figure 5I**), THYM (AUC = 0.802) (**Figure 5J**), STAD (AUC = 0.779) (**Figure 5K**), LAML (AUC = 0.935) (**Figure 5L**), TGCT (AUC = 0.994) (**Figure 5M**), OC (AUC = 0.996) (**Figure 5N**), and UCS (AUC = 0.980) (**Figure 5O**). Among them, CLDN6 had a high accuracy (AUC > 0.9) in predicting LAML, TGCT, OC, and UCS.

Prognostic Value of CLDN6 in Cancers

The expression level of CLDN6 was notably correlated with the OS, DSS, and PFI of UCEC, ACC, BLCA, and STAD. For UCEC, Cox regression results showed that the higher expression of CLDN6 had a worse prognosis, including OS

TABLE 1 | Clinical characteristics of UCEC patients.

Characteristic	Low expression of CLDN6	High expression of CLDN6	*p*
n	**276**	**276**	
Clinical stage, n (%)			<0.001
Stage I	196 (35.5%)	146 (26.4%)	
Stage II	25 (4.5%)	26 (4.7%)	
Stage III	47 (8.5%)	83 (15%)	
Stage IV	8 (1.4%)	21 (3.8%)	
Age, *n* (%)			<0.001
≤60	124 (22.6%)	82 (14.9%)	
>60	150 (27.3%)	193 (35.2%)	
Weight, n (%)			0.036
≤80	109 (20.6%)	134 (25.4%)	
>80	155 (29.4%)	130 (24.6%)	
Histological type, *n* (%)			<0.001
Endometrioid	257 (46.6%)	153 (27.7%)	
Mixed	8 (1.4%)	16 (2.9%)	
Serous	11 (2%)	107 (19.4%)	
Histologic grade, *n* (%)			<0.001
G1	68 (12.6%)	30 (5.5%)	
G2	83 (15.3%)	37 (6.8%)	
G3	122 (22.6%)	201 (37.2%)	
Menopause status, *n* (%)			0.005
Pre	17 (3.4%)	18 (3.6%)	
Peri	15 (3%)	2 (0.4%)	
Post	217 (42.9%)	237 (46.8%)	

[hazard ratio (HR) = 2.12, 95% confidence interval (CI): 1.38–3.25, $p = 0.001$] (**Figure 6A**), DSS (HR = 2.38, 95% CI: 1.40–4.05, $p = 0.001$) (**Figure 6B**), and PFI (HR = 2.21, 95% CI: 1.53–3.18, $p < 0.001$) (**Figure 6C**). Also, for ACC, Cox regression results showed that the higher expression of CLDN6 had a worse prognosis, including OS (HR = 2.19, 95% CI: 1.01–4.75, $p = 0.047$) (**Figure 6D**), DSS (HR = 2.63, 95% CI: 1.14–6.06, $p = 0.023$) (**Figure 6E**), and PFI (HR = 2.10, 95% CI: 1.12–3.94, $p = 0.021$) (**Figure 6F**). For BLCA, Cox regression results showed that the higher expression of CLDN6 had a worse prognosis, including OS (HR = 1.59, 95% CI: 1.18–2.13, $p = 0.002$) (**Figure 6G**), DSS (HR = 1.66, 95% CI: 1.16–2.38, $p = 0.005$) (**Figure 6H**), and PFI (HR = 1.54, 95% CI: 1.15–2.08, $p = 0.004$) (**Figure 6I**). For STAD, Cox regression results showed that the higher expression of CLDN6 had a worse prognosis, including OS (HR = 1.44, 95% CI: 1.04–2.00, $p = 0.03$) (**Figure 6J**), DSS (HR = 1.94, 95% CI: 1.26–2.97, $p = 0.003$) (**Figure 6K**), and PFI (HR = 1.78, 95% CI: 1.24–2.55, $p = 0.002$) (**Figure 6L**).

Furthermore, we investigated the correlations between CLDN6 and prognosis (OS, DSS, and PFI) in different clinical subgroups of UCEC. The results showed that the higher expression of CLDN6 had a worse OS in most clinical subgroups, including subgroup of age > 60 (**Figure 7A**), subgroup of weight > 80 (**Figure 7B**), subgroup of BMI > 30 (**Figure 7C**), subgroup of tumor invasion ≥ 50% (**Figure 7D**), subgroup of postmenopause (**Figure 7E**), subgroup of primary therapy outcome (CR) (**Figure 7F**), subgroup of residual tumor (R0) (**Figure 7G**), subgroup of radiation therapy (No) (**Figure 7H**), and subgroup of diabetes (No) (**Figure 7I**). For DSS, the higher expression of CLDN6 had a worse DSS in subgroup of age > 60 (**Figure 8A**), subgroup of weight > 80 (**Figure 8B**), subgroup of BMI > 30 (**Figure 8C**), subgroup of tumor invasion ≥ 50% (**Figure 8D**), subgroup of postmenopause (**Figure 8E**), subgroup of primary therapy outcome (CR) (**Figure 8F**), and subgroup of radiation therapy (No) (**Figure 8G**). For PFI, the higher expression of CLDN6 had a worse PFI in subgroup of weight > 80 (**Figure 9A**), subgroup of BMI > 30 (**Figure 9B**), subgroup of postmenopause (**Figure 9C**), subgroup of primary therapy outcome (CR) (**Figure 9D**), subgroup of residual tumor (R0) (**Figure 9E**), and subgroup of clinical stage III (**Figure 9F**).

CLDN6 Is Correlated With Different Clinical Characteristics in UCEC

We further presented the associations between CLDN6 and different clinical characteristics in UCEC and found that CLDN6 expression was significantly related to age, stage, weight, histological type, histologic grade, and menopause status of UCEC (**Table 1**). Moreover, CLDN6 was expressed higher in patients with age > 60 (**Figure 10A**), stage III/IV (**Figure 10D**), histologic grade 3 (**Figure 10E**), and histological type of serous (**Figure 10F**), while it was expressed lower in patients with weight > 80 (**Figure 10B**) and primary therapy outcome (CR) (**Figure 10C**), respectively.

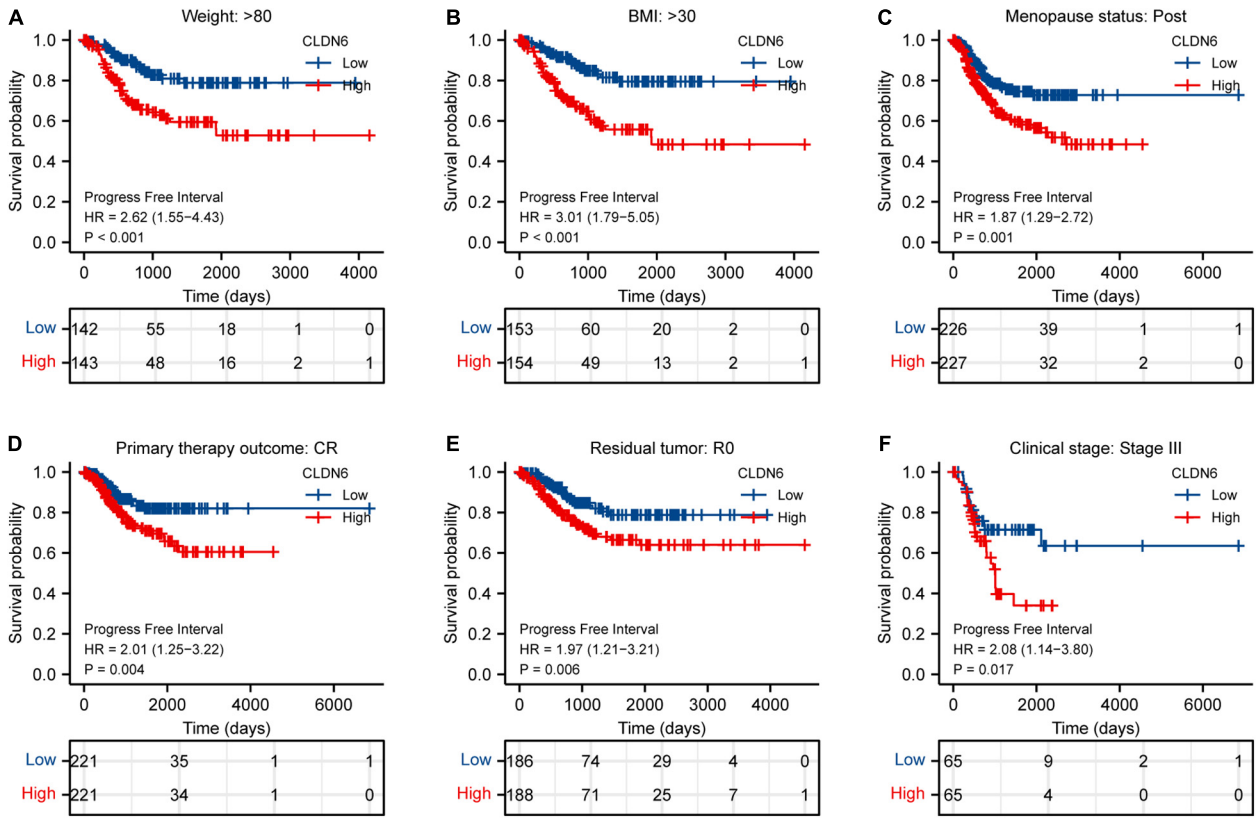


FIGURE 10 | Associations between CLDN6 expression and the PFI in different clinical subgroups of UCEC. **(A)** Weight > 80; **(B)** BMI > 30; **(C)** postmenopause; **(D)** primary therapy outcome (CR); **(E)** residual tumor (R0); **(F)** stage III.

Univariate and Multivariate Cox Regression Analyses in UCEC

We implemented univariate and multivariate Cox regression analyses of CLDN6 and clinical characteristics in UCEC. In the univariate and multivariate Cox regression analyses, clinical stage, primary therapy outcome, age, and CLDN6 were significantly associated with the OS (**Table 2**), while clinical stage, primary therapy outcome, histologic grade, and CLDN6 were significantly associated with DSS (**Supplementary Table 1**), and clinical stage, primary therapy outcome, and CLDN6 were significantly associated with PFI (**Supplementary Table 2**).

Co-expression Gene Analysis of CLDN6 in UCEC

We explored the top 50 co-expression genes positively or negatively correlated with CLDN6 expression in UCEC and displayed the correlations between CLDN6 expression and top 10 genes expression in the heatmap. In the heatmap of positive correlation (**Figure 11A**), we obtained the top 10 genes, including HIF3A (r = 0.71) (**Figure 11B**), GAL3ST3 (r = 0.66) (**Figure 11C**), PNOC (r = 0.64) (**Figure 11D**), AC068987.3 (r = 0.65) (**Figure 11E**), L1CAM (r = 0.66) (**Figure 11F**), AP000662.1 (r = 0.63) (**Figure 11G**), HMGA2 (r = 0.62) (**Figure 11H**), CLDN9 (r = 0.62) (**Figure 11I**), CLDN19 (r = 0.61) (**Figure 11J**), and EPHB2 (r = 0.61) (**Figure 11K**). In the heatmap of negative correlation (**Figure 12A**), we obtained the top 10 genes, including ELAPOR1 (r = −0.61) (**Figure 12B**), PIGR (r = −0.57) (**Figure 12C**), MLPH (r = −0.56) (**Figure 12D**), SPDEF (r = −0.56) (**Figure 12E**), FOXA2 (r = −0.54) (**Figure 12F**), TFF3 (r = −0.54) (**Figure 12G**), GLYATL2 (r = −0.53) (**Figure 12H**), IL20RA (r = −0.53) (**Figure 12I**), PLPP2 (r = −0.52) (**Figure 12J**), and LNCTAM34A (r = −0.50) (**Figure 12K**).

DEGs Between CLDN6 High Expression Group and Low Expression Groups in UCEC

A total of 2,884 DEGs were acquired with the threshold values of |log2 fold-change (FC)| > 1.0 and adjusted p-value < 0.05, including 2,005 upregulated genes and 879 downregulated genes (**Figure 13A**). Among them, 446 DEGs were obtained with the threshold values of |log2 fold-change (FC)| > 2.0 and adjusted p-value < 0.05, including 388 upregulated genes and 58 downregulated genes. Then, we conducted the GO and KEGG enrichment analyses (**Figures 13B,C**) of DEGs, revealing that the primary BP contained complement activation, classical pathway, humoral immune response mediated by circulating immunoglobulin, complement

TABLE 2 | Univariate and multivariate Cox regression analyses of clinical characteristics associated with OS of UCEC.

Characteristics	Total (*N*)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	*p*-Value	Hazard ratio (95% CI)	*p*-Value
Clinical stage (stage II and stage III and stage IV vs. stage I)	551	3.270 (2.145–4.984)	<0.001	3.225 (1.940–5.362)	<0.001
Primary therapy outcome (CR vs. PD and SD and PR)	480	0.129 (0.078–0.215)	<0.001	0.168 (0.100–0.283)	<0.001
Age (>60 vs. ≤60)	549	1.847 (1.160–2.940)	0.010	1.754 (1.049–2.931)	0.032
CLDN6 (high vs. low)	551	2.121 (1.383–3.250)	<0.001	1.669 (1.020–2.732)	0.041

activation, proteinactivation cascade, and humoral immune response. The CC was mainly enriched in immunoglobulin complex, immunoglobulin complex, circulating, transmembrane transporter complex, transporter complex, and ion channel complex. The MF was primarily involved in antigen binding, immunoglobulin receptor binding, receptor ligand activity, gated channel activity, and passive transmembrane transporter activity. The KEGG pathway enrichment was mainly related to neuroactive ligand–receptor interaction, cytokine–cytokine receptor interaction, cAMP signaling pathway, adrenergic signaling in cardiomyocytes, and protein digestion and absorption.

Furthermore, we obtained the top 10 hub genes (**Figure 13D**) of 446 DEGs, including CCK, NTS, AVP, ADRA1B, NPFFR2, CASR, GCGR, TACR3, UTS2R, and NTSR1. Among them, the top five hub genes (**Figure 13E**) were CCK, NTS, AVP, ADRA1B, and NPFFR2. Additionally, the top three hub genes (**Figure 13F**) were CCK, NTS, and AVP.

DISCUSSION

Claudins, together with occludins and junctional adhesion molecules, are three transmembrane components of tight junctions, playing the role of paracellular barrier and intracellular signaling, while regulating the proliferation, differentiation, and apoptosis of the epithelial cell (Severson and Parkos, 2009; Liu F. et al., 2016; Osanai et al., 2017; Kage et al., 2019; Phattarataratip and Sappayatosok, 2020). It has been demonstrated that claudins are intimately involved in the majority of epithelial-derived cancers by altering their expression manners (Baumholtz et al., 2017; Osanai et al., 2017; Singh et al., 2017; Tabaries and Siegel, 2017; Cherradi et al., 2019; Gowrikumar et al., 2019; Kage et al., 2019; Auguste et al., 2020; Koziel et al., 2020; Phattarataratip and Sappayatosok, 2020; Kolchakova et al., 2021). Recent studies have confirmed that CLDN6 could promote the cell migration and invasion of certain cancers, including breast cancer (Liu Y. et al., 2016; Lu et al., 2017), gastric cancer (Torres-Martinez et al., 2017; Yu et al., 2019), hepatocellular carcinoma (Huang et al., 2020; Kong et al., 2020), and endometrial cancer (Kojima et al., 2021), and can be vigorously stimulated in human-induced pluripotent stem cells developed from fibroblasts, thus might present to be a promising target candidate antibody for cancer therapy (Türeci et al., 2018).

However, there is no existing study to our knowledge to evaluate the significance of CLDN6 in pan-cancer on the whole scale. In the present study, to assess the expression level of CLDN6 across pan-cancer, we examined CCLE database, TCGA database, and GTEx database and found that it was significantly upregulated in 20 types of human cancers, including ACC, BLCA, BRCA, CHOL, COAD, ESCA, HNSC, LIHC, LUAD, LUSC, OV, PAAD, PCPG, READ, STAD, TGCT, THCA, THYM, UCEC, and UCS, while it was downregulated in GBM, KICH, KIRC, LAML, and LGG. The finding reveals that CLDN6 serves as a cancer-promoting gene in the majority of malignant tumors and might be involved in tumor formation or cancer development. In addition, there are meaningful associations between CLDN6 expression level and molecular subtypes of seven cancer types, including UCEC, BRCA, ESCA, LUSC, HNSC, OV, and STAD. For instance, CLDN6 was observed to express the highest in the molecular subtype of CN_HIGH in UCEC, in the molecular subtype of basal in BRCA, and in the molecular subtype of CIN in STAD. Meanwhile, we observed that CLDN6 was significantly correlated with different immune subtypes in nine cancer types. It is worth pointing out that CLDN6 is closely correlated with both molecular subtype and immune subtype in six types of cancers, including UCEC, BRCA, STAD, OV, LUSC, and HNSC. As the previous study demonstrated, CLDN6 is confirmed to be a possible single prognostic marker and promising therapeutic target for a subgroup of intestinal type gastric cancer (Kohmoto et al., 2020). Hence, the research focused on a distinct molecular subtype or immune subtype of cancers may provide an appropriate entry point to explore the role of CLDN6. Then, to examine the diagnostic and prognostic value of CLDN6, we performed the ROC curve and the Kaplan–Meier survival curve in pan-cancer and found that CLDN6 had a certain accuracy (AUC > 0.7) in predicting 15 cancer types, especially had a high accuracy (AUC > 0.9) in predicting LAML, TGCT, OV, and UCS. Additionally, CLDN6 was significantly correlated with the OS, DSS, and PFI in UCEC, ACC, BLCA, and STAD. The results demonstrate here that CLDN6 presents great diagnostic and prognostic significance in the above cancers and may be a potential biomarker or therapeutic target for precision oncology. To deeply investigate the biofunction of CLDN6, we conducted the GO and KEGG pathway enrichment analyses of its 50 targeted binding proteins, revealing that the BP was major in cell–cell adhesion or bicellular tight junction, its MF was primarily involved in virus receptor activity, hijacked MF, miRNA binding, and regulatory RNA binding, and the main pathways were enriched in leukocyte transendothelial migration, tight

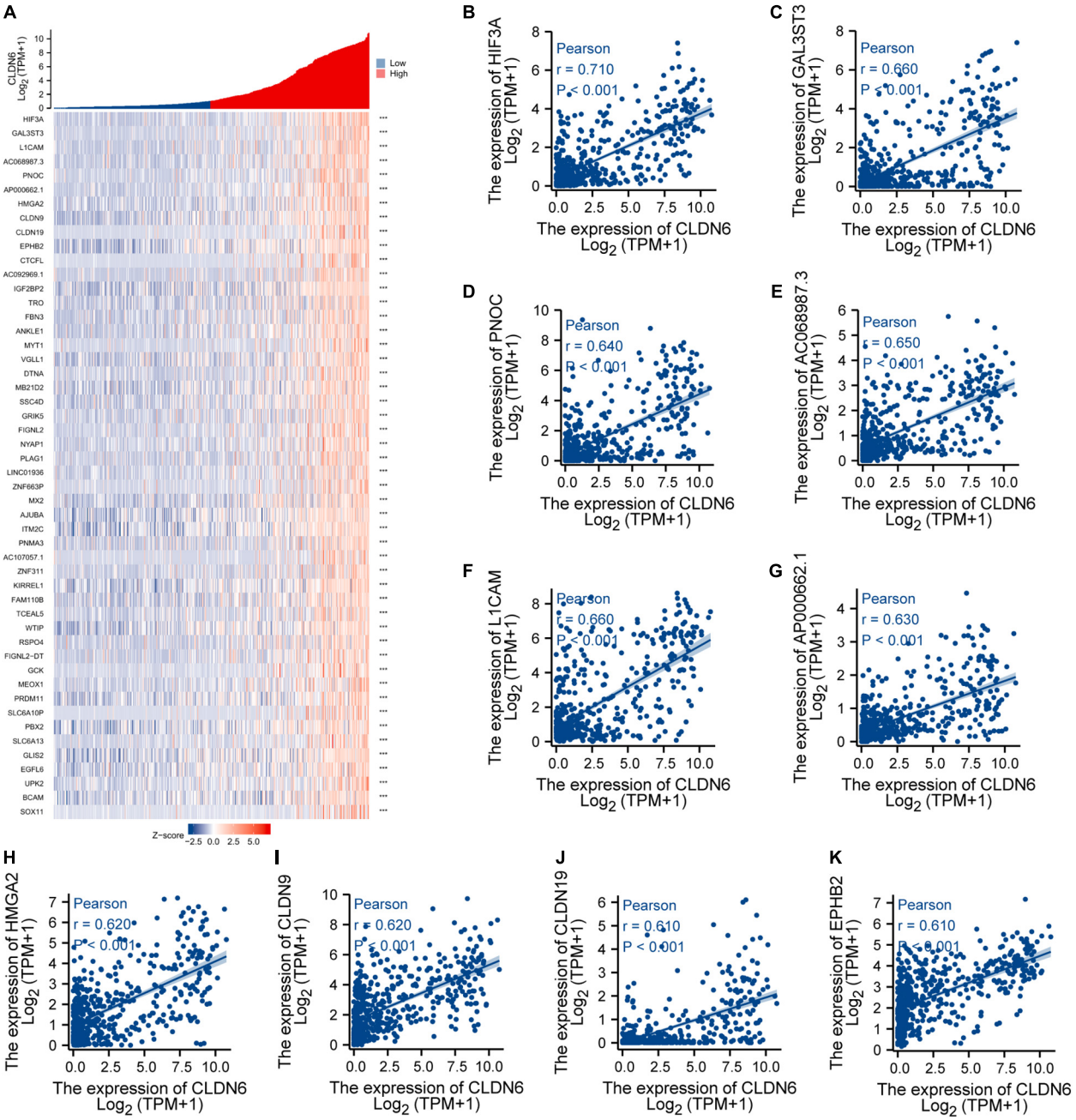


FIGURE 11 | Top 50 genes positively correlated with CLDN6 expression in UCEC. **(A)** The gene co-expression heatmap of the top 50 genes positively correlated with CLDN6 in UCEC; **(B–K)** correlation analysis of the top 10 genes and CLDN6 in the heatmap.

junction, cell adhesion molecules, hepatitis C, and pathogenic *E. coli* infection. It should be emphasized that CLDN6 is not only essential for cell–cell adhesion and tight junction but also crucial for outside–in cell signaling.

Furthermore, we primarily analyzed the role of CLDN6 played in UCEC and identified the significant correlations between CLDN6 expression level and age, stage, weight, histological type, histologic grade, and menopause status. Subsequently, we discovered that CLDN6 high expression could cause a worse OS, DSS, or PFI in a variety of clinical subgroups of UCEC, yet cause a worse all of the OS, DSS, and PFI only in clinical subgroups of weight > 80, BMI > 30, postmenopause, or primary therapy outcome of CR. Since then, we confirmed clinical stage, primary therapy outcome, and CLDN6 expression level as independent risk factors in OS, DSS, and PFI of UCEC through univariate and multivariate Cox regression analyses. The results are further supported by the prior research involving prognostic significance of aberrant CLDN6 expression in UCEC (Kojima et al., 2020),

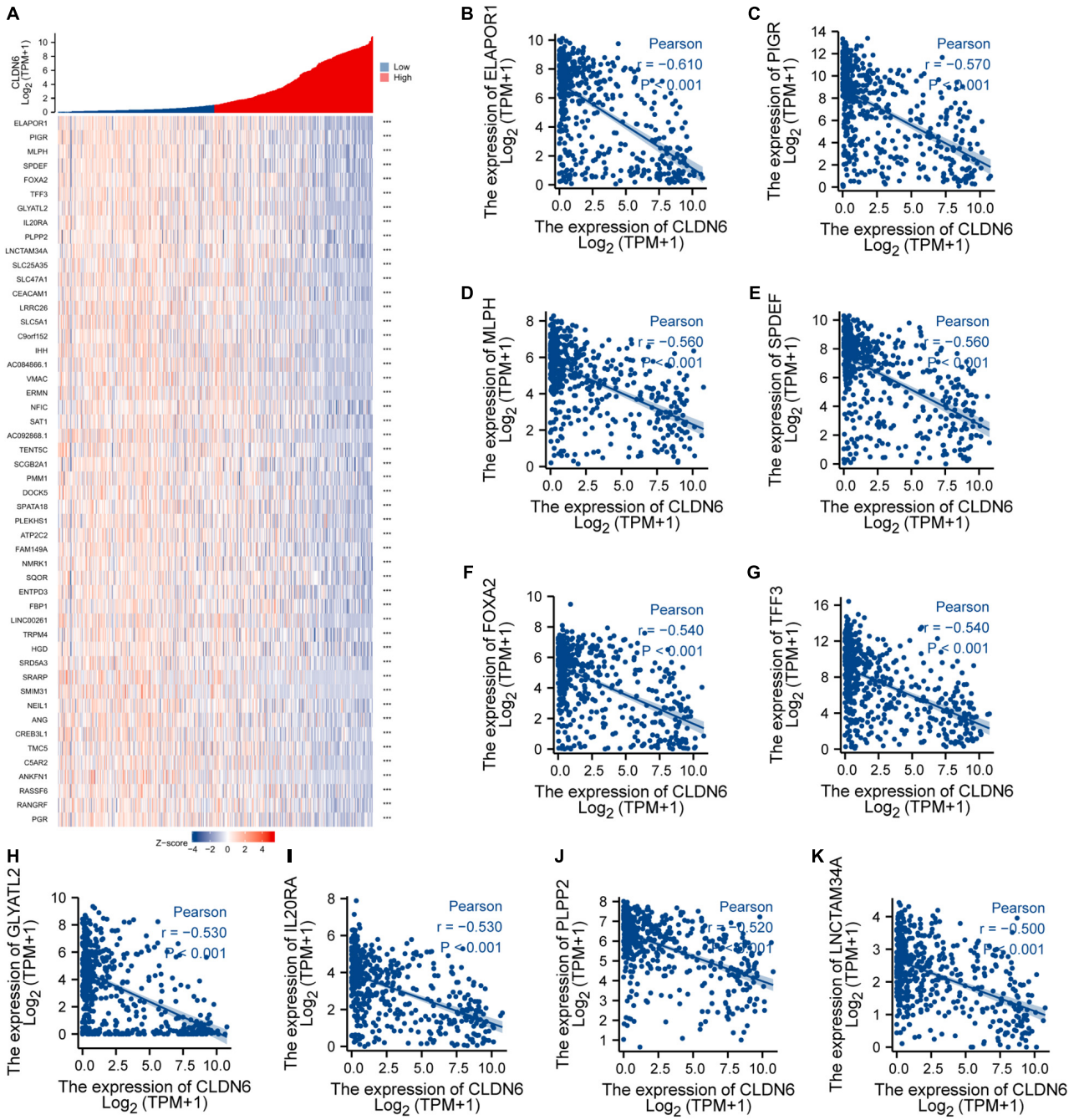


FIGURE 12 | Top 50 genes negatively correlated with CLDN6 expression in UCEC. **(A)** The gene co-expression heatmap of the top 50 genes negatively correlated with CLDN6 in UCEC; **(B–K)** correlation analysis of the top 10 genes and CLDN6 in the heatmap.

which established a monoclonal antibody, then performed immunohistochemical staining and semi-quantification to assess the associations between CLDN6 and clinical characteristics in 173 cases, and finally identified stages III/IV, distant metastasis, and high CLDN6 expression as independent prognostic variables for the OS. It is of great importance that our findings provide a more comprehensive and further supplement and strengthen the role of CLDN6 played in UCEC, and that we precisely analyze the correlations between CLDN6 expression and different prognosis conditions (OS, DSS, or PFI) in various clinical subgroups of UCEC. In addition, we further obtained co-expression genes of CLDN6, and the top 10 co-expression genes positively correlated with CLDN6 contained HIF3A, GAL3ST3, PNOC, AC068987.3 (Lnc-SCN8A-2), L1CAM, AP000662.1 (Lnc-CLP1-3), HMGA2, CLDN9, CLDN19, and EPHB2, whereas the top 10 co-expression genes negatively correlated with CLDN6 contained ELAPOR1,

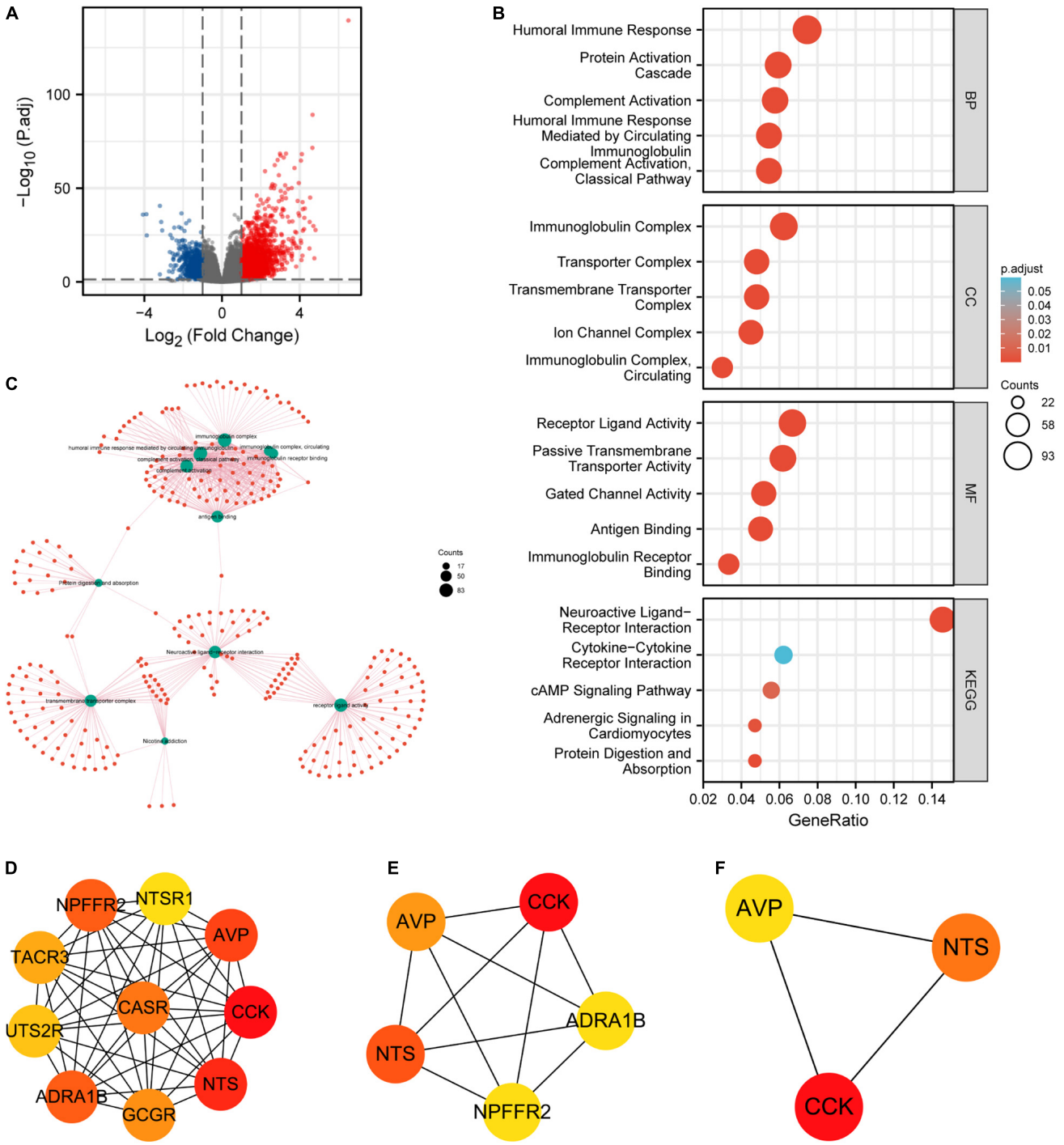


FIGURE 13 | Protein–protein interaction (PPI) network building and GO and KEGG analyses of DEGs between CLDN6 high expression and low expression groups in UCEC. **(A)** The volcano map of DEGs (red: upregulation; blue: downregulation); **(B,C)** GO and KEGG analyses of DEGs; **(D–F)** hub genes of PPI network and MCODE2 components identified in the gene lists.

PIGR, MLPH, SPDEF, FOXA2, TFF3, GLYATL2, IL20RA, PLPP2, and LNCTAM34A. Besides, we explored the DEGs between CLDN6 high expression group and low expression group, then performed the GO and KEGG pathway analyses of the DEGs, and found that the BP mainly linked to complement activation, classical pathway, humoral immune response mediated by circulating immunoglobulin, complement activation, protein activation cascade, and humoral immune response. The MF was

primarily involved in antigen binding, immunoglobulin receptor binding, receptor ligand activity, gated channel activity, and passive transmembrane transporter activity. Finally, we screened out the hub genes of DEGs, including CCK, NTS, AVP, ADRA1B, NPFFR2, CASR, GCGR, TACR3, UTS2R, and NTSR1.

There are still several limitations in the present study. On the one hand, we set out to explore CLDN6 only using CCLE database, TCGA database, and GTEx database while lacking actual clinical data. On the other hand, the precise verification and high-quality evidence should be further performed and provided by biological experiments.

Many computational methods have been applied in bioinformatics research, such as lncRNA–miRNA interaction predictions (Liu et al., 2020; Zhang et al., 2021a,b), the identification of microRNA combinatorial biomarkers (Yang et al., 2017a,b; Liu and Yang, 2018), and so on (Wang et al., 2020; Zhang et al., 2020). In the next step, we will consider the use of the machine learning and deep learning methods in the research of CLDN6. In summary, the identification of CLDN6 in diagnostic and prognostic significance across pan-cancer, together with its further exploration in UCEC, could add a new dimension to the comprehensive understanding of its critical role in tumor promotion and suppression and provide an integrative analyzing basis for deep verification of molecular biology experiments, even for future clinical application of cancer therapies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

QZ and LO contributed to the design of the study protocol. CZ performed the statistical analysis and drew the pictures. YL and KL contributed to the writing of the study protocol. CG downloaded the data. All authors approved the final version of the manuscript.

FUNDING

This work was supported by the Scientific Research Funding Project of Liaoning Provincial Department of Science and Technology under Grant No. 2020JH2/10300050.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.726656/full#supplementary-material

REFERENCES

Auguste, A., Blanc-Durand, F., Deloger, M., Le Formal, A., Bareja, R., Wilkes, D. C., et al. (2020). Small cell carcinoma of the ovary, hypercalcemic type (SCCOHT) beyond SMARCA4 mutations: a comprehensive genomic analysis. *Cells* 9:1496. doi: 10.3390/cells9061496

Barmeyer, C., Schulzke, J. D., and Fromm, M. (2015). Claudin-related intestinal diseases. *Semin. Cell Dev. Biol.* 42, 30–38. doi: 10.1016/j.semcdb.2015.05.006

Baumholtz, A. I., Gupta, I. R., and Ryan, A. K. (2017). Claudins in morphogenesis: forming an epithelial tube. *Tissue Barriers* 5:e1361899. doi: 10.1080/21688370.2017.1361899

Birks, D. K., Kleinschmidt-Demasters, B. K., Donson, A. M., Barton, V. N., Mcnatt, S. A., Foreman, N. K., et al. (2010). Claudin 6 is a positive marker for atypical teratoid/rhabdoid tumors. *Brain Pathol.* 20, 140–150. doi: 10.1111/j.1750-3639.2008.00255.x

Cherradi, S., Martineau, P., Gongora, C., and Del Rio, M. (2019). Claudin gene expression profiles and clinical value in colorectal tumors classified according to their molecular subtype. *Cancer Manag. Res.* 11, 1337–1348. doi: 10.2147/CMAR.S188192

Gowrikumar, S., Singh, A. B., and Dhawan, P. (2019). Role of claudin proteins in regulating cancer stem cells and chemoresistance-potential implication in disease prognosis and therapy. *Int. J. Mol. Sci.* 21:53. doi: 10.3390/ijms21010053

Huang, L., Zhao, C., Sun, K., Yang, D., Yan, L., Luo, D., et al. (2020). Downregulation of CLDN6 inhibits cell proliferation, migration, and invasion via regulating EGFR/AKT/mTOR signalling pathway in hepatocellular carcinoma. *Cell Biochem. Funct.* 38, 541–548. doi: 10.1002/cbf.3489

Kage, H., Flodby, P., Zhou, B., and Borok, Z. (2019). Dichotomous roles of claudins as tumor promoters or suppressors: lessons from knockout mice. *Cell. Mol. Life Sci.* 76, 4663–4672. doi: 10.1007/s00018-019-03238-7

Kohmoto, T., Masuda, K., Shoda, K., Takahashi, R., Ujiro, S., Tange, S., et al. (2020). Claudin-6 is a single prognostic marker and functions as a tumor-promoting gene in a subgroup of intestinal type gastric cancer. *Gastric Cancer* 23, 403–417. doi: 10.1007/s10120-019-01014-x

Kojima, M., Sugimoto, K., Kobayashi, M., Ichikawa-Tomikawa, N., Kashiwagi, K., Watanabe, T., et al. (2021). Aberrant claudin-6-adhesion signaling promotes endometrial cancer progression via estrogen receptor alpha. *Mol. Cancer Res.* 19, 1208–1220. doi: 10.1158/1541-7786.MCR-20-0835

Kojima, M., Sugimoto, K., Tanaka, M., Endo, Y., Kato, H., Honda, T., et al. (2020). Prognostic significance of aberrant claudin-6 expression in endometrial cancer. *Cancers (Basel)* 12:2748. doi: 10.3390/cancers12102748

Kolchakova, D., Moten, D., Batsalova, T., and Dzhambazov, B. (2021). Tight junction protein claudin-12 is involved in cell migration during metastasis. *Biomolecules* 11:636. doi: 10.3390/biom11050636

Kong, F. E., Tang, Y. Q., Gong, Y. F., Mo, J. Q., Zhao, Y., Li, M. M., et al. (2020). Identification of prognostic claudins signature in hepatocellular carcinoma from a hepatocyte differentiation model. *Hepatol. Int.* 14, 521–533. doi: 10.1007/s12072-020-10035-z

Koziel, M. J., Kowalska, K., and Piastowska-Ciesielska, A. W. (2020). Claudins: new players in human fertility and reproductive system cancers. *Cancers (Basel)* 12:711. doi: 10.3390/cancers12030711

Lal-Nag, M., and Morin, P. J. (2009). The claudins. *Genome Biol.* 10:235. doi: 10.1186/gb-2009-10-8-235

Liu, F., Koval, M., Ranganathan, S., Fanayan, S., Hancock, W. S., Lundberg, E. K., et al. (2016). Systems proteomics view of the endogenous human claudin protein family. *J. Proteome Res.* 15, 339–359. doi: 10.1021/acs.jproteome.5b00769

Liu, H., Ren, G., Chen, H., Liu, Q., Yang, Y., and Zhao, Q. (2020). Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized. *Knowl. Based Syst.* 191:105261. doi: 10.1016/j.knosys.2019.105261

Liu, K., and Yang, Y. (2018). Incorporating link information in feature selection for identifying tumor biomarkers by using miRNA-mRNA paired expression data. *Curr. Proteomics* 15, 165–171. doi: 10.2174/1570164614666171031160232

Liu, Y., Jin, X., Li, Y., Ruan, Y., Lu, Y., Yang, M., et al. (2016). DNA methylation of claudin-6 promotes breast cancer cell migration and invasion by recruiting MeCP2 and deacetylating H3Ac and H4Ac. *J. Exp. Clin. Cancer Res.* 35:120. doi: 10.1186/s13046-016-0396-x

Lu, Y., Wang, L., Li, H., Li, Y., Ruan, Y., Lin, D., et al. (2017). SMAD2 inactivation inhibits CLDN6 methylation to suppress migration and invasion of breast cancer cells. *Int. J. Mol. Sci.* 18:1863. doi: 10.3390/ijms18091863

Osanai, M., Takasawa, A., Murata, M., and Sawada, N. (2017). Claudins in cancer: bench to bedside. *Pflugers Arch.* 469, 55–67. doi: 10.1007/s00424-016-1877-7

Peter, Y., and Goodenough, D. (2004). Claudins. *Curr. Biol.* 14, R293–R294. doi: 10.1016/j.cub.2004.03.045

Phattarataratip, E., and Sappayatosok, K. (2020). The significance of relative claudin expression in odontogenic tumors. *Head Neck Pathol.* 14, 480–488. doi: 10.1007/s12105-019-01072-8

Reinhard, K., Rengstl, B., Oehm, P., Michel, K., Billmeier, A., Hayduk, N., et al. (2020). An RNA vaccine drives expansion and efficacy of claudin-CAR-T cells against solid tumors. *Science* 367, 446–453. doi: 10.1126/science.aay5967

Ru, B., Wong, C. N., Tong, Y., Zhong, J. Y., Zhong, S. S. W., Wu, W. C., et al. (2019). TISIDB: an integrated repository portal for tumor-immune system interactions. *Bioinformatics* 35, 4200–4202. doi: 10.1093/bioinformatics/btz210

Severson, E. A., and Parkos, C. A. (2009). Mechanisms of outside-in signaling at the tight junction by junctional adhesion molecule A. *Ann. N. Y. Acad. Sci.* 1165, 10–18. doi: 10.1111/j.1749-6632.2009.04034.x

Singh, A. B., Uppada, S. B., and Dhawan, P. (2017). Claudin proteins, outside-in signaling, and carcinogenesis. *Pflugers Arch.* 469, 69–75. doi: 10.1007/s00424-016-1919-1

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sugimoto, K., Ichikawa-Tomikawa, N., Kashiwagi, K., Endo, C., Tanaka, S., Sawada, N., et al. (2019). Cell adhesion signals regulate the nuclear receptor activity. *Proc. Natl. Acad. Sci. U.S.A.* 116, 24600–24609. doi: 10.1073/pnas.1913346116

Tabaries, S., and Siegel, P. M. (2017). The role of claudins in cancer metastasis. *Oncogene* 36, 1176–1190. doi: 10.1038/onc.2016.289

Torres-Martinez, A. C., Gallardo-Vera, J. F., Lara-Holguin, A. N., Montano, L. F., and Rendon-Huerta, E. P. (2017). Claudin-6 enhances cell invasiveness through claudin-1 in AGS human adenocarcinoma gastric cancer cells. *Exp. Cell Res.* 350, 226–235. doi: 10.1016/j.yexcr.2016.11.025

Türeci, Ö, Wagner, M., Paret, C., Kreuzberg, M. M., Wöll, S., Walter, K., et al. (2018). Claudin 6 is a carcinoembryonic antigen with cancer stem cell marker features. *Cancer Res.* 78, 1907–1907. doi: 10.1158/1538-7445.Am2018-1907

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., et al. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* 35, 314–316. doi: 10.1038/nbt.3772

Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of membrane protein types via multivariate information fusion with Hilbert–Schmidt independence criterion. *Neurocomputing* 383, 257–269. doi: 10.1016/j.neucom.2019.11.103

Yang, Y., Huang, N., Hao, L., and Kong, W. (2017a). A clustering-based approach for efficient identification of microRNA combinatorial biomarkers. *BMC Genomics* 18:210. doi: 10.1186/s12864-017-3498-8

Yang, Y., Xiao, Y., Cao, T., and Kong, W. (2017b). MiRFFS: a functional group-based feature selection method for the identification of microRNA biomarkers. *Int. J. Data Min. Bioinformatics* 18, 40–55. doi: 10.1504/IJDMB.2017.086099

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, S., Zhang, Y., Li, Q., Zhang, Z., Zhao, G., and Xu, J. (2019). CLDN6 promotes tumor progression through the YAP1-snail1 axis in gastric cancer. *Cell Death Dis.* 10:949. doi: 10.1038/s41419-019-2168-y

Zhang, J., Zhang, Z., Pu, L., Tang, J., and Guo, F. (2020). AIEpred: an ensemble predictive model of classifier chain to identify anti-inflammatory peptides. *IEEE ACM Trans. Comput. Biol. Bioinformatics* 1–1. doi: 10.1109/TCBB.2020.2968419

Zhang, L., Liu, T., Chen, H., Zhao, Q., and Liu, H. (2021a). Predicting lncRNA–miRNA interactions based on interactome network and graphlet interaction. *Genomics* 113, 874–880. doi: 10.1016/j.ygeno.2021.02.002

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021b). Using network distance analysis to predict lncRNA–miRNA interactions. *Interdiscip. Sci. Comput. Life Sci.* doi: 10.1007/s12539-021-00458-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Zhang, Guo, Li, Liu, Zhao and Ouyang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/fcell.2021.700355

Prediction of Genetic Factors of Hyperthyroidism Based on Gene Interaction Network

Fei Shen[1,2†], Wensong Cai[1,2†], Xiaoxiong Gan[1,2], Jianhua Feng[1,2], Zhen Chen[1,2], Mengli Guo[1,2], Fang Wei[3], Jie Cao[3] and Bo Xu[1,2*]*

[1] Department of Thyroid Surgery, School of Medicine, Guangzhou First People's Hospital, South China University of Technology, Guangzhou, China, [2] Department of Thyroid Surgery, Guangzhou First People's Hospital, Guangzhou Medical University, Guangzhou, China, [3] Department of General Surgery, School of Medicine, Guangzhou First People's Hospital, South China University of Technology, Guangzhou, China

The number of hyperthyroidism patients is increasing these years. As a disease that can lead to cardiovascular disease, it brings great potential health risks to humans. Since hyperthyroidism can induce the occurrence of many diseases, studying its genetic factors will promote the early diagnosis and treatment of hyperthyroidism and its related diseases. Previous studies have used genome-wide association analysis (GWAS) to identify genes related to hyperthyroidism. However, these studies only identify significant sites related to the disease from a statistical point of view and ignore the complex regulation relationship between genes. In addition, mutation is not the only genetic factor of causing hyperthyroidism. Identifying hyperthyroidism-related genes from gene interactions would help researchers discover the disease mechanism. In this paper, we purposed a novel machine learning method for identifying hyperthyroidism-related genes based on gene interaction network. The method, which is called "RW-RVM," is a combination of Random Walk (RW) and Relevance Vector Machines (RVM). RW was implemented to encode the gene interaction network. The features of genes were the regulation relationship between genes and non-coding RNAs. Finally, multiple RVMs were applied to identify hyperthyroidism-related genes. The result of 10-cross validation shows that the area under the receiver operating characteristic curve (AUC) of our method reached 0.9, and area under the precision-recall curve (AUPR) was 0.87. Seventy-eight novel genes were found to be related to hyperthyroidism. We investigated two genes of these novel genes with existing literature, which proved the accuracy of our result and method.

Keywords: hyperthyroidism, gene interaction network, random walk, a novel machine learning method, multiple Relevance Vector Machines

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Haixiu Yang,
Harbin Medical University, China
Lingling Zhao,
Harbin Institute of Technology, China

****Correspondence:***
Jie Cao
eycaojie@scut.edu.cn
Bo Xu
eyboxu@scut.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *26 April 2021*
Accepted: *02 June 2021*
Published: *02 August 2021*

Citation:
Shen F, Cai W, Gan X, Feng J, Chen Z, Guo M, Wei F, Cao J and Xu B (2021) Prediction of Genetic Factors of Hyperthyroidism Based on Gene Interaction Network. Front. Cell Dev. Biol. 9:700355. doi: 10.3389/fcell.2021.700355

INTRODUCTION

Hyperthyroidism refers to a type of common endocrinology disease in which the level of thyroid hormone in the body is abnormally increased (Ross, 2011). The main manifestations of patients are palpitations, sweating, increased eating but weight loss, etc., which are harmful to the cardiovascular system (Ertek and Cicero, 2013), digestive system (Journy et al., 2017), nervous

system (Ku et al., 2005), etc. Hyperthyroidism will cause a certain degree of damage, which will seriously affect the quality of life of patients. Hyperthyroidism occurs mostly in the elderly and pregnant women. Hyperthyroidism has significant adverse effects on the heart, bones, and cognitive functions. Standardized management is of great significance for improving the clinical diagnosis and treatment of thyroid diseases in the elderly and protecting the health of the elderly (Martin and Deam, 1996). The thyroid hormone is very important for the health of pregnant women and the growth and development of the fetus. During pregnancy, thyroid function will undergo a series of physiological changes (Moleti et al., 2019). Hyperthyroidism during pregnancy can cause adverse pregnancy outcomes and maternal and infant complications.

The common causes of hyperthyroidism in the elderly include Graves' disease (GD) (Okosieme et al., 2019), toxic multinodular goiter (TMNG) (Azizi et al., 2019), and toxic adenomas (TA) (Sharma and Abraham, 2020). Other rare causes include pituitary thyroid-stimulating hormone (TSH) tumors, nourishment cell layer tumors, and metastatic differentiated thyroid cancer. The etiology spectrum of elderly hyperthyroidism is not completely consistent with that of non-elderly patients: GD is mainly observed in iodine-sufficient areas (Liu et al., 2017), and it decreases with age after the age of 40 years; in iodine-deficient areas, TMNG is the main source, and the proportion of people over 60 years old with hyperthyroidism is 28–65%, while that of people under 40 years old with hyperthyroidism is only 5–10%. In pregnant women the blood TSH drops by 30–50% in the first trimester and gradually rises in the second trimester (Smaniotto et al., 2017). However, it is worth noting that the TSH of some pregnant women cannot return to the non-pregnancy level during the third trimester. The magnitude of this physiological change of TSH is also different in different races. These physiological changes leading to the diagnosis of abnormal thyroid function during pregnancy must use the pregnancy-specific reference range. Compared with the reference range of TSH for non-pregnant women (Moon et al., 2015), the lower limit of the TSH reference range for pregnant women is reduced by about 0.1–0.2 mU/L, and the upper limit of TSH is reduced by about 0.5–1.0 mU/L. And, the decrease in TSH is more obvious than in the middle and late pregnancy (Asik et al., 2014). Women with twin and multiple pregnancies have higher hCG concentrations in their bodies, and their TSH also decreases more significantly than women with single pregnancies.

Although several clinical studies have researched the causal reason of hyperthyroidism, the genetic factors of hyperthyroidism are still unclear. GD is caused by genetic and environmental factors, and 79% susceptibility can be attributed to genetic factors (Kuś et al., 2015). Through candidate gene analysis, genome-wide association study, and some functional studies, researchers had identified several susceptible genes of GD, such as HLA, CTLA4, PTPN22, FCRL3, RNASET2, and TSHR (Chu et al., 2011; Simmonds, 2013). GD can occur at any age, but the cutoff age of early onset GD has not been clearly defined (most researchers chose patients younger than or 30 years old as early onset GD). Research about the unique loci of early onset GD were reported several years ago, but considering their insufficient sample size and deficient analysis methods, these loci should be verified in more studies. A study did genotyping of 196,524 polymorphisms in 106 early onset patients (onset at age < 30 years) and 855 healthy subjects through Illumina Infinium Immunochip and performed case–control association analyses, which finally found 30 specific single-nucleotide polymorphisms (SNPs) in early onset patients (these SNPs were located in the genes of HLA-I, HLA-II, BTNL2, NOTCH4, TNFAIP3, and CXCR4). Another study had reported that FOXP3 in the X chromosome was a unique gene of early onset patients (onset at age ≤ 30 years).

Although genome-wide association analysis (GWAS) can reveal susceptible genes of hyperthyroidism, this method does not take gene interaction into account. However, the regulation relationship between genes plays a vital role in the occurrence and development of diseases (Maurano et al., 2012; Zhao et al., 2020b). Previous studies have proven the effectiveness of biological network in identifying diseases-related molecules (Peng and Zhao, 2020). Multiple computational methods have been developed to process the biological network, which achieved drug–target interaction prediction (Tianyi et al., 2020), regulatory relationship prediction (Zhao et al., 2021a), diseases-related metabolites (Zhao et al., 2020a), etc. Random Walk (RW) is a common method to deal with biological networks. Researchers have applied this method to identify miRNAs (Niu et al., 2019), proteins (Rathore and de Pablo, 2002), gene expression (Zhao et al., 2021b), and multiple diseases-related molecules. In addition, Relevance Vector Machines (RVM) has been also widely used in disease-related research. Veeramani and Muthusamy (2016) used RVM-classified lung image to identify lung cancer. Wang et al. (2013) implemented early diagnosis of Parkinson's disease by RVM. In this paper, we propose a novel method called "RW-RVM" to predict hyperthyroidism. Using RW-RVM, we extracted the topological relations of gene interaction network and fused them with regulation relationship of non-coding RNAs. Multiple RVM models were implemented to predict hyperthyroidism-related genes by weight voting.

MATERIALS AND METHODS

Construction of Gene Network

First, we obtained hyperthyroidism-related known genes from DisGeNET (Piñero et al., 2020). Two hundred sixty-nine have been reported to be related to hyperthyroidism. These genes were input into the String database (Szklarczyk et al., 2016), and the gene interaction network was constructed as shown in **Figure 1**.

As we can see in **Figure 1**, most of these genes have close interactions with each other, which supports our hypothesis: Disease-related genes are related to gene interactions. Therefore, the genes that can interact with these 269 genes are potential genes related to hyperthyroidism. Our aim is to determine the possibility that these genes are associated with hyperthyroidism.

From HumanNet database (Hwang et al., 2019), we obtained 1,517 genes that can interact with these 269 genes. Using these 1,786 genes, we constructed a gene interaction network.

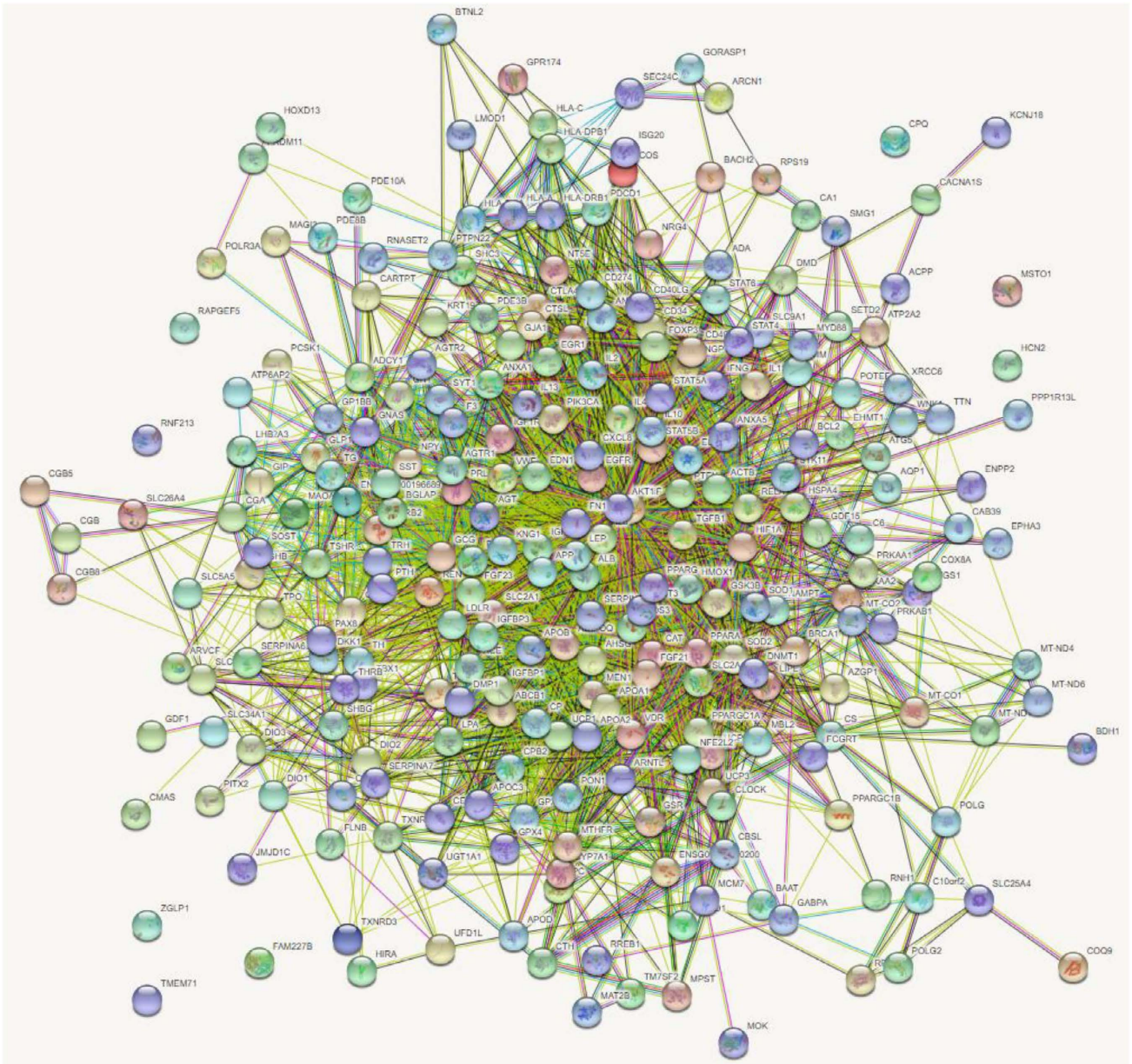


FIGURE 1 | Two hundred sixty-nine hyperthyroidism-related genes network.

The nodes in this network are genes, and the edges are interactions between genes.

Random Walk

The edges of the gene network have different weights. Based on the edges and known 269 genes, we implemented RW to traverse this network to reach steady state. Because the gene interaction network we constructed is a two-dimensional graph, when we travel in the form of probability according to the interaction scores and we know the current gene node information, the historical gene node traversal information has nothing to do with the future gene node traversal path. Therefore, we can regard the method of mining disease-related genes based on RW as a Markov chain. In each step of the Markov chain, the probability distribution of hyperthyroidism-related genes can change from one state to another or maintain the current state. The change of state is called transition, and the probability associated with different states is called transition probability.

If A is the adjacency matrix of the gene interaction network, we can normalize A to:

$$P = D^{-1}A \tag{1}$$

D is a diagonal matrix, which is the degree matrix of the gene interaction network, and its diagonal element is

$D(i, i) = \sum A(i, j)$. Here, P is a RW matrix, and the sum of transition probability of each node is 1, which is the probability matrix of hyperthyroidism associated with all genes.

A RW matrix corresponds to a Markov chain, and the probability distribution of hyperthyroidism-related genes changes with the change of state in the Markov chain. The probability from any state to the next state is as follows:

$$P_{t+1} = A_t P \tag{2}$$

This process continues, and the relationship between hyperthyroidism and genes is also changing. After a period of time, it reaches equilibrium. The equilibrium state, also known as steady state, means that the probability distribution of the association between hyperthyroidism and genes does not change. The calculation method of the steady state is as follows:

$$\pi = D(i, j) / \sum_i \sum_j A(i, j) \tag{3}$$

When $\pi P = \pi$, the whole system reaches steady state. This steady state is the final calculated correlation between hyperthyroidism and genes. The workflow of encoding the gene interaction network is shown in **Table 1**.

Extracting Gene Feature

Since the relationship between genes and diseases is related to the gene regulation, we extract the regulatory relationship between non-coding RNAs and genes as a feature of genes. The features of each gene could be represented as follows:

$$gene_t = \begin{bmatrix} l_1, l_2, ..., l_n \\ m_1, m_2, ..., m_n \end{bmatrix} \tag{4}$$

l_1 represents the regulatory relationship between lncRNA1 and $gene_t$, and m_1 represents the regulatory relationship between miRNA1 and $gene_t$. If lncRNA1 can interact with $gene_t$, $l_1 = 1$; otherwise, $l_1 = 0$

After obtaining all features of the genes, the final features of genes are the following:

$$Feature = P_{t+1} \bullet gene \tag{5}$$

TABLE 1 | Encoding gene interaction network based on random walk.

INPUT:Probability matrix of initial hyperthyroidism-related genes $_{P_0}$, parameter γ,control accuracy is $_{\ell}$, iteration times is $_N$, gene interaction network G(V, E, W)
OUTPUT:Probability matrix of hyperthyroidism-related genes

1.FOR i = 0 to $_N$\{
2. Update the probability matrix of hyperthyroidism-related genes $_{P_{t+1}=(1-\gamma)AP_t+\gamma P_0}$;
3. IF $||P_{t+1} - P_t|| > \ell$\{
4. $_{P_{t+1}=(1-\gamma)AP_t+\gamma P_0}$;
5. \}
6. ELSE \{
7. RETURN Probability matrix $_{P_{t+1}}$;
8. \}
9. \}

Therefore, the final features not only contain the regulatory relationship between non-coding RNAs and genes but also the topological characteristics of the gene interaction network. In the next step, these features could be input into RVM models to obtain the probability of genes associated with hyperthyroidism.

Construction of RVM Model

The kernel function of the RVM is not limited by Mercer conditions; it is more sparse and has less super-parameters so it reduces the computational burden of kernel functions.

For a given dataset $\{x_i, t_i\} i = 1^N$, $x_i \in \mathrm{R}^d$, the non-linear model is:

$$t = y(x) + \varepsilon \tag{6}$$

N is the sample number, y is the non-linear function, ε is the noise, $\varepsilon \sim N(0, \sigma^2)$.

The final function of RVM is:

$$t = \Phi\omega + \varepsilon \tag{7}$$

$\omega = (\omega_0, \cdots, \omega_N)^{\mathrm{T}}$ is the weight, and Φ is the matrix of the kernel function. $K()$ is the kernel function. $\phi_i(x_i) = [1, K(x_i, x_1), \cdots, K(x_i, x_N)]$, $i = 1, 2, \cdots, N$.

The distribution of $p(t|x)$ obeys $N(t|y(x), \sigma^2)$. The likelihood estimation of data is:

$$p(t|\omega, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\{-||t - \Phi\omega||^2/(2\sigma^2)\} \tag{8}$$

Tipping defines a zero mean Gauss-type prior distribution on ω :

$$p(\omega/\alpha) = \prod_0^N N(\omega_i|0, \alpha_i^{-1}) = \prod_0^N \frac{\alpha_i}{\sqrt{2\pi}} \exp(\frac{\omega_i^2\alpha_i}{2}) \tag{9}$$

α is the super-parameter; it has a one-to-one correspondence with the weight.

α and the variance of noise σ^2 meet the Gamma distribution.

$$\begin{aligned} p(\alpha) &= \textstyle\prod_{i=0}^N \mathrm{Gamma}(\alpha_i|a, b) \\ p(\sigma^2) &= \textstyle\prod_{i=0}^N \mathrm{Gamma}(\beta|c, d) \end{aligned} \tag{10}$$

The prediction based on the sparse Bayesian learning framework can be expressed as follows:

$$p(t_{N+1}|t) = \int p(t_{N+1}|\omega, \alpha, \sigma^2) p(\omega, \alpha, \sigma^2|t) d\omega d\alpha d\sigma^2 \tag{11}$$

t_{N+1} is the target value of the new observation x_{N+1}.

For a new set of inputs x_*, the output t_* should meet the distribution $p(t_* |t) \sim N(\mu^{\mathrm{T}} \Phi (x^*), \sigma_*^2)$.

$$t_* = \mu^{\mathrm{T}} \Phi(x_*) \tag{12}$$

$$\sigma_*^2 = \sigma_{MP}^2 + \Phi(x_*)^{\mathrm{T}} \sum \Phi(x_*) \tag{13}$$

σ_{MP}^2 is the final variance of noise.

To accomplish the construction of the RVM model, we also need to set the various parameters in **Table 2**.

TABLE 2 | Parameters and functions of RVM.

Setting items	The value set
Max iterations	100
Kernel function	Gauss

Since the output of the RW is not stable, we constructed three RVM models to obtain the final prediction. These three models have equal weights, so the average of the predictions is the final probability of genes associated with hyperthyroidism.

The workflow of the RVM is shown in **Figure 2**.

As we can see from **Figure 2**, we generated three kinds of features from three RW times. Then, the features of the genes were input into the kernel function, and likelihood estimation was done. After obtaining the prior distribution, three RVM models could be built. Finally, we could obtain the final prediction by averaging the three predicted values.

RESULTS

Results of 10-Cross Validation

Since we only know the genes which are reported to be related to hyperthyroidism, we cannot define whether other genes are related to hyperthyroidism. Therefore, we randomly select negative samples from 1,517 genes. As we have 269 positive samples, we randomly select 269 negative samples each time and repeat 10 times. For each negative dataset, we combined it with the positive samples as a whole dataset and implemented 10-cross validation. Therefore, we did 10-cross validation a total of 10 times.

The process of 10-cross validation is as follows: (1) Randomly group negative dataset and positive dataset into 10 groups. (2) Select nine negative groups and nine positive groups to build the RW-RVM model. (3) Input the other negative group and positive group to test the RW-RVM model. (4) Repeat steps 2 and 3 10 times.

The results of performing 10-cross validation 10 times are shown in **Table 3**.

As we can see from **Table 3**, the performance of the RW-RVM model is stable. The average area under the receiver operating characteristic curve (AUC) is 0.90, and the AUPR is 0.87.

Compare With Other Methods

We compared RW-RVM with several other methods, such as random forest (RF), naïve Bayes (NB), and artificial neural network (ANN). The AUC and AUPR curves of these methods are shown in **Figures 3**, **4**.

As shown in **Figures 3**, **4**, RW-RVM performed best among these methods, and RF ranked second. Due to the small sample set, ANN cannot be fully trained, so the AUC and AUPR of the ANN are not satisfactory. The theory of RVM is similar to support vector machine (SVM), which is very suitable for small-sample modeling. Therefore, the performance of the RVM is the best.

TABLE 3 | The AUC and AUPR of 10-cross validation.

Experiment	AUC	AUPR
1	0.89 ± 0.010	0.84 ± 0.012
2	0.92 ± 0.008	0.87 ± 0.013
3	0.89 ± 0.009	0.85 ± 0.011
4	0.88 ± 0.011	0.85 ± 0.012
5	0.90 ± 0.008	0.86 ± 0.013
6	0.91 ± 0.008	0.88 ± 0.012
7	0.91 ± 0.009	0.89 ± 0.017
8	0.89 ± 0.008	0.87 ± 0.011
9	0.90 ± 0.010	0.86 ± 0.009
10	0.90 ± 0.009	0.87 ± 0.011

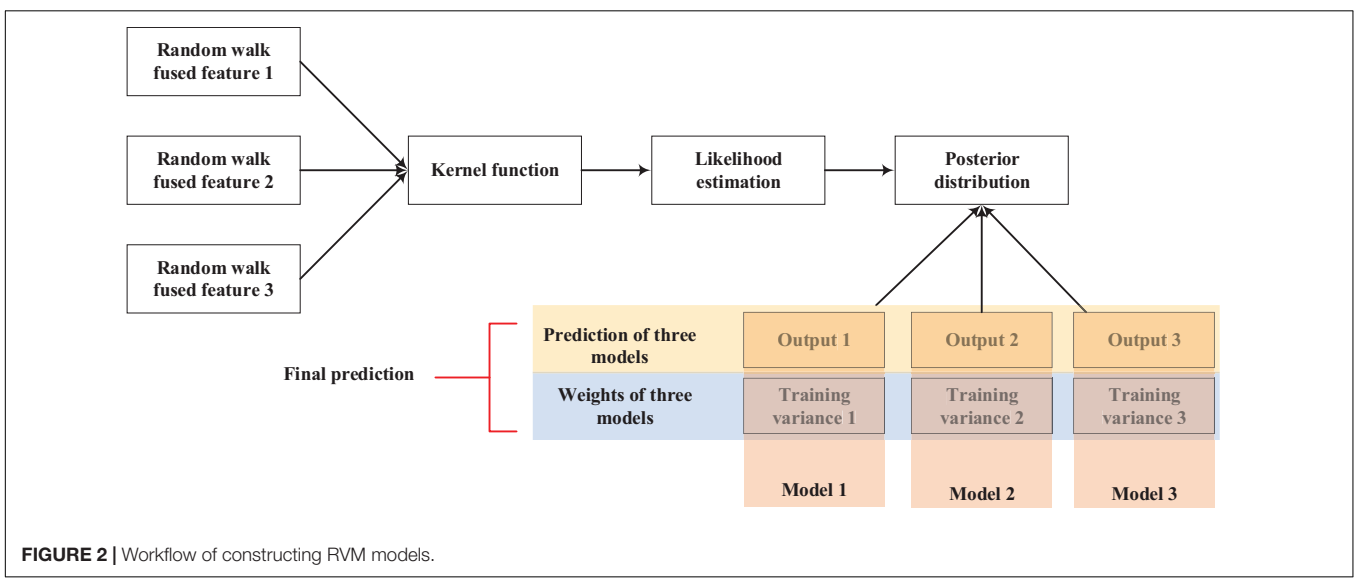


FIGURE 2 | Workflow of constructing RVM models.

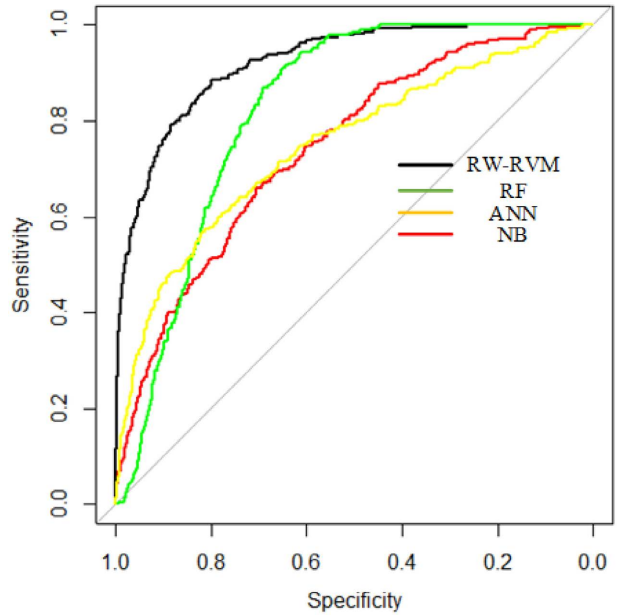


FIGURE 3 | AUC curves of four methods.

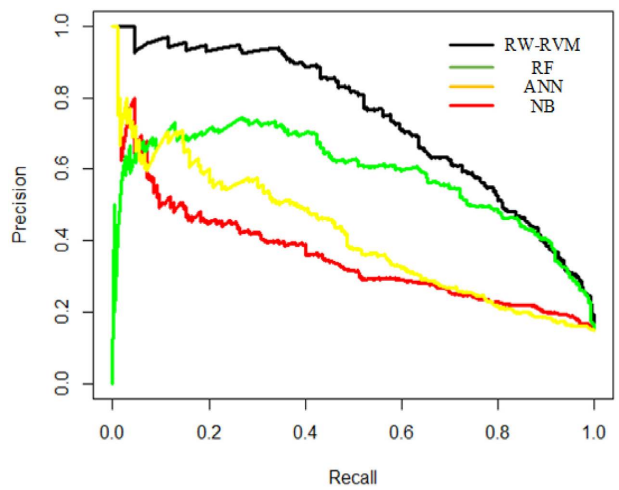


FIGURE 4 | AUPR curves of four methods.

DISCUSSION

Biological experiments have found that hundreds of genes are associated with hyperthyroidism. However, as the cost of sequencing continues to decrease and the amount of data continues to grow, computing methods have already dug out knowledge on a large scale from massive amounts of data. Gene regulation information could help researchers reveal pathogenic mechanisms. Therefore, in this paper, we proposed a novel method, "RW-RVM," to identify hyperthyroidism-related genes. We constructed a gene interaction network based on the known gene interactions and used RW to encode this network to obtain the network topology relationship. In addition, we considered the regulation relationship between genes and non-coding RNAs. After fusing network topology relationship with regulation relationship, we built RVM models to identify hyperthyroidism-related genes. We performed 10-cross validation 10 times to verify the effectiveness of our method. We compared our method with several other methods and found our method performed best in both AUC and AUPR. Seventy-eight novel genes were found to be related to hyperthyroidism (**Supplementary Table 1**).

To verify the accuracy of our results, we did case studies. We found CD34 and MBL2 were associated with hyperthyroidism. Nagura et al. (2001) found that endothelial cells showing CD34 positivity were frequently observed in GD tissue. Filho et al. (2012) reported that MBL2 gene exon 1 variants are related to thyroid disease by sequencing 163 Brazilian patients and 214 healthy controls.

Overall, RW-RVM is a useful tool for discovering hyperthyroidism-related genes in large scale.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

FS, WC, JC, and BX participated in designing the study. FS, WC, XG, JF, ZC, MG, and FW analyzed the data. FS and WC wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Guangzhou Medicine and Health Care Technology Projects (20211A011010) and the Guangzhou Science and Technology Plan Project (202102080170).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.700355/full#supplementary-material

REFERENCES

Asik, M., Gunes, F., Binnetoglu, E., Eroglu, M., Bozkurt, N., Sen, H., et al. (2014). Decrease in TSH levels after lactose restriction in Hashimoto's thyroiditis patients with lactose intolerance. *Endocrine* 46, 279–284. doi: 10.1007/s12020-013-0065-1

Azizi, F., Takyar, M., Madreseh, E., and Amouzegar, A. (2019). Treatment of toxic multinodular goiter: comparison of radioiodine and long-term methimazole treatment. *Thyroid* 29, 625–630. doi: 10.1089/thy.2018.0397

Chu, X., Pan, C.-M., Zhao, S.-X., Liang, J., Gao, G.-Q., Zhang, X.-M., et al. (2011). A genome-wide association study identifies two new risk loci for Graves' disease. *Nat. Genet.* 43:897. doi: 10.1038/ng.898

Ertek, S., and Cicero, A. F. (2013). Hyperthyroidism and cardiovascular complications: a narrative review on the basis of pathophysiology. *Arch. Med. Sci.* 9:944. doi: 10.5114/aoms.2013.38685

Filho, C., Rodrigues, F., Segat, L., Fonseca, A., Araujo, J., Arahata, C., et al. (2012). Association of MBL2 gene exon 1 variants with autoimmune thyroid disease in Brazilian patients. *Int. J. Immunogenetics* 39, 357–361. doi: 10.1111/j.1744-313x.2012.01102.x

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2019). HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* 47, D573–D580. doi: 10.1093/nar/gky1126

Journy, N. M., Bernier, M.-O., Doody, M. M., Alexander, B. H., Linet, M. S., and Kitahara, C. M. (2017). Hyperthyroidism, hypothyroidism, and cause-specific mortality in a large cohort of women. *Thyroid* 27, 1001–1010. doi: 10.1089/thy.2017.0063

Ku, C. F., Lo, C. Y., Chan, W. F., Kung, A. W., and Lam, K. S. (2005). Total thyroidectomy replaces subtotal thyroidectomy as the preferred surgical treatment for Graves' disease. *ANZ J. Surg.* 75, 528–531. doi: 10.1111/j.1445-2197.2005.03441.x

Kuś, A., Szymański, K., Peeters, R. P., Miśkiewicz, P., Porcu, E., Pistis, G., et al. (2015). The association of thyroid peroxidase antibody risk loci with susceptibility to and phenotype of Graves' disease. *Clin. Endocrinol.* 83, 556–562. doi: 10.1111/cen.12640

Liu, J., Fu, J., Xu, Y., and Wang, G. (2017). Antithyroid drug therapy for Graves' disease and implications for recurrence. *Int. J. Endocrinol.* 2017:3813540. doi: 10.1155/2017/3813540

Martin, F., and Deam, D. R. (1996). Hyperthyroidism in elderly hospitalised patients: clinical features and treatment outcomes. *Med. J. Australia* 164, 200–203. doi: 10.5694/j.1326-5377.1996.tb94135.x

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

Moleti, M., Di Mauro, M., Sturniolo, G., Russo, M., and Vermiglio, F. (2019). Hyperthyroidism in the pregnant woman: maternal and fetal aspects. *J. clin. Transl. Endocrinol.* 16:100190. doi: 10.1016/j.jcte.2019.100190

Moon, H.-W., Chung, H.-J., Park, C.-M., Hur, M., and Yun, Y.-M. (2015). Establishment of trimester-specific reference intervals for thyroid hormones in Korean pregnant women. *Ann Lab Med.* 35:198. doi: 10.3343/alm.2015.35.2.198

Nagura, S., Katoh, R., Miyagi, E., Shibuya, M., and Kawaoi, A. (2001). Expression of vascular endothelial growth factor (VEGF) and VEGF receptor-1 (Flt-1) in Graves disease possibly correlated with increased vascular density. *Human Pathol.* 32, 10–17. doi: 10.1053/hupa.2001.21139

Niu, Y.-W., Wang, G.-H., Yan, G.-Y., and Chen, X. (2019). Integrating random walk and binary regression to identify novel miRNA-disease association. *BMC Bioinforma.* 20:59. doi: 10.1186/s12859-019-2640-9

Okosieme, O. E., Taylor, P. N., Evans, C., Thayer, D., Chai, A., Khan, I., et al. (2019). Primary therapy of Graves' disease and cardiovascular morbidity and mortality: a linked-record cohort study. *Lancet Diabetes Endocrinol.* 7, 278–287. doi: 10.1016/s2213-8587(19)30059-2

Peng, J., and Zhao, T. (2020). Reduction in TOM1 expression exacerbates Alzheimer's disease. *Proc. Nat. Acad. Sci.* 117, 3915–3916. doi: 10.1073/pnas.1917589117

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855.

Rathore, N., and de Pablo, J. J. (2002). Monte Carlo simulation of proteins through a random walk in energy space. *J. Chem. Phys.* 116, 7225–7230. doi: 10.1063/1.1463059

Ross, D. S. (2011). Radioiodine therapy for hyperthyroidism. *N. Engl. J. Med.* 364, 542–550. doi: 10.1056/NEJMct1007101

Sharma, A., and Abraham, D. (2020). Vascularity-targeted percutaneous ethanol injection of toxic thyroid adenomas: outcomes of a feasibility study performed in the USA. *Endocr. Prac.* 26, 22–29. doi: 10.4158/ep-2019-0329

Simmonds, M. J. (2013). GWAS in autoimmune thyroid disease: redefining our understanding of pathogenesis. *Nat. Rev. Endocrinol.* 9:277. doi: 10.1038/nrendo.2013.56

Smaniotto, A., Mezalira, D. Z., Zapp, E., Gallardo, H., and Vieira, I. C. (2017). Electrochemical immunosensor based on an azo compound for thyroid-stimulating hormone detection. *Microchem. J.* 133, 510–517. doi: 10.1016/j.microc.2017.04.010

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinformatics* 22:bbaa044. doi: 10.1093/bib/bbaa044

Veeramani, S. K., and Muthusamy, E. (2016). Detection of abnormalities in ultrasound lung image using multi-level RVM classification. *J. Matern. Fetal Neonatal Med.* 29, 1844–1852. doi: 10.3109/14767058.2015.1064888

Wang, J., Long, D., and Chen, Z. (2013). "Early diagnosis of Parkinson's disease patients using RVM-based classification with multi-characteristics," in *Proceedings of the IEEE 3rd International Conference on Information Science and Technology (ICIST)*, (Yangzhou: IEEE), 54–58.

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinformatics* 2020:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021a). Prediction and collection of protein–metabolite interactions. *Brief. Bioinformatics* 2021:bbab014. doi: 10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Shen, Cai, Gan, Feng, Chen, Guo, Wei, Cao and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 03 August 2021
doi: 10.3389/fcell.2021.724172

Identification of Biomarkers for Cervical Cancer Radiotherapy Resistance Based on RNA Sequencing Data

Yue Feng, Zhao Wang, Nan Yang, Sijia Liu, Jiazhuo Yan, Jiayu Song, Shanshan Yang and Yunyan Zhang**

Department of Gynecological Radiotherapy, Harbin Medical University Cancer Hospital, Harbin, China

Cervical cancer as a common gynecological malignancy threatens the health and lives of women. Resistance to radiotherapy is the primary cause of treatment failure and is mainly related to difference in the inherent vulnerability of tumors after radiotherapy. Here, we investigated signature genes associated with poor response to radiotherapy by analyzing an independent cervical cancer dataset from the Gene Expression Omnibus, including pre-irradiation and mid-irradiation information. A total of 316 differentially expressed genes were significantly identified. The correlations between these genes were investigated through the Pearson correlation analysis. Subsequently, random forest model was used in determining cancer-related genes, and all genes were ranked by random forest scoring. The top 30 candidate genes were selected for uncovering their biological functions. Functional enrichment analysis revealed that the biological functions chiefly enriched in tumor immune responses, such as cellular defense response, negative regulation of immune system process, T cell activation, neutrophil activation involved in immune response, regulation of antigen processing and presentation, and peptidyl-tyrosine autophosphorylation. Finally, the top 30 genes were screened and analyzed through literature verification. After validation, 10 genes (KLRK1, LCK, KIF20A, CD247, FASLG, CD163, ZAP70, CD8B, ZNF683, and F10) were to our objective. Overall, the present research confirmed that integrated bioinformatics methods can contribute to the understanding of the molecular mechanisms and potential therapeutic targets underlying radiotherapy resistance in cervical cancer.

Keywords: biomarkers, cervical cancer, radiotherapy resistance, bioinformatic, RNA sequencing data

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Cheng Liang,
Shandong Normal University, China
Yunlong Ma,
Zhejiang University, China

****Correspondence:***
Yunyan Zhang
zhangyunyan@hrbmu.edu.cn
Shanshan Yang
yangshanshan@hrbmu.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *12 June 2021*
Accepted: *14 July 2021*
Published: *03 August 2021*

Citation:
Feng Y, Wang Z, Yang N, Liu S, Yan J, Song J, Yang S and Zhang Y (2021) Identification of Biomarkers for Cervical Cancer Radiotherapy Resistance Based on RNA Sequencing Data. Front. Cell Dev. Biol. 9:724172. doi: 10.3389/fcell.2021.724172

INTRODUCTION

Cervical cancer is the fourth most common malignancy and the fourth leading cause of cancer-related mortality in women globally (Barker et al., 2015; Doja et al., 2020; Sahu and Pattanayak, 2020). Among women between the ages of 20 and 39, cervical cancer is the second leading cause of cancer deaths (Siegel et al., 2020). To date, approximately 80% of patients with cervical cancer need radiotherapy in the process of treatment (Jemal et al., 2011). Radiotherapy resistance, also known as radiosensitivity, remains a significant hurdle to cervical cancer therapeutics. Since 1999,

the cisplatin-based concurrent chemoradiotherapy (CCRT) is the standard of treatment for advanced cervical cancer (Rose et al., 1999). Compared with radiotherapy alone, CCRT increases patients' local control rates and can improve prognosis. However, cisplatin increases the incidence of acute hematological toxicity, which may lead to the interruption of CCRT treatment and poor prognosis in patients with cervical cancer (Green et al., 2001; Dueñas-González et al., 2011; Bazan et al., 2013; Liu et al., 2020). The 5 years survival rate for advanced cervical cancer is still only 5–15% (Chopra et al., 2018). Therefore, exploring the mechanisms underlying radiosensitivity at the molecular level is the essential strategy for increasing long-term survival of cervical cancer.

Resistance to radiotherapy can be intrinsic or acquired, but its specific mechanism is not yet clear. As a promising predictive application, bioinformatics has led to breakthroughs in the field of medical research (Kanehisa and Bork, 2003; Wadlow and Ramaswamy, 2005; Oliver et al., 2015; Tang et al., 2018; Gauthier et al., 2019; Yang and Deng, 2020; Zhang, 2020; Zhang et al., 2020b). For decades, many bioinformatics studies have been performed to investigate radiotherapy resistance-related mechanisms and detect molecular biomarkers (Zhang et al., 2020c,d). However, in radiotherapy, a tumor may gradually adapt to changes in physical and chemical environments and acquire radiotherapy resistance (Baskar et al., 2012; Stapleton et al., 2017). Reliable studies that focus on changes in gene expression during radiotherapy treatment based on bioinformatics are few. Hence, key genes for revealing molecular mechanisms to overcome cervical cancer radiotherapy resistance according to genomic changes should be identified.

In the current study, we acquired a cervical cancer-associated gene expression dataset, comparing cervix cancer tissue between pre-irradiation and mid-irradiation status to investigate the phenomenon of changes in gene expression during radiotherapy. Second, we performed gene expression profile analysis and identification of differentially expressed genes (DEGs). Correlations between these genes were discovered through the Pearson correlation analysis. Subsequently, random forest was used in optimizing risk genes. According to the random forest scoring results, the top 30 genes were selected as candidate genes for verification analysis. Finally, a total of 10 signature genes were verified through comprehensive literature searching as biomarkers for radiotherapy resistance in cervical cancer.

MATERIALS AND METHODS

The whole designation was performed according to the flow chart (**Figure 1**).

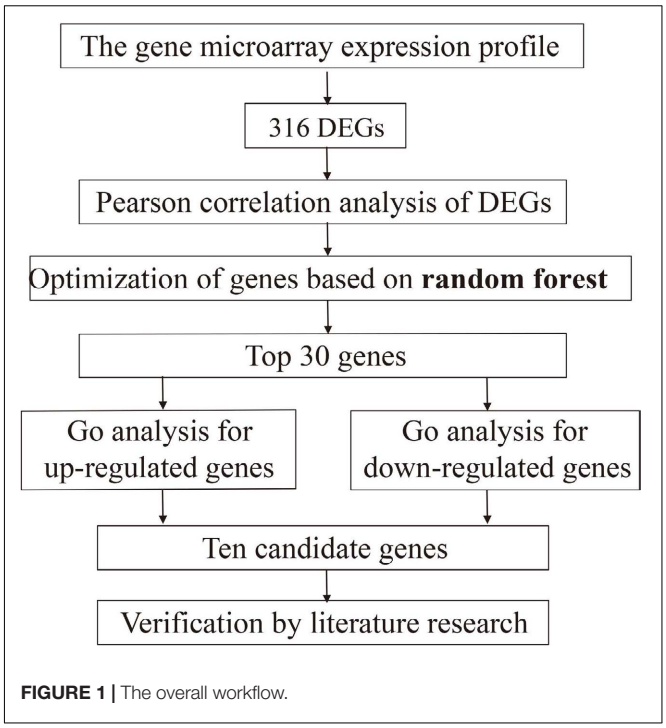


FIGURE 1 | The overall workflow.

Microarray Data Collection and Preprocessing

The gene expression profile was downloaded from the Gene Expression Omnibus (GEO) database[1] which includes a variety of gene expression status of public genomics data (Dueñas-González et al., 2011). We retrieved the keyword "cervical cancer" in this functional database to obtain relevant information about radiotherapy resistance. The inclusion criteria were as follows: (a) the sample was treated with CCRT and without neoadjuvant chemotherapy; (b) number of samples were not less than 50; and (c) sample data were reliable.

The raw GEO dataset of GSE3578 was downloaded as the validation set. It had a total of 156 samples (78 pre-irradiation and 78 mid-irradiation samples). The dataset was based on the platform GPL2895, GE Healthcare/Amer sham Biosciences Code Link Human Whole Genome Bio array. Next, we preprocessed the data and converted each probe ID into a gene symbol. Probes without corresponding or mapping to multiple gene symbols were rejected to prevent ambiguity (Zhang et al., 2018).

Gene Expression Profile Analysis and Identification of Differentially Expressed Genes

The limma package was used in processing the downloaded expression profile and converting and filtering unqualified data to obtain DEGs (Ritchie et al., 2015; Liu et al., 2019b). The original data were chip data, and thus the limma package was used for our analysis. DEGs between the pre-irradiation and mid-irradiation samples were identified, and a P value of <0.05 and $|\log2\text{fold change (FC)}|$ of >0.5 were used as the cutoff thresholds.

Pearson Correlation Analysis

Pearson correlation analysis is a statistical method with powerful functions for accurately analyzing gene or protein co-expression correlations (Zhang et al., 2020a; Tang et al., 2021). In this study,

[1] https://www.ncbi.nlm.nih.gov/gds/

it was used in assessing the relationships among the 316 DEGs obtained. According to the results of the analysis, genes with Pearson correlation coefficient threshold absolute values of >0.8 were selected as candidate genes.

Random Forest

Random forest is a generally acknowledged ensemble classifier for machine learning and can exploit large data repositories for the analysis of risk predictors and their intimate interactions and advancement risk prediction capability (Yang et al., 2021). It was used in evaluating cervical cancer candidate genes as the main objects of our research (Nunn et al., 2012; Wang et al., 2018; Deng et al., 2019b; Ru et al., 2019; Lv et al., 2020; Yang et al., 2020). In the random forest setting, we divided the expression profile data of 103 genes according to the properties of the samples and considered the samples before radiotherapy as "good" and the samples during radiotherapy "poor," respectively. Then, 156 samples from GSE3578 were randomly divided into three groups. One group (52 samples) was used as the training set, whereas the other two groups (104 samples) were as the test sets. All samples were used for random forest scoring (Rigatti, 2017; Wang and Zhou, 2017).

Enrichment Analysis

Gene ontology (GO) enrichment was performed for the annotation of the three types of functions of genes, namely, biological process (BP), cytological component, and molecular function (Peng et al., 2017; Deng et al., 2019a). The gene list containing the top 30 scores in the random forest was regarded as an input file and utilized for GO function enrichment in the R package. On the basis of the reconstructed enrichment analysis, the results of pathways on cervical cancer radiosensitivity were further found.

Screening and Analysis of the Top 30 Genes

Medical literature retrieval is the basic and authoritative verification method in the verification process. After obtaining 30 candidate genes, we extensively reviewed the literature to determine whether the selected genes are indeed signature genes for cervical cancer radiotherapy resistance. We retrieved the name of the genes with the keywords "cervical cancer," "radiotherapy resistance," "irradiation," and "radiosensitivity" in the databases to obtain relevant information about the top 30 genes. Ultimately, we obtained the signature genes for cervical cancer radiotherapy resistance through comprehensive analysis.

RESULTS

Identification of Differentially Expressed Genes

First, to investigate the radiosensitivity in cervical cancer, we processed and converted the raw gene microarray expression dataset GSE3578 associated with the cancer. We extracted 17,097 genes from the obtained microarray dataset. Significance analysis with the limma package indicated that 316 DEGs were present in the samples, of which 237 were obviously down-regulated and 79 were up-regulated.

Pearson Correlation Analysis of DEGs

Pearson correlation analysis was subsequently performed between 316 differential genes for the assessment of the association (Cheng et al., 2019). Genes with a correlation coefficient threshold greater than 0.8 in absolute value were selected on the basis of the results. The analysis results were provided in **Figure 2**, and 103 genes met the threshold setting requirements. All the small nodes were the differential genes we calculated and each node represented a gene with a correlation coefficient greater than 0.8. According to the Pearson correlation coefficient, we obtained the correlations between genes, that is, the greater the connectivity, the greater the node size of the gene displayed.

Optimization of Candidate Genes With Random Forest

Random forest classification is a universally accepted machine learning procedure for regression tree classification and regression and model prediction (Cheng et al., 2018; Sapir-Pichhadze and Kaplan, 2020; Ubels et al., 2020; Zhao et al., 2020). It was first proposed by Breiman in 2001. It takes advantage of binary splits on predictor variables for value outcome forecasts (Steffens et al., 2006). To achieve cervical cancer-related candidate genes, we built a random forest model. We divided the 156 samples into three parts, of which one was the training set (52 samples) and the other was the test set (104 samples). In the test set, the "good group" contained 51 samples, meanwhile, the "poor group" contained 56 samples. Regarding the differential genes as features, the test set was classified by random forest method. When the number of "ntrees" was 1,000, we obtained the most ideal training result with a classification error rate of 9.62%, of which five cases were misclassified in the "good group" and five cases were misclassified in the "poor group." As the results were obtained through the classification of the test set, we regarded this classified procedure noble ideal.

Furthermore, we comprehensively utilized random forest scoring on all samples to further verify the accuracy of our data. The results revealed that the minimum error rate of the classification results was only 1.92% and 78 samples were in the "good group" and one was misclassified. The "poor group" contained 78 samples, including two misclassified samples. After several times of analysis, the error rate of random forest classification results was less than 5%, indicating that the differential genes as features can classify samples well. Therefore, sample classification with the genes had a low error rate, indicating that our classification results were of great significance and had the potential as molecular markers of cervical cancer radiotherapy resistance.

Mean decrease gini (MDG), one of importance measures in random forest scoring, was employed to rank variables and for variable selection. In this study, MDG marked the degree of contributions of genes that could distinguish the characteristics

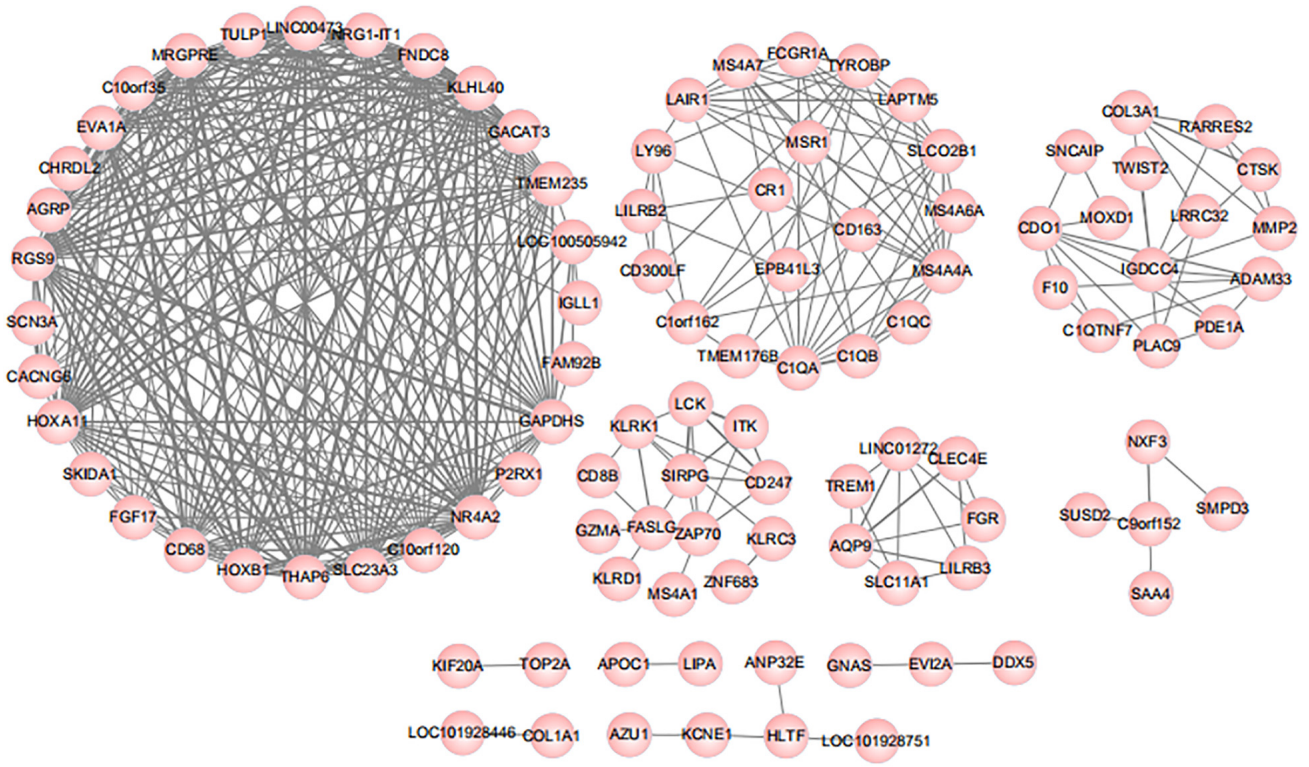


FIGURE 2 | Results of Pearson correlation analysis. According to the Pearson correlation coefficient, we obtained the correlations between genes, that is, the greater the connectivity, the greater the node size of the gene displayed.

of samples before and during radiotherapy. As a result, all genes were sorted in descending order according to MDG, and the top 30 candidate genes for radiotherapy resistance in cervical cancer were selected and shown in **Figure 3**.

Enrichment Functional Analysis of Candidate Genes

To further uncover the molecular biology functions regulated by the genes, the top-ranked 30 genes according to the random forest scoring results were subsequently employed for the functional analysis of BP. As shown in **Figure 4**, the results indicated that biological functions were chiefly enriched in the corresponding tumor immune process pathways with the high fold overexpression, including cellular defense response, negative regulation of immune system process, T cell activation, neutrophil activation involved in immune response, regulation of antigen processing and presentation, and peptidyl-tyrosine autophosphorylation. These immune biological functions are significantly correlated with the examination of radiotherapy resistance. Researches on the radiotherapy resistance of cervical cancer from the perspective of the tumor immune microenvironment is extensive (Barker et al., 2015), showing that tumor radiotherapy has a certain impact on tumor microenvironment (TME), and changes in the TME, such as the generation of tumor-associated macrophages (TAMs) and cancer-associated fibroblasts can in turn affect the sensitivity of radiotherapy (Chu et al., 2014; Zhang et al., 2017; Choo et al., 2018). This shows that our enrichment results are very valuable. Given that the satisfying pathways were enriched, we used the results for the follow-up verification of the signature genes.

Validation of Candidate Genes in Literature Retrieval

Considering the specific functions of the candidate genes in the sensitivity of cervical cancer radiotherapy, we comprehensively conducted a systematic medical literature search based on the top 30 genes. After verification in the literature, the candidate genes ranked 1, 5, 7, 9, 11, 13, 14, 18, 24, and 29 were confirmed with relevant experiments with cancers, demonstrating that these genes were our signature genes for cervical cancer radiotherapy sensitivity. In details, Killer cell lectin like receptor K1 (KLRK1), lymphocyte protein tyrosine kinase (LCK), kinesin family member 20A (KIF20A), CD247 molecule (CD247), Fas ligand gene (FASLG), CD163 molecule (CD163), zeta-chain associated protein 70 (ZAP70), CD8b molecule (CD8B), zinc finger protein 683 (ZNF683), and coagulation factor X (F10) were the top 10 genes. The scores of the 11th and 13th genes were closely associated to cervical cancer radiotherapy resistance, the 7th, 24th, and 29th genes were associated with cervical cancer, the first gene was correlated with cervical intraepithelial neoplasia, the 5th, 14th, and 18th genes were related to radiosensitivity for other cancers, and the 9th gene was associated to tumor

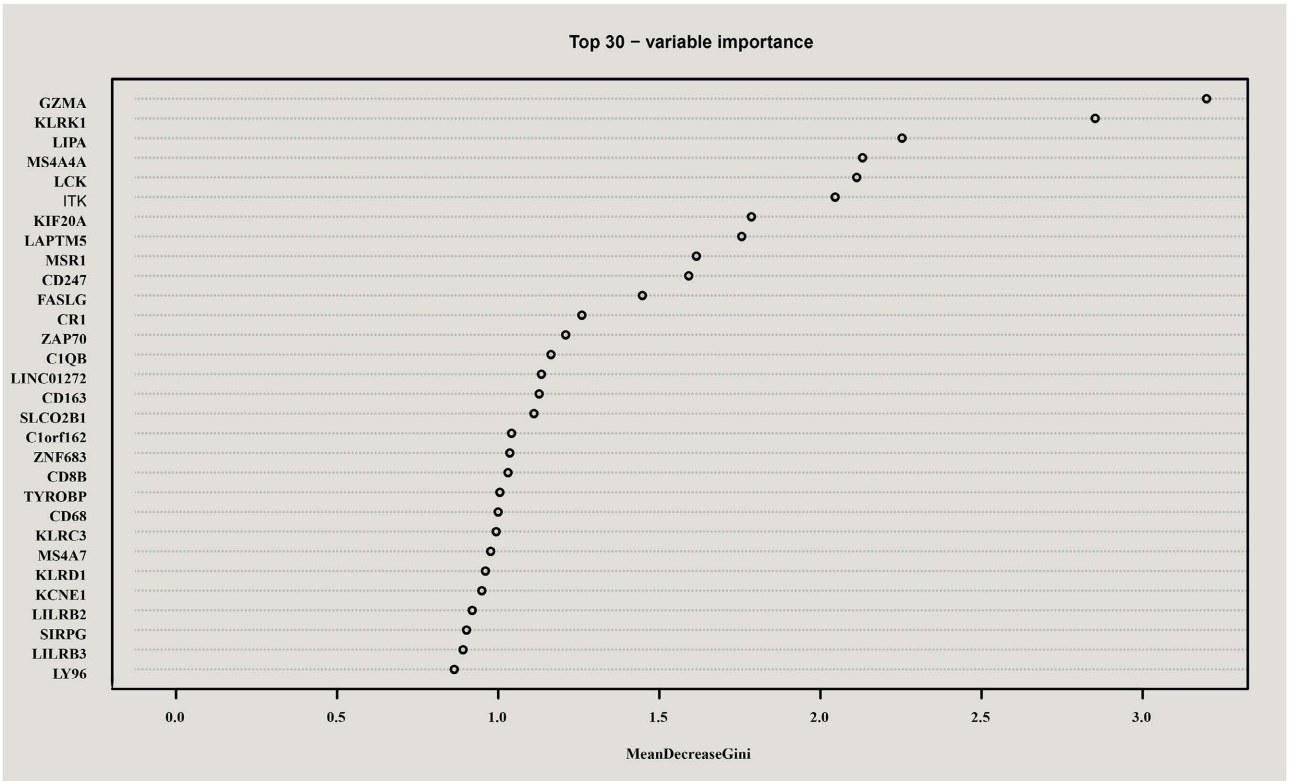


FIGURE 3 | Top 30 genes in random forest scoring results. All genes were sorted in descending order according to Mean decrease gini (MDG), and the top 30 candidate genes for radiotherapy resistance in cervical cancer were selected.

immune microenvironment. In conclusion, this suggests that signature genes are expected to be biomarkers for cervical cancer radiotherapy resistance based on RNA sequencing data.

DISCUSSION

Although radiotherapy for cervical cancer has been widely used, the emergence of radioresistance has deeply affected its treatment effect (Kaidar-Person et al., 2012; Han et al., 2021). Consequently, overcoming the radiotherapy resistance of cervical cancer and finding related accurate biomarkers are essential and urgently needed. Different gene expression characteristics may lead to different types of tumors with varying degrees of sensitivity to radiotherapy (Jiang et al., 2017; Prevo et al., 2017; Zhou et al., 2017). With the development of bioinformatics, gene expression microarrays have been widely used in studying cancer-related genes and considered promising tools for molecular prediction (Cheng et al., 2017; Liu et al., 2019a; Xing et al., 2020).

In the present study, the data from GSE3578 were used in identifying 316 DEGs, which we regarded as the origins. The relationships among these genes were considered, and Pearson correlation analysis was employed. A total of 103 genes with correlation coefficient thresholds >0.8 in absolute value were used for subsequent gene optimization. Identifying cancer-related candidate genes through the random forest method is important. The expression profile data of 103 genes were classified according to the properties of the samples. Then, the 156 samples in GSE3578 were randomly divided into a training set and two test sets, and all the samples were used for random forest scoring. The error rate of sample classification based on these genes was sufficiently small, which can help us to distinguish between normal and disease samples.

Gene ontology term enrichment analysis was conducted using the top 30 genes for the recognition of the roles of DEGs in radiotherapy initiation and progression in cervical cancer. The BP enrichment results were closely related to radiotherapy resistance in cervical carcinoma, suggesting that cervical cancer radiotherapy and immune system are mutually influential. That is, these results showed the specific molecular mechanisms of radiosensitivity in cervical cancer to a certain degree. According to the enriched GO terms, these genes were selected for subsequent verification and analysis. After thoroughly verifying the above genes in the literature, we observed that 10 genes were significant. These genes were selected as our signature genes for cervical cancer radiotherapy sensitivity.

Many studies have been carried out on the 10 candidate genes, and the ranking of the 11th and 13th genes were the most meaningful for our purpose. First, the genes related to cervical cancer or radiotherapy sensitivity were explored. FASLG, a type II membrane protein in the tumor necrosis factor superfamily of death receptors, can combine with and trimerize the death

receptor Fas on cell surfaces and thereby activate authoritative extrinsic pathways for apoptosis (Nagata, 1997; Fulda and Debatin, 2006). Cervical cancer is sensitive to radiotherapy because this gene has different levels of expression. Di et al. stated that the proportion of the pro-apoptotic protein FASLG is up-regulated by the radiotherapy sensitizer diallyl disulfide, which induce the activation of the apoptotic pathway in HeLa cells; they also showed that FASLG plays a direct role in the study of human cervical cancer cell apoptosis induced by radiosensitizers (Fulda and Debatin, 2006). In addition, radiotherapy plays an important role in the adjuvant treatment of glioma cancer; Huang et al. (2019) showed that the radiosensitivity of human glioma cells increases through the FASLG expression-associated signaling pathway. FASLG may be a candidate signature for the mechanism of radiotherapy response in cervical cancer.

Changes in the TME are closely related to the biological effects of radiation and are important factors that affect the sensitivity of tumor cells to radiotherapy (Barker et al., 2015). CD163, one of the immune cell markers for macrophages in the TME, is often used in determining M1 and M2 macrophages (de Vos van Steenwijk et al., 2013). The biological behavior of FASLG is closely correlated with the radiosensitivity of cervical cancer. Lippens et al. (2020) found that patients with reducing CD163 scores showed more pathological complete response, besides, patients with high levels of CD163 cells before treatment that were decreased after therapy had better prognosis. Another study focused on impacts on TME subjected to radiotherapy against head and neck squamous cell carcinoma and pointed out the up-regulation of M2 macrophage-related genes (CD163 and CD206) in irradiated tumors (Moreira et al., 2021). M2 phenotype TAMs are associated with decrease in disease-free survival in advanced breast cancer, and the M2-shifted subgroup is unsuitable for partial breast irradiation (Schnellhardt et al., 2020). Radiation therapy, in addition to killing cancer cells, affects the TME, particularly contributing to tumor blood vessel damage and immunosuppressive pathway activation (Jarosz-Biej et al., 2019). The accumulation of cells can lead to the suppression of radioactive response specifically in TAMs with the M2 phenotype, which facilitates the survival of cancer cells (Fridman et al., 2017). Previous reports and our results have pointed out that CD163 may have a key role in exploring the effectiveness of radiotherapy in cervical cancer radiotherapy.

Four of the 10 candidate genes have been studied in cervical cancer, but their correlations with cervical cancer radiotherapy is unclear (Verhey and Hammond, 2009). KIF20A has a microtubule-dependent force, participating in different cellular processes, such as the formation of the biological behavior of spindle structures and chromosome composition. Overexpressed KIF20A increases the risk of cervical cancer. Through bioinformatics methods and meta-analysis, KIF20A expression was found to significantly vary among 363 normal tissues and cervical cancer tissues and thus considered correlated with cervical cancer development (Wu et al., 2018). He et al. (2020) detected the gene and protein expression levels of KIF20A and found that the up-regulation of KIF20A expression is positively associated with the proliferation and invasion of cervical cancer cells. KIF20A affects the proliferation, migration, and invasion of cervical cancer by participating in a new type of circ_0005576/miR-153-3p/KIF20A axis (Ma et al., 2019), and KIF20A expression is involved in HPV infection, clinical stage, tumor recurrence, pelvic lymph node metastasis, and poor outcomes in early-stage cervical squamous cell cancer (Zhang et al., 2016). Thus, aberrant KIF20A expression is a potential independent prognostic factor for cervical cancer and is correlated with other cancers; for instance, radiation resistance may be enhanced partly by inducing the expression of KIF20A (Xiu et al., 2018). Irradiation-induced IF20A overexpression is associated with liver fibrosis, and KIF20A is associated with advanced pancreatic cancer (Taniuchi et al., 2005). Therefore, KIF20A-targeted treatment may be of great significance and may serve as a potential method for cervical cancer or its radiotherapy.

ZNF683, as a transcription factor, disrupts transcriptional processes in tissue-resident memory T and natural killer T cells, supplying immune protection when the body is re-infected (Van Gisbergen et al., 2012; Mackay et al., 2016). Liu et al. (2018) conducted analysis to clarify the potential mechanism of biological behavior factors that cause cervical cancer. Their results elucidated that the ZNF683 mRNAs are linked to the GO term "adaptive immune response." This result is consistent with ours. In our study, based on the gene list of the top 30 scores in the random forest model, GO function enrichment was constructed using the R package for the prediction of pathways on cervical cancer radiosensitivity and pathways related to immune response. ZNF683 probably mediated the regulation of immune response in malignant cervical cancer. F10 gene may lead to the malignant transformation of the hydatidiform mole and occurrence of gynecological malignancies (Zhang and Pang, 2017). In patients with F10 overexpression, poorly differentiated cervical cancer tissues are higher in proportion compared with well-differentiated cancer tissues, and the negative correlation may be related to the occurrence and development of cervical cancer. KLRK1 encodes NKG2D, serving as an activating receptor expressed by NK cells and T cell subsets (Raulet, 2003). KLRK1 was thought to promote cervical cancer and CIN lesion susceptibility in 195 patients from southern Brazil.

Although not related to cervical cancer, most of the other parts of the genes we obtained through bioinformatics analysis were found to be related to radiosensitivity for other cancers. LCK was verified as a member of the Src family of non-receptor tyrosine kinases (Veillette et al., 1991). In Chinese hamster V79 cells, the expression of LCK gene is a key response to ionizing radiation-induced apoptosis and increases with radiation dose and incubation period. The Src-like tyrosine kinase p56/Lck is involved in the regulation of apoptosis caused by ionizing radiation. Similarly, in radioresistant B-lymphoma cells, p56/Lck promotes radiation-associated cell apoptosis (Waddick et al., 1993; Belka et al., 1999). The B-cell receptor-associated Src-like kinase Btk exhibits this function in B cells (Uckun et al., 1996).

ZAP70 is involved in T cell receptor signal transduction. The ability of T lymphocytes to convert signals to the nucleus is essential to the initiation and preservation of immune response, and the stimulation of T-cell receptors requires ligands to stimulate events in cells organized by ZAP70 and other proteins (Mozaffari et al., 2007). According to our results, ZAP70 plays

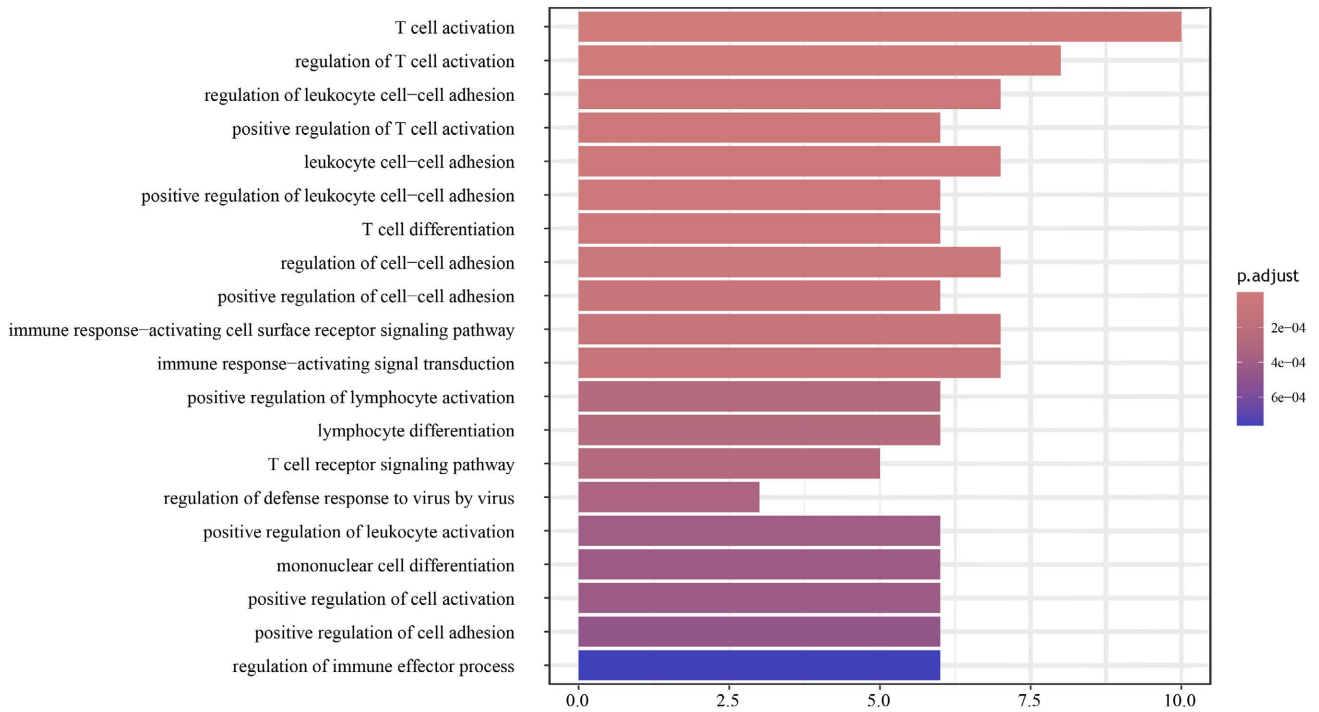


FIGURE 4 | Enrichment analysis results of biological process (BP) was constructed by R package. The biological functions chiefly enriched in tumor immune responses.

an important role in radiotherapy-related immune response pathway. Similarly, a longitudinal study focused on immunologic system effects of adjuvant chemotherapy and radiotherapy in breast cancer showed that regulatory T cells and ZAP70 remain in their low grade forms after adjuvant chemoradiotherapy. In radiotherapy for colorectal cancer, through RNA extraction and following bioinformatic analysis, overexpressed ZAP70 was found to influence prognostic factors, such as tumor size and lymph node metastasis and differentiation; this finding showed that ZAP70 gene may act as a sensitivity biomarker for radiotherapy in colorectal cancer (Huang et al., 2011). Most of our information about PD-1 signaling was obtained through the exploration of activated T cells. In addition, the inhibition of ZAP70 impedes the activation of transcription factors in PD-1 signaling transduction during the stimulation of T cells (Wu et al., 2019).

As for CD247 and CD8B, related reports have been published. Mori et al. (2021) performed univariate experiments and analysis and found that reduction in PD-L1 mRNA/CD8B mRNA in tumor tissues may be a specific prognostic factor for salvage surgery for local recurrence after definitive concurrent radiotherapy and chemotherapy. CD247, also called CD3z-chain or TCR-Z, is a 16 kDa molecule participating in the TCR complex (Germain and Stefanová, 1999). Abnormal CD247 causes the irregular stimulation of T cells upon appointment with the TCR. CD247 down-regulation in TILs and peripheral blood lymphocytes is related to several types of cancers, including gastric carcinoma, head and neck cancer, and ovarian cancer, suggesting that CD247 is a therapeutic target for these cancer types (Ishigami et al., 2002; Ye et al., 2019). However, the relationships of the four genes with cervical cancer are rarely investigated despite that they may play vital role in cervical cancer. Additional tests are needed to verify the mechanism of these genes in the radiosensitivity of cervical cancer.

We employed Pearson correlation analysis to assess the meaningful relationships among the DEGs. Random forest was used in revealing the error rate of classifying samples. The results are satisfying and suitable for distinguishing cervical cancer-related genes from the samples. The results of the top 30 gene enrichment pathways have been verified by medical literature search to confirm their correlation with the sensitivity of cervical cancer radiotherapy, indicating that our results are precise. Notably, the medical literature search further confirmed that the expression levels of the 10 risk genes are associated with cervical cancer radiotherapy response or different types of tumors to a certain degree. Hence, the results suggested that our analytical application is of great value in identifying radiotherapy resistant molecular biomarkers for cervical cancer.

However, our current study has limitations and drawbacks. In this research, we verified the top 30 genes through literature searching because this method is highly comprehensive and can be easily conducted. In subsequent studies, we will perform accurate experimental techniques in pre-irradiation and mid-irradiation cervical cancer cell lines to validate the mRNA and protein expression levels of the 10 signature genes in cervical cancer. According to the results of BP analyses, we concluded

that our candidate genes were mainly abundant in the immune responses of tumors, which is a new perspective for our further study of the mechanisms. We will preserve on valuable researches of cervical cancer radiotherapy resistance to explore the potential therapeutic targets for it.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3578.

AUTHOR CONTRIBUTIONS

YF and YZ project design. NY and ZW project supervision. YF, SL, JY, and JS data generation and analysis. YF and YZ revisions of the manuscript. All authors reviewed the manuscript.

FUNDING

This study was supported by the National Natural Science Foundation (U20A20339).

REFERENCES

Barker, H. E., Paget, J. T., Khan, A. A., and Harrington, K. J. (2015). The tumour microenvironment after radiotherapy: mechanisms of resistance and recurrence. *Nat. Rev. Cancer* 15, 409–425. doi: 10.1038/nrc3958

Baskar, R., Lee, K. A., Yeo, R., and Yeoh, K. W. (2012). Cancer and radiation therapy: current advances and future directions. *Int. J. Med. Sci*. 9, 193–199. doi: 10.7150/ijms.3635

Bazan, J. G., Luxton, G., Kozak, M. M., Anderson, E. M., Hancock, S. L., Kapp, D. S., et al. (2013). Impact of chemotherapy on normal tissue complication probability models of acute hematologic toxicity in patients receiving pelvic intensity modulated radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys*. 87, 983–991. doi: 10.1016/j.ijrobp.2013.09.017

Belka, C., Marini, P., Lepple-Wienhues, A., Budach, W., Jekle, A., Los, M., et al. (1999). The tyrosine kinase lck is required for CD95-independent caspase-8 activation and apoptosis in response to ionizing radiation. *Oncogene* 18, 4983–4992. doi: 10.1038/sj.onc.1202878

Cheng, J., Lu, X., Wang, J., Zhang, H., Duan, P., and Li, C. (2017). Interactome analysis of gene expression profiles of cervical cancer reveals dysregulated mitotic gene clusters. *Am. J. Transl. Res*. 9, 3048–3059.

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief. Bioinform*. 20, 203–209. doi: 10.1093/bib/bbx103

Choo, Y. W., Kang, M., Kim, H. Y., Han, J., Kang, S., Lee, J. R., et al. (2018). M1 macrophage-derived nanovesicles potentiate the anticancer efficacy of immune checkpoint inhibitors. *ACS Nano* 12, 8977–8993. doi: 10.1021/acsnano.8b02446

Chopra, S., Gupta, M., Mathew, A., Mahantshetty, U., Engineer, R., Lavanya, G., et al. (2018). Locally advanced cervical cancer: a study of 5-year outcomes. *Indian J. Cancer* 55, 45–49. doi: 10.4103/ijc.IJC_428_17

Chu, T. Y., Yang, J. T., Huang, T. H., and Liu, H. W. (2014). Crosstalk with cancer-associated fibroblasts increases the growth and radiation survival of cervical cancer cells. *Radiat. Res*. 181, 540–547. doi: 10.1667/rr13583.1

de Vos van Steenwijk, P. J., Ramwadhdoebe, T. H., Goedemans, R., Doorduijn, E. M., Van Ham, J. J., Gorter, A., et al. (2013). Tumor-infiltrating CD14-positive myeloid cells and CD8-positive T-cells prolong survival in patients with cervical carcinoma. *Int. J. Cancer* 133, 2884–2894. doi: 10.1002/ijc.28309

Deng, L., Wang, J., and Zhang, J. (2019a). Predicting gene ontology function of human MicroRNAs by integrating multiple networks. *Front. Genet*. 10:3. doi: 10.3389/fgene.2019.00003

Deng, L., Yang, W., and Liu, H. (2019b). PredPRBA: prediction of protein-RNA binding affinity using gradient boosted regression trees. *Front. Genet*. 10:637. doi: 10.3389/fgene.2019.00637

Doja, M. N., Kaur, I., and Ahmad, T. (2020). Current state of the art for survival prediction in cancer using data mining techniques. *Curr. Bioinform*. 15, 174–186. doi: 10.2174/1574893614666190902152142

Dueñas-González, A., Zarbá, J. J., Patel, F., Alcedo, J. C., Beslija, S., Casanova, L., et al. (2011). Phase III, open-label, randomized study comparing concurrent gemcitabine plus cisplatin and radiation followed by adjuvant gemcitabine and cisplatin versus concurrent cisplatin and radiation in patients with stage IIB to IVA carcinoma of the cervix. *J. Clin. Oncol*. 29, 1678–1685. doi: 10.1200/jco.2009.25.9663

Fridman, W. H., Zitvogel, L., Sautès-Fridman, C., and Kroemer, G. (2017). The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol*. 14, 717–734. doi: 10.1038/nrclinonc.2017.101

Fulda, S., and Debatin, K. M. (2006). Extrinsic versus intrinsic apoptosis pathways in anticancer chemotherapy. *Oncogene* 25, 4798–4811. doi: 10.1038/sj.onc.1209608

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). A brief history of bioinformatics. *Brief. Bioinform*. 20, 1981–1996. doi: 10.1093/bib/bby063

Germain, R. N., and Stefanová, I. (1999). The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation. *Annu. Rev. Immunol*. 17, 467–522. doi: 10.1146/annurev.immunol.17.1.467

Green, J. A., Kirwan, J. M., Tierney, J. F., Symonds, P., Fresco, L., Collingwood, M., et al. (2001). Survival and recurrence after concomitant chemotherapy and radiotherapy for cancer of the uterine cervix: a systematic review and meta-analysis. *Lancet* 358, 781–786. doi: 10.1016/s0140-6736(01)05965-7

Han, X., Kong, Q., Liu, C., Cheng, L., and Han, J. (2021). SubtypeDrug: a software package for prioritization of candidate cancer subtype-specific drugs. *Bioinformatics*. doi: 10.1093/bioinformatics/btab011 Online ahead of print,

He, Q., Meng, J., Liu, S., Zeng, Q., Zhu, Q., Wei, Z., et al. (2020). Long non-coding RNA UCA1 upregulates KIF20A expression to promote cell proliferation and invasion via sponging miR-204 in cervical cancer. *Cell Cycle* 19, 2486–2495. doi: 10.1080/15384101.2020.1807666

Huang, C., Chen, D., Zhu, H., Lv, S., Li, Q., and Li, G. (2019). LITAF enhances radiosensitivity of human glioma cells via the FoxO1 pathway. *Cell Mol. Neurobiol*. 39, 871–882. doi: 10.1007/s10571-019-00686-4

Huang, M. Y., Wang, J. Y., Chang, H. J., Kuo, C. W., Tok, T. S., and Lin, S. R. (2011). CDC25A, VAV1, TP73, BRCA1 and ZAP70 gene overexpression correlates with radiation response in colorectal cancer. *Oncol. Rep*. 25, 1297–1306. doi: 10.3892/or.2011.1193

Ishigami, S., Natsugoe, S., Tokuda, K., Nakajo, A., Higashi, H., Iwashige, H., et al. (2002). CD3-zetachain expression of intratumoral lymphocytes is closely related to survival in gastric carcinoma patients. *Cancer* 94, 1437–1442. doi: 10.1002/cncr.10346

Jarosz-Biej, M., Smolarczyk, R., Cichoń, T., and Kułach, N. (2019). Tumor microenvironment as a "Game Changer" in cancer radiotherapy. *Int. J. Mol. Sci*. 20:3212. doi: 10.3390/ijms20133212

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin*. 61, 69–90. doi: 10.3322/caac.20107

Jiang, L. P., He, C. Y., and Zhu, Z. T. (2017). Role of microRNA-21 in radiosensitivity in non-small cell lung cancer cells by targeting PDCD4 gene. *Oncotarget* 8, 23675–23689. doi: 10.18632/oncotarget.15644

Kaidar-Person, O., Bortnyak-Abdah, R., Amit, A., Berniger, A., Ben-Yosef, R., and Kuten, A. (2012). Current principles for radiotherapy in cervical cancer. *Med. Oncol*. 29, 2919–2922. doi: 10.1007/s12032-012-0170-7

Kanehisa, M., and Bork, P. (2003). Bioinformatics in the post-sequence era. *Nat. Genet*. 33(Suppl.), 305–310. doi: 10.1038/ng1109

Lippens, L., Van Bockstal, M., De Jaeghere, E. A., Tummers, P., Makar, A., De Geyter, S., et al. (2020). Immunologic impact of chemoradiation in cervical

cancer and how immune cell infiltration could lead toward personalized treatment. *Int. J. Cancer* 147, 554–564. doi: 10.1002/ijc.32893

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., and Deng, L. (2020). DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res*. 48, D871–D881.

Liu, J., Nie, S., Gao, M., Jiang, Y., Wan, Y., Ma, X., et al. (2019a). Identification of EPHX2 and RMI2 as two novel key genes in cervical squamous cell carcinoma by an integrated bioinformatic analysis. *J. Cell Physiol*. 234, 21260–21273. doi: 10.1002/jcp.28731

Liu, J., Zhou, S., Li, S., Jiang, Y., Wan, Y., Ma, X., et al. (2019b). Eleven genes associated with progression and prognosis of endometrial cancer (EC) identified by comprehensive bioinformatics analysis. *Cancer Cell Int*. 19:136. doi: 10.1186/s12935-019-0859-1

Liu, Q., Lu, C., Dai, W., Li, K., Xu, J., Huang, Y., et al. (2018). Association of biobehavioral factors with non-coding RNAs in cervical cancer. *Biosci. Trends* 12, 24–31. doi: 10.5582/bst.2017.01325

Lv, Z. B., Zhang, J., Ding, H., and Zou, Q. (2020). RF-PseU: a random forest predictor for RNA pseudouridine sites. *Front. Bioeng. Biotechnol*. 8:10. doi: 10.3389/fbioe.2020.00134

Ma, H., Tian, T., Liu, X., Xia, M., Chen, C., Mai, L., et al. (2019). Upregulated circ_0005576 facilitates cervical cancer progression via the miR-153/KIF20A axis. *Biomed. Pharmacother*. 118:109311. doi: 10.1016/j.biopha.2019.109311

Mackay, L. K., Minnich, M., Kragten, N. A., Liao, Y., Nota, B., Seillet, C., et al. (2016). Hobit and Blimp1 instruct a universal transcriptional program of tissue residency in lymphocytes. *Science* 352, 459–463. doi: 10.1126/science.aad2035

Moreira, D., Sampath, S., Won, H., White, S. V., Su, Y. L., Alcantara, M., et al. (2021). Myeloid cell-targeted STAT3 inhibition sensitizes head and neck cancers to radiotherapy and T cell-mediated immunity. *J. Clin. Invest*. 131:e137001. doi: 10.1172/jci137001

Mori, T., Kumagai, K., Nasu, K., Yoshizawa, T., Kuwano, K., Hamada, Y., et al. (2021). Clonal expansion of tumor-infiltrating T cells and analysis of the tumor microenvironment within esophageal squamous cell carcinoma relapsed after definitive chemoradiation therapy. *Int. J. Mol. Sci*. 22:1098. doi: 10.3390/ijms22031098

Mozaffari, F., Lindemalm, C., Choudhury, A., Granstam-Björneklett, H., Helander, I., Lekander, M., et al. (2007). NK-cell and T-cell functions in patients with breast cancer: effects of surgery and adjuvant chemo- and radiotherapy. *Br. J. Cancer* 97, 105–111. doi: 10.1038/sj.bjc.6603840

Nagata, S. (1997). Apoptosis by death factor. *Cell* 88, 355–365. doi: 10.1016/s0092-8674(00)81874-7

Nunn, M. E., Fan, J., Su, X., Levine, R. A., Lee, H. J., and Mcguire, M. K. (2012). Development of prognostic indicators using classification and regression trees for survival. *Periodontology 2000* 58, 134–142. doi: 10.1111/j.1600-0757.2011.00421.x

Oliver, G. R., Hart, S. N., and Klee, E. W. (2015). Bioinformatics for clinical next generation sequencing. *Clin. Chem*. 61, 124–135. doi: 10.1373/clinchem.2014.224360

Peng, J., Wang, H., Lu, J., Hui, W., Wang, Y., and Shang, X. (2017). Identifying term relations cross different gene ontology categories. *BMC Bioinform*. 18:573. doi: 10.1186/s12859-017-1959-3

Prevo, R., Tiwana, G. S., Maughan, T. S., Buffa, F. M., Mckenna, W. G., and Higgins, G. S. (2017). Depletion of signal recognition particle 72kDa increases radiosensitivity. *Cancer Biol. Ther*. 18, 425–432. doi: 10.1080/15384047.2017.1323587

Raulet, D. H. (2003). Roles of the NKG2D immunoreceptor and its ligands. *Nat. Rev. Immunol*. 3, 781–790. doi: 10.1038/nri1199

Rigatti, S. J. (2017). Random forest. *J. Insur. Med*. 47, 31–39. doi: 10.17849/insm-47-01-31-39.1

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 43:e47. doi: 10.1093/nar/gkv007

Rose, P. G., Bundy, B. N., Watkins, E. B., Thigpen, J. T., Deppe, G., Maiman, M. A., et al. (1999). Concurrent cisplatin-based radiotherapy and chemotherapy for locally advanced cervical cancer. *N. Engl. J. Med*. 340, 1144–1153. doi: 10.1056/nejm199904153401502

Ru, X. Q., Li, L. H., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteome Res*. 18, 2931–2939. doi: 10.1021/acs.jproteome.9b00250

Sahu, R., and Pattanayak, S. P. (2020). Strategic developments & future perspective on gene therapy for breast cancer: role of mTOR and Brk/PTK6 as molecular targets. *Curr. Gene Ther*. 20, 237–258. doi: 10.2174/1566523220999200731002408

Sapir-Pichhadze, R., and Kaplan, B. (2020). Seeing the forest for the trees: random forest models for predicting survival in kidney transplant recipients. *Transplantation* 104, 905–906. doi: 10.1097/tp.0000000000002923

Schnellhardt, S., Erber, R., Büttner-Herold, M., Rosahl, M. C., Ott, O. J., Strnad, V., et al. (2020). Accelerated partial breast irradiation: macrophage polarisation shift classification identifies high-risk tumours in early hormone receptor-positive breast cancer. *Cancers* 12:446. doi: 10.3390/cancers12020446

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin*. 70, 7–30. doi: 10.3322/caac.21590

Stapleton, S., Jaffray, D., and Milosevic, M. (2017). Radiation effects on the tumor microenvironment: implications for nanomedicine delivery. *Adv. Drug Deliv. Rev*. 109, 119–130. doi: 10.1016/j.addr.2016.05.021

Steffens, M., Lamina, C., Illig, T., Bettecken, T., Vogler, R., Entz, P., et al. (2006). SNP-based analysis of genetic substructure in the German population. *Hum. Hered*. 62, 20–29. doi: 10.1159/000095850

Tang, Q., Nie, F., Kang, J., and Chen, W. (2021). mRNALocater: Enhance the prediction accuracy of eukaryotic mRNA subcellular localization by using model fusion strategy. *Mol. Ther*. doi: 10.1016/j.ymthe.2021.04.004 Online ahead of print,

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Taniuchi, K., Nakagawa, H., Nakamura, T., Eguchi, H., Ohigashi, H., Ishikawa, O., et al. (2005). Down-regulation of RAB6KIFL/KIF20A, a kinesin involved with membrane trafficking of discs large homologue 5, can attenuate growth of pancreatic cancer cell. *Cancer Res*. 65, 105–112.

Ubels, J., Schaefers, T., Punt, C., Guchelaar, H. J., and De Ridder, J. (2020). RAINFOREST: a random forest approach to predict treatment benefit in data from (failed) clinical drug trials. *Bioinformatics* 36, i601–i609. doi: 10.1093/bioinformatics/btaa799

Uckun, F. M., Waddick, K. G., Mahajan, S., Jun, X., Takata, M., Bolen, J., et al. (1996). BTK as a mediator of radiation-induced apoptosis in DT-40 lymphoma B cells. *Science* 273, 1096–1100. doi: 10.1126/science.273.5278.1096

Van Gisbergen, K. P., Kragten, N. A., Hertoghs, K. M., Wensveen, F. M., Jonjic, S., Hamann, J., et al. (2012). Mouse hobit is a homolog of the transcriptional repressor Blimp-1 that regulates NKT cell effector differentiation. *Nat. Immunol*. 13, 864–871. doi: 10.1038/ni.2393

Veillette, A., Abraham, N., Caron, L., and Davidson, D. (1991). The lymphocyte-specific tyrosine protein kinase p56lck. *Semin. Immunol*. 3, 143–152.

Verhey, K. J., and Hammond, J. W. (2009). Traffic control: regulation of kinesin motors. *Nat. Rev. Mol. Cell Biol*. 10, 765–777. doi: 10.1038/nrm2782

Waddick, K. G., Chae, H. P., Tuel-Ahlgren, L., Jarvis, L. J., Dibirdik, I., Myers, D. E., et al. (1993). Engagement of the CD19 receptor on human B-lineage leukemia cells activates LCK tyrosine kinase and facilitates radiation-induced apoptosis. *Radiat. Res*. 136, 313–319. doi: 10.2307/3578542

Wadlow, R., and Ramaswamy, S. (2005). DNA microarrays in clinical cancer research. *Curr. Mol. Med*. 5, 111–120. doi: 10.2174/1566524053152834

Wang, H., and Zhou, L. (2017). Random survival forest with space extensions for censored data. *Artif. Intell. Med*. 79, 52–61. doi: 10.1016/j.artmed.2017.06.005

Wang, H., Liu, C., and Deng, L. (2018). Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci. Rep*. 8:14285.

Wu, D. M., Shi, J., Liu, T., Deng, S. H., Han, R., and Xu, Y. (2018). Integrated analysis reveals down-regulation of SPARCL1 is correlated with cervical cancer development and progression. *Cancer Biomark*. 21, 355–365. doi: 10.3233/cbm-170501

Wu, X., Gu, Z., Chen, Y., Chen, B., Chen, W., Weng, L., et al. (2019). Application of PD-1 Blockade in Cancer Immunotherapy. *Comput. Struct. Biotechnol. J*. 17, 661–674. doi: 10.1016/j.csbj.2019.03.006

Xing, L., Mi, W., Zhang, Y., Tian, S., Zhang, Y., Qi, R., et al. (2020). The identification of six risk genes for ovarian cancer platinum response based on global network algorithm and verification analysis. *J. Cell Mol. Med*. 24, 9839–9852. doi: 10.1111/jcmm.15567

Xiu, G., Sui, X., Wang, Y., and Zhang, Z. (2018). FOXM1 regulates radiosensitivity of lung cancer cell partly by upregulating KIF20A. *Eur. J. Pharmacol.* 833, 79–85. doi: 10.1016/j.ejphar.2018.04.021

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 75, 140–149. doi: 10.1016/j.inffus.2021.02.015

Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., et al. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci .Rep*. 10:5245. doi: 10.1038/s41598-020-62133-5

Yang, W., and Deng, L. (2020). PreDBA: a heterogeneous ensemble approach for predicting protein-DNA binding affinity. *Sci. Rep*. 10:1278.

Ye, W., Zhou, Y., Xu, B., Zhu, D., Rui, X., Xu, M., et al. (2019). CD247 expression is associated with differentiation and classification in ovarian cancer. *Medicine* 98:e18407. doi: 10.1097/md.0000000000018407

Zhang, H., Yue, J., Jiang, Z., Zhou, R., Xie, R., Xu, Y., et al. (2017). CAF-secreted CXCL1 conferred radioresistance by regulating DNA damage response in a ROS-dependent manner in esophageal squamous cell carcinoma. *Cell Death Dis*. 8:e2790. doi: 10.1038/cddis.2017.180

Zhang, L., Jiang, X., Li, Y., Fan, Q., Li, H., Jin, L., et al. (2020a). Clinical correlation of Wnt2 and COL8A1 with colon adenocarcinoma prognosis. *Front. Oncol.* 10:1504. doi: 10.3389/fonc.2020.01504

Zhang, M. Z., and Pang, Z. J. (2017). [F10 expressions in cervical cancer tissues]. *Nan Fang Yi Ke Da Xue Xue Bao*. 37, 792–796. doi: 10.3969/j.issn.1673-4254.2017.06.13

Zhang, S., He, X., Zhang, R., and Deng, W. (2020b). LncR2metasta: a manually curated database for experimentally supported lncRNAs during various cancer metastatic events. *Brief. Bioinform*. 20:bbaa178.

Zhang, W., He, W., Shi, Y., Gu, H., Li, M., Liu, Z., et al. (2016). High expression of KIF20A is associated with poor overall survival and tumor progression in early-stage cervical squamous cell carcinoma. *PLoS One* 11:e0167449. doi: 10.1371/journal.pone.0167449

Zhang, X., Feng, H., Li, Z., Li, D., Liu, S., Huang, H., et al. (2018). Application of weighted gene co-expression network analysis to identify key modules and hub genes in oral squamous cell carcinoma tumorigenesis. *Oncol. Targets Ther*. 11, 6001–6021. doi: 10.2147/ott.S171791

Zhang, Y. (2020). Artificial intelligence for bioinformatics and biomedicine. *Curr. Bioinform*. 15, 801–802. doi: 10.2174/1574893615082012210923330

Zhang, Z. M., Tan, J. X., Wang, F., Dao, F. Y., Zhang, Z. Y., and Lin, H. (2020c). Early diagnosis of hepatocellular carcinoma using machine learning method. *Front. Bioeng. Biotechnol.* 8:254. doi: 10.3389/fbioe.2020.00254

Zhang, Z. M., Wang, J. S., Zulfiqar, H., Lv, H., Dao, F. Y., and Lin, H. (2020d). Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with machine-learning method. *Front. Cell Dev. Biol.* 8:582864. doi: 10.3389/fcell.2020.582864

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhou, Y., Wang, C., Liu, X., Wu, C., and Yin, H. (2017). Long non-coding RNA HOTAIR enhances radioresistance in MDA-MB231 breast cancer cells. *Oncol. Lett*. 13, 1143–1148. doi: 10.3892/ol.2017.5587

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Feng, Wang, Yang, Liu, Yan, Song, Yang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 06 August 2021
doi: 10.3389/fcell.2021.684322

Comprehensive Analysis of Glycolysis-Related Genes for Prognosis, Immune Features, and Candidate Drug Development in Colon Cancer

Zhongqi Cui[1†], Guifeng Sun[1†], Ramesh Bhandari[1,2†], Jiayi Lu[1], Mengmei Zhang[1], Rajeev Bhandari[3], Fenyong Sun[1], Zhongchen Liu[4] and Shasha Zhao[1]**

[1] Department of Clinical Laboratory, Shanghai Tenth People's Hospital, Tongji University, Shanghai, China, [2] Department of Pathology, Universal College of Medical Sciences, Bhairahawa, Nepal, [3] Department of General Surgery, Universal College of Medical Sciences, Bhairahawa, Nepal, [4] Department of Gastrointestinal Surgery, Shanghai Tenth People's Hospital Affiliated to Tongji University, Shanghai, China

OPEN ACCESS

Edited by:
Chuan-Xing Li,
Karolinska Institutet, Sweden

Reviewed by:
Shunrong Ji,
Fudan University, China
Mariateresa Giuliano,
Università degli Studi della Campania "Luigi Vanvitelli", Italy

***Correspondence:**
Shasha Zhao
zhaosha2012@126.com
Zhongchen Liu
liuzhongchen_10th@163.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 25 March 2021
Accepted: 12 July 2021
Published: 06 August 2021

Citation:
Cui Z, Sun G, Bhandari R, Lu J, Zhang M, Bhandari R, Sun F, Liu Z and Zhao S (2021) Comprehensive Analysis of Glycolysis-Related Genes for Prognosis, Immune Features, and Candidate Drug Development in Colon Cancer. Front. Cell Dev. Biol. 9:684322. doi: 10.3389/fcell.2021.684322

The dysregulated expression of glycolysis-related genes (GRGs) is closely related to the occurrence of diverse tumors and regarded as a novel target of tumor therapy. However, the role of GRGs in colon cancer is unclear. We obtained 226 differential GRGs (DE-GRGs) from The Cancer Genome Atlas (TCGA) database. Cox regression analysis was used to construct a DE-GRG prognostic model, including P4HA1, PMM2, PGM2, PPARGC1A, PPP2CB, STC2, ENO3, and CHPF2. The model could accurately predict the overall survival rate of TCGA and GSE17536 patient cohorts. The risk score of the model was closely related to a variety of clinical traits and was an independent risk factor for prognosis. Enrichment analysis revealed the activation of a variety of glycolysis metabolism and immune-related signaling pathways in the high-risk group. High-risk patients displayed low expression of CD4+ memory resting T cells and resting dendritic cells and high expression of macrophages M0 compared with the expression levels in the low-risk patients. Furthermore, patients in the high-risk group had a higher tumor mutation load and tumor stem cell index and were less sensitive to a variety of chemotherapeutic drugs. Quantitative reverse transcription polymerase chain reaction and immunohistochemistry analyses validated the expression of eight GRGs in 43 paired clinical samples. This is the first multi-omics study on the GRGs of colon cancer. The establishment of the risk model may benefit the prognosis and drug treatment of patients.

Keywords: glycolysis, colon cancer, prognosis model, immune microenvironment, tumor mutation burden, chemotherapy response

INTRODUCTION

Colon cancer is the most prevalent gastrointestinal malignancy. It is the third-ranked malignancy-related death globally (Arnold et al., 2020). According to the National Cancer Institute and GLOBOCAN data, over one million new cases of colorectal cancer (CRC) are diagnosed annually, and the estimated death rate is 33%. Colon cancer is a multifactorial disease, and its causes could

include genetic predisposition, lifestyle changes, and diet (Pilleron et al., 2019; Vuik et al., 2019). Despite recent advancements in medicine, diagnostics tools, and surgical techniques, the annual incidence and mortality rates of colon cancer continue to increase appreciably, and the 5-year survival rate is less than 15% (Kalyan et al., 2018; Ladabaum et al., 2020). Therefore, it is crucial to discover new biomarkers and therapeutic drug targets for the rapid diagnosis and treatment of colon cancer.

Increasing evidence on cellular metabolism has revealed the alteration in energy metabolism from oxidative phosphorylation (OXPHOS) to aerobic glycolysis that occurs in response to the increased bioenergetic demands of rapidly proliferating and invading cancerous cells. This altered cellular energy metabolism is an important hallmark for cancer development (Cairns et al., 2011; Hanahan and Weinberg, 2011). Unlike normal cells, tumor cells also derive the energy from glucose glycolysis, which is termed as the Warburg effect (Gatenby and Gillies, 2004; Vander Heiden et al., 2009). Lactate produced by the Warburg effect creates a suitable acidic tumor microenvironment (TME) that facilities tumor progression, invasion, and immune response (Corbet and Feron, 2017). Therefore, targeting the glycolysis pathway could be beneficial for tumor prognosis and effective cancer treatment (DeBerardinis et al., 2008; Martinez-Outschoorn et al., 2017).

Glycolysis can alter the biological function of immune cells within the TME and can help cancer cells escape the host body defense mechanism, facilitating cancer progression (Gill et al., 2016). However, only a few comprehensive analyses have explored the role and relationship of tumor glycolysis and immune cell infiltration in colon cancer development.

In the current study, we systematically analyzed the glycolysis-related genomic expression profile in colon adenocarcinoma (COAD) samples from The Cancer Genome Atlas (TCGA) database. The 226 differential glycolysis-related genes (DE-GRGs) that were identified included 17 survival-related GRGs. A risk model incorporating eight GRGs (P4HA1, STC2, PPARGC1A, PPP2CB, ENO3, CHPF2, PMM2, and PGM2) was constructed. The model displayed a high predictive value for the overall survival (OS) rate of colon cancer in both the training and validation groups. Furthermore, the risk model was closely related to a variety of clinical risk factors and could be used as an independent prognostic index for patients with colon cancer. This model may also reflect the dysregulation of immune cell infiltration, the change of gene copy number, and the increase of tumor mutation load and stem cell index. The findings indicate the utility of the model as a powerful tool for the prognosis and treatment of colon cancer.

MATERIALS AND METHODS

Data Collection and Acquisition of Glycolysis-Related Genes

The expressed gene RNA-sequencing dataset (fragments per kilobase of transcript per million [FPKM] value) and corresponding TCGA clinical information of 473 COAD and 71 non-tumor tissue samples were extracted[1]. A total of 298 GRGs, of which 226 were DE-GRGs, were identified by gene set enrichment analysis (GSEA) of glycolysis-related pathways (**Supplementary Figure 1**).

Identification of Differential Glycolysis-Related Gene and Function Analyses

Differential glycolysis-related genes in TCGA COAD tumor tissues and adjacent non-tumor tissues with a false discovery rate (FDR) < 0.05 were identified using the limma R package. The 226 DE-GRGs identified were displayed by Venn diagram (**Supplementary Table 4**). Biological process (BP) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were performed by the clusterProfiler R package (Liao et al., 2019) to identify the functionally enriched genes and classify the gene clusters. A q-value < 0.05 was considered statistically significant.

Construction and Visualization of Protein–Protein Interaction Network Module

The PPI of the 226 DE-GRGs was analyzed in the String online database[2] as previously described (Szklarczyk et al., 2019). The network was visualized by Cytoscape 3.7.2 software[3]. The key modules in the PPI network were filtered according to the score and the criterion of node counts >5 (Bader and Hogue, 2003) using the MCODE plug-in in Cytoscape. A p-value ≤ 0.05 denoted a significant difference.

Construction of Prognostic Model

Key GRGs identified using the survival R package were further subjected to univariate Cox regression analysis to screen and identify significant prognosis candidates. The candidate genes were incorporated into the multivariate Cox risk model. Eight GRGs associated with significant prognosis potential were retained in the process of multiple calculations. The risk score for each patient was calculated as follows:

$$Risk\ score = \sum regression\ coefficient(genei) \times expression\ value(genei)$$

Colon adenocarcinoma patients in the training group were categorized into the low- and high-risk groups based on their median risk score. The OS rates of the two subgroups were compared by the Kaplan–Meier method and the log-rank test. A receiver operating characteristic (ROC) curve was constructed by SurvivalROC R package to evaluate the prognosis ability of the aforementioned model. In addition, 177 COAD patient samples from the GSE17536 dataset (Li et al., 2018) were used to validate the predictive ability of this prognostic model. Finally, the RMS-R package was used to construct a nomogram with calibration plots to show the likelihood OS rate.

[1]https://portal.gdc.cancer.gov/repository

[2]https://string-db.org/

[3]http://apps.cytoscape.org/apps/iregulon

Patients and Tissue Sampling

A total of 43 pairs of colon tumor tissues and adjacent non-tumor tissues were collected from patients who underwent colon cancer surgery at Shanghai Tenth People's Hospital from September 2019 to June 2020. After surgical resection, the colon cancer tissue samples were immediately cleaned in normal saline and cryopreserved. Prior to the surgery, none of the patients had received radiation and chemotherapy. This study was sanctioned by the Institutional Research Ethical Committee (IREC) of Shanghai Tenth People's Hospital of Tongji University of Medicine, Shanghai. All patients provided their verbal and written consent.

RNA Extraction and Quantitative Reverse Transcription Polymerase Chain Reaction Analysis

Total RNA from the colon cancer tissue and CRC cell lines was extracted using TRIzol reagent (Invitrogen, Thermo Scientific, Shanghai, China). Total mRNA was reverse-transcribed into cDNA using the prime script Reverse Transcription reagent kit (TaKaRa Bio, Shiga, Japan). Relative expression of eight potentially prognostic GRG markers was determined by qRT-PCR using the SYBR Green reagent kit (TaKaRa Bio) in an ABI 7500 PCR system (Applied Biosystems). Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) was used to normalize gene expression. Primer sequences are listed in **Supplementary Table 1**.

Immunohistochemistry Staining of Colon Cancer Tissue Sections

Paraffin-embedded colon cancer tissue was cut into 5-μm-thin sections. The sections were deparaffinized and rehydrated. Antigen retrieval was performed by exposure of the sections to 3% hydrogen peroxide along with normal goat serum. Antigen retrieval tissue sections were initially treated with antibody to stanniocalcin-2 (STC2, 1:200 dilution), prolyl 4-hydroxylase subunit alpha 1 (P4HA1, 1:200), protein phosphatase 2 catalytic subunit beta (PPP2CB, 1:200 or 1:250), peroxisome proliferator-activated receptor gamma coactivator 1-alpha (PPARGC1A, 1:250), enolase 3 (ENO3, 1:200), chondroitin polymerizing factor 2 (CHPF2, 1:500), phosphoglucomutase 2 (PGM2, 1:200), and phosphomannomutase 2 (PMM2, 1:200) (all from Abcam, Cambridge, United Kingdom). The sections were subsequently treated with biotinylated secondary rabbit antibodies conjugated with streptavidin–horseradish peroxidase (1:200). The binding of secondary antibody was revealed by reaction with added 3,3'-diaminobenzidine (DAB) followed by counterstaining with hematoxylin. The stained slides were visualized and photographed by optical microscopy (Leica, Tokyo, Japan). Image-Pro Plus 6.0 Software (Media Cybernetics, Rockville, MD, United States) was used to analyze protein expression in the stained tissue sections.

Cell Culture and Transient Transfection

HCT116 and SW480 human CRC cell lines were cultured at 37°C in a 5% CO_2 incubator in Roswell Park Memorial Institute (RPMI) 1,640 medium or Dulbecco's modified Eagle's medium (both from Gibco, Grand Island, NY, United States) supplemented with 10% fetal bovine serum (FBS). For transient transfection, prior to the day of transfection 2.5×10^5/ml of HTC116 and SW480 cells were seeded in six-well plates and incubated overnight to allow development of 40 to 60% confluent growth. These cells were transfected with small interfering RNA (siRNA) to STC2 (Gene Pharma, Shanghai, China) using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, United States) and incubated further for 48 h. Target sequences for siRNAs were as follows:

STC2-si1: GCGUGUUUGAAUGUUUCGATT and *STC2-si2: GGGUGAUAGUGGAGAUGAUTT*.

Cell Proliferation and Colony Formation Assay

Proliferation of STC2-siRNA-treated HCT116 and SW480 cells was determined using the Cell Counting Kit-8 (Dojindo Molecular Technologies, Inc., Rockville, MD, United States) as previously described (Liu et al., 2019). A clonogenic assay was performed to measure and the monitor the colony-forming capabilities of STC2-siRNA CRC cell (Zhen et al., 2019).

Cell Invasion and Migration Assay

Matrigel precoated Transwell inserts were placed into 24-well plate filled with 500 μl of culture medium supplemented with 10% FBS. Aliquots (400 μl containing 5 × 104 SW480 or HCT116 cells) of STC2-siRNA-treated cell suspensions were added to the upper chamber of Transwell units and incubated at 37°C for 20–24 h in a 5% CO_2 incubator. Cells that traversed the membrane separating the upper and lower chambers of each unit (i.e., invading cells) were fixed with 4% paraformaldehyde (PFA) for 15 min at room temperature and stained with 0.1% crystal violet. Cells that did not invade were gently removed from the membrane surface exposed to the upper chamber using a wet cotton swab. Each Transwell unit was allowed to dry and examined by inverted microscopy (Leica). Quantitative analysis of invaded cell was performed using ImageJ software (NIH, Bethesda, MD, United States).

For the cell migration assay, SW480 and HCT116 cells were transfected with STC2-siRNA and cultured on solid medium. Twenty-four hours later, a wound was created horizontally by gently pressing a 200-μl sterile micropipette tip on the surface of the confluent cell monolayer on six-well plates. Each well was washed twice with serum-free medium. Photographs were taken with an inverted microscope (Leica) at 0 and 24 h. The cell migration rate was determined by time-lapse analysis using ImageJ software.

Evaluation of Immune Cell Infiltration

The CIBERSORT[4] online bioinformatics analytical tool was used to estimate and distinguish 22 commonly infiltrating immune cells that included B cells, B memory cells, plasma cells, CD8+ and naïve CD4+ T cells, natural killer cells, macrophages, dendritic

[4]https://cibersort.standford.edu/

cells, and mast cells in COAD patients with low- and high-risk scores (Newman et al., 2015).

Estimation of Tumor Mutation Burden, Copy Number Variation, and Tumor Stemness

Tumor mutation burden (TMB) on the target gene was calculated by dividing the total number of mutations by the size of the coding region of the target gene. The GenVisR R package was used for the evaluation and analysis of the top 30 frequently mutated genes among the low- and high-risk groups (Skidmore et al., 2016). The copy number variation (CNV) in these groups of patients was calculated using chi-square test (Feng et al., 2020). The Rcircos R package was used to construct Circos plots to determine the CNV in chromosomes (Zhang et al., 2013). Utilization of the mRNA expression-based stemness index (mRNAsi) and epidermal growth factor receptor (EGFR)-mRNAsi data obtained from Tathiane Malta, University of São Paulo, was used to estimate the tumor stemness index to assess the dedifferentiation potential of oncogenic cells (Malta et al., 2018).

Prediction of Responses of Antineoplastic Drugs and Small Candidate Molecules

Based on the Genomics of Drug Sensitivity in Cancer (GDSC)[5], the efficacy of distinct antineoplastic drugs on COAD TCGA samples was determined using the pRRophetic R package. The half-maximal inhibitory concentration (IC50) of a specific chemotherapeutic drug was obtained by ridge regression analysis. The prediction accuracy was measured by the 10-fold cross-validation of the GDSC cell line expression profile data. In

[5]https://www.cancerrxgene.org

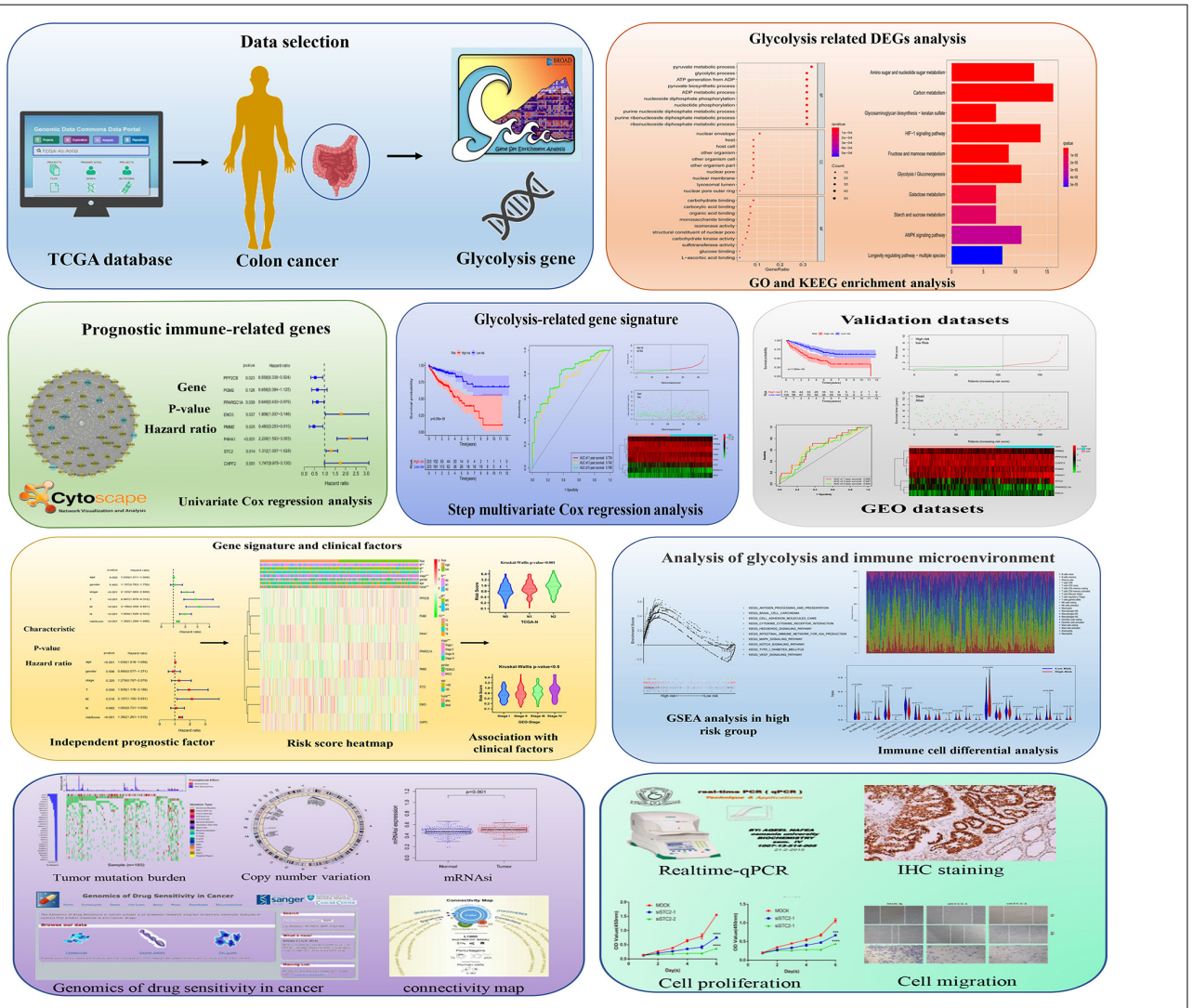


FIGURE 1 | Schematic diagram of the comprehensive analysis of glycolysis-related genes (GRGs) for prognosis prediction and drug selection in colon adenocarcinoma (COAD) patients.

addition, the Broad Institute Connectivity Map (CMap[6]) was used to predict the small candidate molecules based on the top 1,000 differentially expressed genes (Subramanian et al., 2017).

Statistical Analyses

All statistical analyses were performed using R software (version 3.6.1). The Wilcoxon test was used for the comparison of two independent non-parametric samples. The Kruskal–Wallis test was used for multiple independent samples. The Kaplan–Meier survival curves were compared with the log-rank test. Univariate and multivariate Cox proportional hazard regressions were performed to identify independent prognostic factors associated with OS. A p-value < 0.05 was considered statistically significant.

RESULTS

As increased aerobic glycolysis activity plays a significant role in cancer development, we explored the prognostic importance of GRGs in colon cancer (**Figure 1**). TCGA data of a total of 473 COAD patients and 177 patients from the Gene Expression Omnibus (GEO) (GSE17536) cohort were included. The demographics and clinical details of the patients are summarized in **Table 1**.

[6]https://clue.io/query

TABLE 1 | Clinical characteristics of the colon cancer patients used in this study.

	TCGA cohort	GSE17536
Number of patients	473	177
Age in years (median, range)	69 (31–90)	66 (26–92)
Gender (%)		
Female	215 (45.5%)	81 (45.8%)
Male	258 (54.5%)	96 (54.2%)
Grade (%)		
1	NA	16 (9.1%)
2	NA	133 (75.7%)
3	NA	27 (15.3%)
4	NA	0 (0%)
Stage (%)		
I	76 (16.1%)	24 (13.6%)
II	177 (37.4%)	57 (32.2%)
III	125 (26.4%)	57 (32.2%)
IV	62 (13.1%)	39 (22.0%)
T		
I	10 (2.11%)	NA
II	77 (16.3%)	NA
III	307 (64.9%)	NA
IV	56 (9.72%)	NA
M		
M0	333 (70.4%)	NA
M1	62 (13.1%)	NA
MX	49 (10.4%)	NA
N		
N0	268 (56.7%)	NA
N1	103 (21.8%)	NA
N2	80 (16.9%)	NA

NA, not applicable; TCGA, The Cancer Genome Atlas.

Identification of Differential Glycolysis-Related Genes and Functional Annotations

The primary screening and analysis of TCGA samples by GSEA identified 226 DE-GRGs in COAD samples compared with normal tissues (**Figure 2A**). Of these, 154 were overexpressed and 72 were underexpressed (**Figures 2B,C**; **Supplementary Table 2**). The DE-GRGs were significantly augmented in BP terms related to carbohydrate metabolism, nucleotide metabolism, AMP-activated protein kinase, and hypoxia-inducible factor signaling pathway (**Figure 2D** and **Supplementary Table 2**). KEGG pathway analysis shows that DE-GRGs are involved in HIF-1 signaling pathway, glycolysis/gluconeogenesis, AMPK signaling pathway, and galactose metabolism (**Figure 2E** and **Supplementary Table 3**).

Establishment of Protein–Protein Interaction Network and Selection of Prognosis-Related Glycolysis-Related Genes

To investigate the role of GRGs in COAD, a PPI network with 177 nodes and 1,216 edges was constructed using Cytoscape String software (**Figure 3A**). In addition, a co-expression network with 69 nodes and 701 edges was created using the MCODE tool (**Figure 3B**) to identify the key modules of GRGs. To explore the prognostic value of these GRGs, univariate Cox regression analysis was performed. Seventeen prognostic-associated candidate hub GRGs were revealed (**Figure 3C**). These candidate hub GRGs were analyzed by multiple stepwise Cox regression to investigate their impact on patient survival and clinical outcomes. Eight hub GRGs were independent predictors in COAD patients (**Figure 3D** and **Table 2**).

Prognosis-Related Glycolysis-Related Genes Risk Score Model Construction and Validation

A prediction model incorporating a prognosis-related signature of the eight hub GRGs was determined with the following calculation of the risk score:

$$Risk\ score = (-0.5826 * ExpPPP2CB) + (-0.4192 * ExpPGM2)$$

$$+(-0.4322 * ExpPPARGC1A) + (0.5911 * ExpENO3) +$$

$$(-0.735 * ExpPMM2) + (0.7924 * ExpP4HA1) +$$

$$(0.2715 * ExpSTC2) + (0.5578 * ExpCHPF2)$$

Survival analysis in the low- and high-risk groups was determined using the Kaplan–Meier survival plot. The analysis revealed a relatively low OS rate in patients with high-risk score compared with the rate in the low-risk group of patients

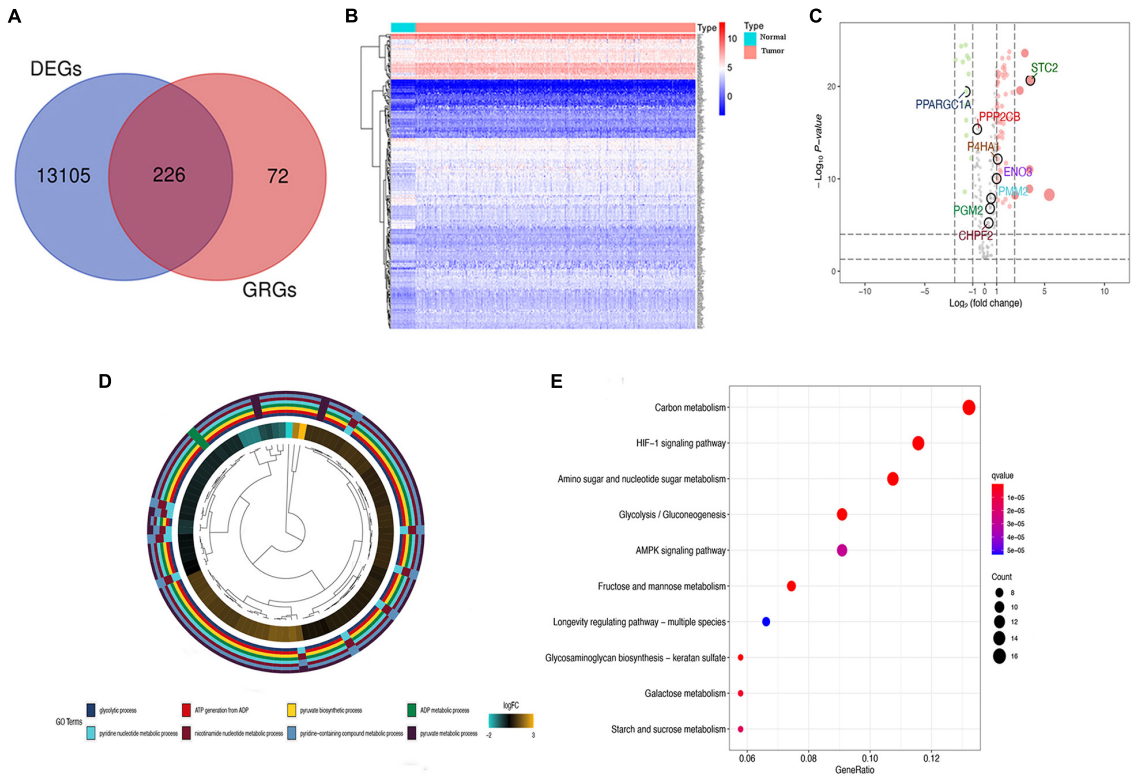


FIGURE 2 | Differential glycolysis-related genes (DE-GRGs) in The Cancer Genome Atlas (TCGA) patients. **(A)** Venn diagram of DE-GRGs among tumor and non-tumor TCGA samples. **(B)** Heatmap of expression of DE-GRGs in colon adenocarcinoma (COAD) sample. **(C)** Volcano plot of DE-GRG expression in tumors and normal tissue samples. Green and red dots indicate downregulated and upregulated GRGs, respectively. **(D)** Biological process (BP) analysis of DE-GRGs. **(E)** Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of DE-GRGs.

(**Figure 4A**). The risk scores, survival scores, and heatmap of GRG expression among the low- and high-risk patients are presented in **Figure 4B**. In addition, the area under the curve (AUC) analysis of COAD patients at 1, 3, and 5 years revealed respective OS rates of 73.4, 74.2, and 78.8% (**Figure 4C**). To verify the effectiveness of this model, a risk score was calculated in the GSE17536 cohort of patients, and they were classified into the high-risk group (n = 71) and low-risk group (n = 106) based on median risk score. Surprisingly, the results obtained from the GSE17536 cohort revealed that the high-risk score patients had lower OS rates than the patients with low-risk score, which was identical to the results obtained in TCGA cohort analysis (**Figures 4D–F**).

The Glycolysis-Related Gene Signature Confers Additional Prognostic Power for Colon Adenocarcinoma Patients

Considering the importance of glycolysis in the prognosis of colon cancer, we further analyzed the relationship between the eight GRGs and the clinical characteristics of colon cancer, including age, sex, tumor stage, and TMN stage. The heatmap (**Figure 5A**) indicated that the expression of CHPF2, ENO3, STC2, and P4HA1 was upregulated in the high-risk group, whereas that of PPP2CB, PGM2, PMM2, and PPARGC1A was downregulated in the high-risk group. Furthermore, a substantial difference in risk scores was evident for the different T, N, and M grades and tumor stages. The risk score increased with tumor progression (**Figure 5B** and **Supplementary Figure 2A**). To better predict the prognosis of the model, stratification analysis was used to confirm whether the model retained the ability to predict OS in different clinical subgroups. The OS rate of colon cancer patients with high-risk score, compared with patients in the low-risk group, was worse in the T3 + T4 and M0 groups (**Figure 5C**). Similarly, the risk model was good at predicting the OS rate of patients >65 or ≤65 years of age, male or female patients, stage I + II or III + IV, grade 1 or grade 2 + 3, and N0 or N1 (**Figure 5C** and **Supplementary Figure 2B**). The collective data indicate that the model may be a good prognostic index for patients with COAD.

Gene Set Enrichment Analysis Enrichment and Immune Cell Infiltration Analysis in the Cancer Genome Atlas and Gene Expression Omnibus Datasets

Gene set enrichment analysis of the GEO and TCGA databases revealed the aberrant activation of multiple signaling pathways. These included oxidative phosphorylation, chemokine receptor interaction, glycerol metabolism, chemokine signal pathway,

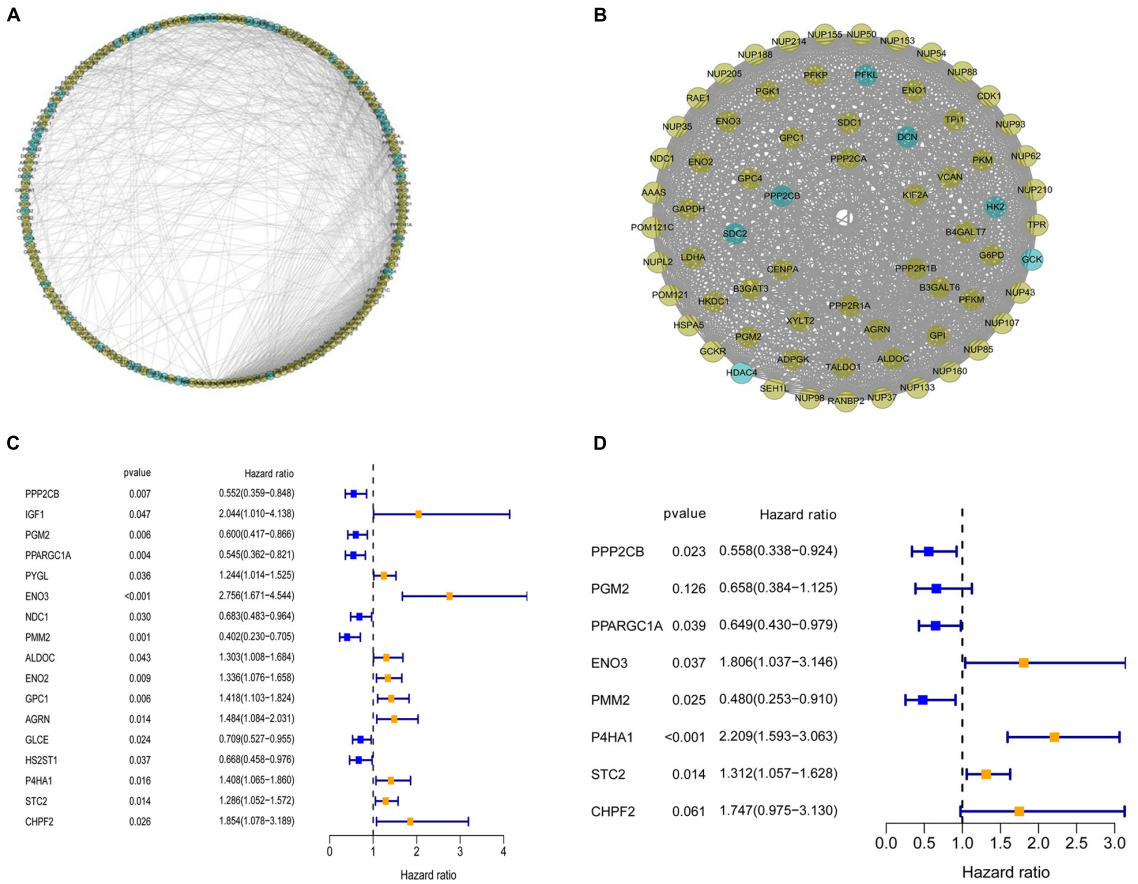


FIGURE 3 | Protein–protein interaction (PPI) network construction and selection of prognosis–related glycolysis-related genes (GRGs). **(A)** PPI network of differential GRGs (DE-GRGs). **(B)** Critical module constructed from PPI network. Green circles indicate downregulation with more than two-fold change. Yellow circles indicate upregulation with more than two-fold change. **(C)** Univariate Cox regression analysis with preliminary identified prognostic related GRGs. **(D)** Multivariate Cox regression analysis with prognosis-related GRGs.

TABLE 2 | Glycolysis related genes Multivariate Cox regression results.

	Multivariate cox analysis	
GRGs	**HR**	***P* value**
PPP2CB	0.558(0.338–0.924)	0.023
PGM2	0.658(0.384–1.125)	0.126
PPARGC1A	0.649(0–430-0.979)	0.039
EN03	1.806(1.037–3.146)	0.037
PMM2	0.480(0.253–0.910)	0.025
P4HA1	2.209(1.593–3.063)	<0.001
STC2	1.312(1.057–1.628)	0.014
CHPF2	1.747(0.975–3.130)	0.061

metabolism or exogenous substances through cytochrome P450, and transforming growth factor-beta signaling pathways in high-risk COAD patients (**Figures 6A,B**).

CIBERSORT along with the LM22 matrix was used to assess immune cell infiltration in the low- and high-risk groups of COAD patients. The GEO and TCGA and dataset analyses indicated that CD4+ and CD8+ T cells, macrophages, and mast cells were the predominant infiltrating immune cells in COAD (**Figures 6C,D**). Furthermore, infiltrations of resting CD4+ memory T cells, resting dendritic cells, and activated dendritic cells were significantly downregulated in the high-risk patients. In contrast, the distributions of M0 macrophages and eosinophils were opposite in the low- and high-risk TCGA groups (**Figure 6E**). Similarly, in the GSE17536 dataset, the infiltrating immune cells comprised significantly low proportions of CD4+ memory resting T cells and resting dendritic cells and high infiltration of M0 macrophages compared with those in the low-risk group (**Figure 6F**). However, there was no significant difference in the infiltrations of activated dendritic cells and eosinophils between the low- and high-risk groups.

Estimation of Mutational Load, Genetic Variation, Stemness Index, and Sensitivity to Antineoplastic Drugs

To explore the underlying reasons for the difference in prognosis of patients in the low- and high-risk groups, we analyzed the TMB, CNV, and tumor stemness among the low- and high-risk

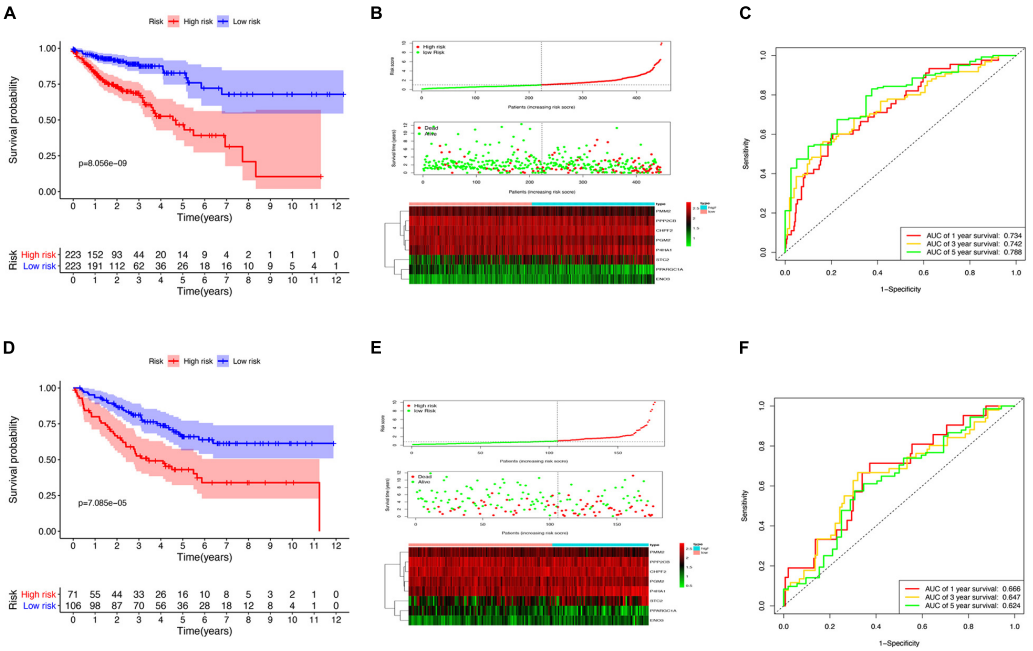


FIGURE 4 | Risk model construction and validation. **(A)** Kaplan–Meier plots illustrating the worse overall survival (OS) rate in the high-risk group compared with the low-risk group patients in The Cancer Genome Atlas (TCGA) cohort. **(B)** Expression heatmap, risk score distribution, and survival status in TCGA cohort. **(C)** Time-dependent area under the curve (AUC) of the receiver operating characteristic (ROC) curve of the 1-, 3-, and 5-year OS rates of TCGA patients. **(D)** Kaplan–Meier plots of worse OS in the high-risk patients compared with low-risk patients in the GSE17536 cohort. **(E)** Expression heatmap, risk score distribution, and survival status in the GSE17536 cohort. **(F)** Time-dependent AUC ROC curve of 1-, 3-, and 5-year OS rates of patients in the GSE17536 cohort.

groups. Waterfall plots were constructed to demonstrate the TMB and to differentiate TMB among the two risk groups. As shown in **Figures 7A,B**, most genes displayed higher mutation frequencies in the high-risk group. These genes included TP53, TTN, and KRAS. We calculated the TMB for each patient in both risk groups. The mean TMB was not significantly different between the two groups ($p > 0.05$; **Figure 7C**). In addition, distinct variations in CNV were evident in 920 genes predominantly located in chromosomes 1, 11, 14, 17, 18, and 19 (**Figure 7D**). A one-class logistic regression analysis was performed to calculate (mRNAsi) and epigenetic regulation-based stemness index (EREG-mRNAsi). The high-risk group had higher stemness indices compared with the low-risk group (**Figures 7E,F**). To evaluate the therapeutic effect of the risk model, the pRRophetic R package was used to analyze data downloaded from the GDSC database for the prediction of the sensitivity of distinct chemotherapeutic drugs on GRGs. Estimated IC50 values indicated that the low-risk patients were more sensitive to the anticancer drugs olaparib, veliparib, axitinib, metformin, and rapamycin ($p < 0.05$; **Figure 7G**).

Glycolysis-Related Gene Prognostic Model Is an Independent Prognostic Factor for Colon Adenocarcinoma Patients

To further judge whether the GRG model can be used as an independent factor to predict the prognosis of patients with colon cancer, Cox regression analysis was performed to reveal an association between risk score and OS of the patients.

Univariate Cox regression analysis revealed the significant association of risk score with OS in both TCGA cohort (hazard ratio [HR] = 1.382, 95% confidence interval [CI] = 1.284–1.488, $p < 0.001$) and GSE17536 cohort (HR = 1.252, 95% CI = 1.125–1.393, $p < 0.001$); **Figures 8A,B**). Multivariate Cox regression analysis with correction for distinct confounding parameters still revealed the risk score as an independent predictor for OS (TCGA cohort: HR = 1.382, 95% CI = 1.261–1.515, $p < 0.001$; GSE 17536 cohort: HR = 1.246, 95% CI = 1.104–1.406, $p < 0.001$; **Figures 8C,D**). Subsequently, ROC analysis was used to explore whether the risk score model was more accurate in predicting OS than clinical characters. As shown in **Figures 8E,F**, the predictive ability of the risk score model was stronger than that of other clinical characteristics. To establish a quantitative tool capable of predicting the clinical application of OS in patients with colon cancer, we established a nomogram integrating the risk score and clinical characters to calculate 1-, 3-, and 5-year OS rates of patients (**Figures 8G,H**). Calibration plots indicated that the nomogram versus an ideal model showed high consistency in TCGA and GSE17536 cohorts (**Figures 8I,J**).

Identification of Potent Drugs Targeting Glycolysis-Related Genes in the CMap Database

Sixty potential small molecule drugs targeting genes for GRGs were identified by CMap database analysis (**Figure 9**). The

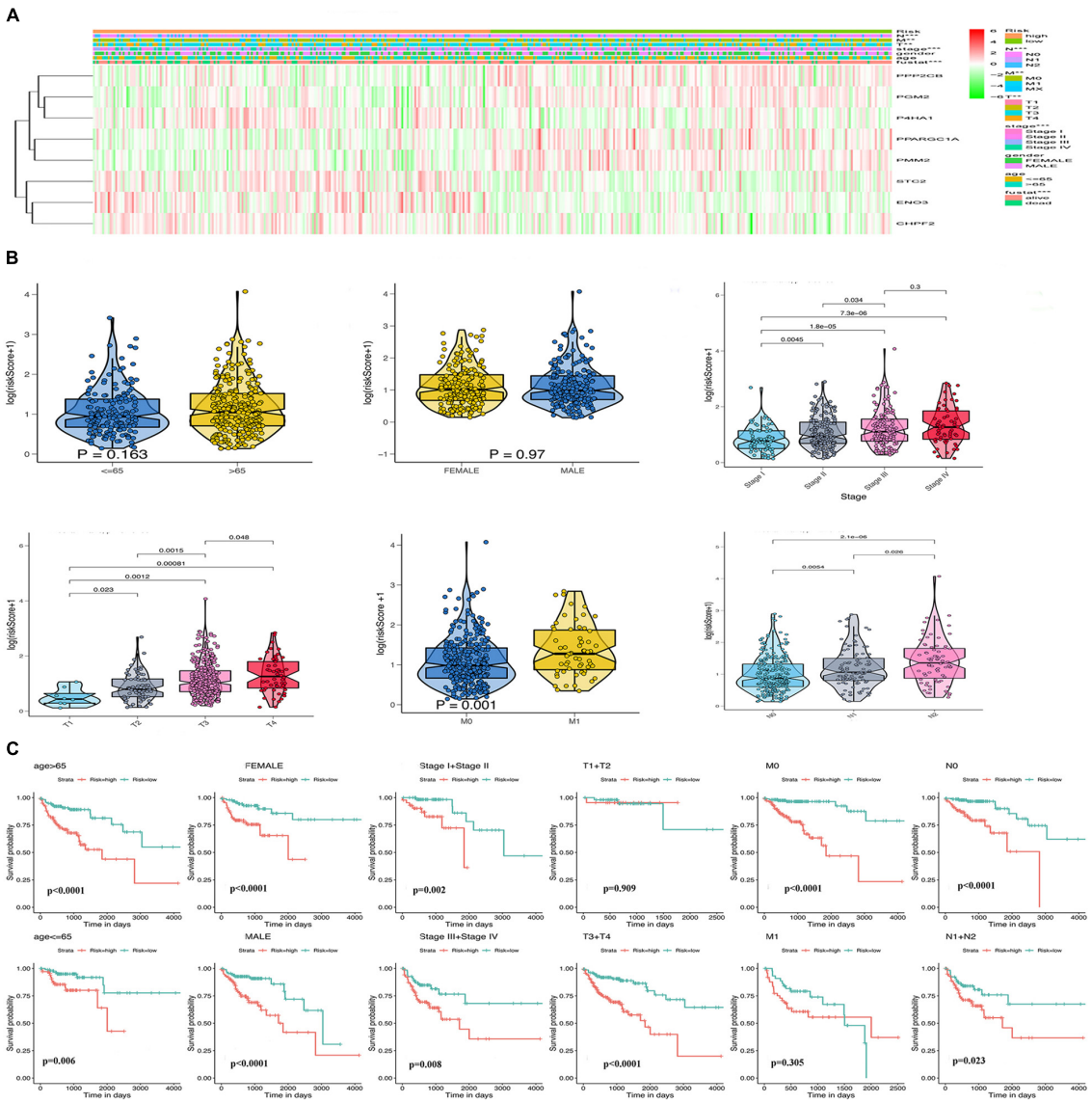


FIGURE 5 | Relationship of the GRG signature with clinicopathological characteristics in The Cancer Genome Atlas (TCGA) data. **(A)** Heatmap of the differential expression of the eight GRGs. **(B)** Box plots showing risk score distribution of different clinical characteristics of the colon adenocarcinoma (COAD) tumors. **(C)** The GRG risk model predicts the overall survival (OS) rate of patients with colon cancer in multiple clinical subgroups.

mechanisms of the drugs were distinct. Among the 60 drugs, three (bicuculline, NCS-382, and pentylenetetrazol) are gamma aminobutyric acid (GABA) receptor antagonists and three other drugs (prednisolone, diflorasone, and fluocinonide) are glucocorticoid receptor agonists. The findings indicated that drugs targeting the signature GRGs might have therapeutic importance in the treatment of colon cancer.

Validation of the Expression and the Prognosis of Eight Glycolysis-Related Genes

qRT-PCR was performed from RNA extracted from 43 pairs of colon cancer and adjacent normal tissues to observe the relative expression levels of eight GRGs (P4HA1, STC2, CHPF2, PMM2, PGM2, ENO3, PPARGC1A, and PPP2CB). P4HA1, STC2, CHPF2, PMM2, PGM2, and ENO3 were expressed at higher levels in COAD tissues, with lower expression of PPP2CB and PPARGC1A (**Figure 10A**). The findings were consistent with the results of analyses of the Gene Expression Profiling Interactive Analysis (GEPIA) and UALCAN online databases (**Supplementary Figure 3**).

Similarly, immunohistochemistry (IHC) staining of colon cancer tissue with antibodies to P4HA1, STC2, CHPF2, PMM2, PGM2, ENO3, PPARGC1A, and PPP2CB had higher expression of the P4HA1, STC2, PMM2, CHPF2, and ENO3 GRGs in COAD tissue and lower expression of the PPP2CB, PPARGC1A, and PGM2 GRGs in the same tissues (**Figures 10B,C**). In addition,

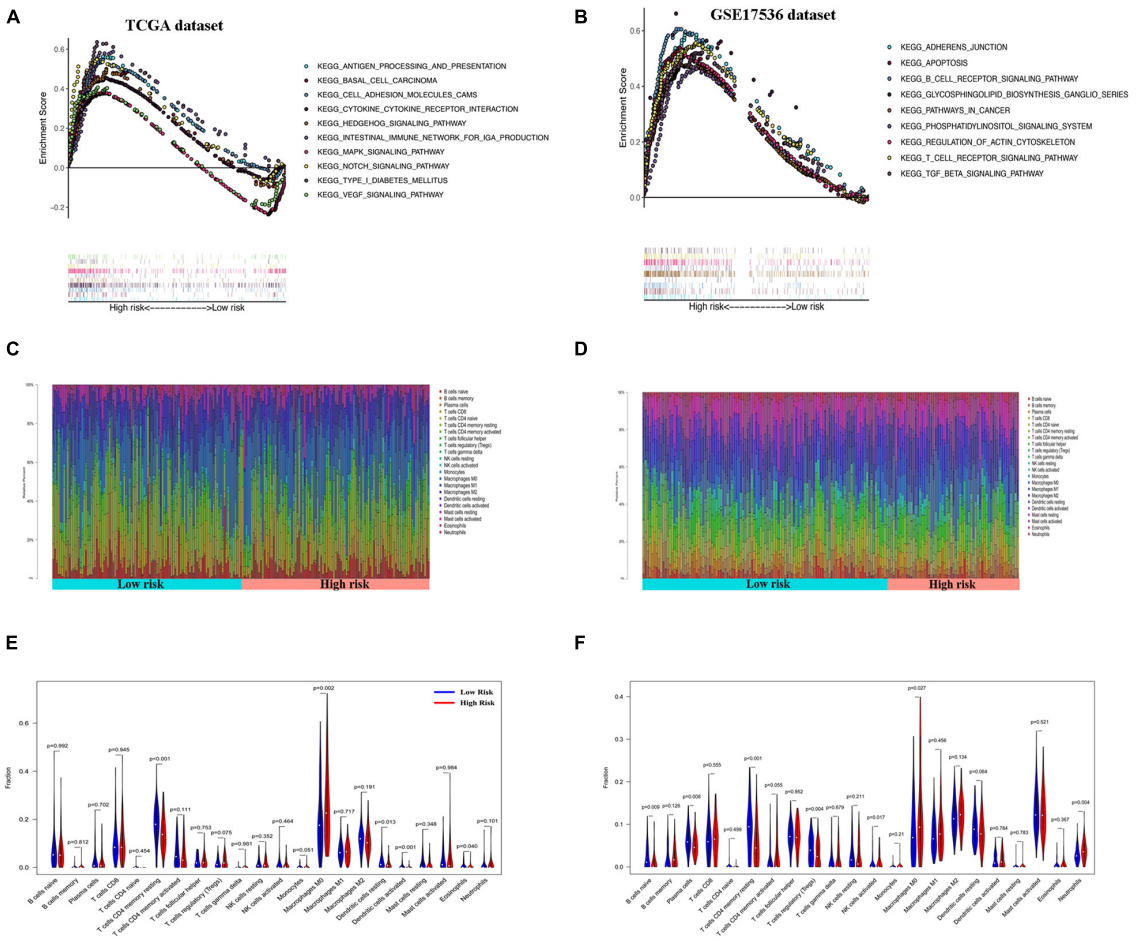


FIGURE 6 | Gene set enrichment analysis (GSEA) and immune cell infiltration analyses in low- and high-risk patients. **(A,B)** Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment of low- and high-risk groups by GSEA in The Cancer Genome Atlas (TCGA) cohort **(A)** and GSE17536 cohort **(B)**. Only gene sets with false discovery rate (FDR) $q < 0.05$ were considered significant. **(C,D)** Relative proportion of infiltrating immune cells in low- and high-risk colon adenocarcinoma (COAD) patients in TCGA cohort **(C)** and GSE17536 cohort **(D)**. **(E,F)** Comparative demonstration of the distinct infiltrating immune cells in low- and high-risk COAD patients in TCGA cohort **(E)** and GSE17536 cohort **(F)**.

there were significant differences in the expression of STC2 in the COAD tissues at different stages of COAD (I, II, III, and IV; **Figure 10D**). These findings suggest that the expression level of STC2 changes with the progression of tumors.

To further explore the potential prognostic value of eight GRGs in COAD patients, we performed the Kaplan–Meier survival analysis to determine the effect of the eight GRGs on OS in TCGA COAD patients. Overexpression of STC2, ENO3, P4HA1, and CHPF2 and low expression of PPP2CB, PPARGC1A, PMM2, and PGM2 are associated with lower OS rate in COAD patients (**Supplementary Figure 4**).

Knockout of STC2 Protein Inhibits the Proliferation and Invasion of Colorectal Cancer Cells

Considering the potential role of STC2 in CRC progression, we performed *in vitro* proliferation, colony formation, migration, and invasion assays. The high expression of STC2 was associated with poor outcomes of CRC patients. We further evaluated the effect of STC2 depletion on the cellular process and BP in the CRC cells. Functionally, knockdown of STC2 by siRNA resulted in decreased cellular proliferation and colony formation by HCT116 and SW480 cells (**Figures 11A,B**). Similarly, knockdown of STC2 in HCT116 and SW480 cells using short hairpin RNA decreased cellular migration and invasion compared with control cells (**Figures 11C,D**). The findings confirmed the potential role of STC2 in the progression or CRC.

DISCUSSION

The present analyses revealed 226 DE-GRGs in TCGA COAD patients. In addition, gene ontology and KEGG analyses of DE-GRGs revealed their involvement in regulating the distinct functions and pathways involved in abnormal energy metabolism during tumor development and metastasis. Furthermore, eight

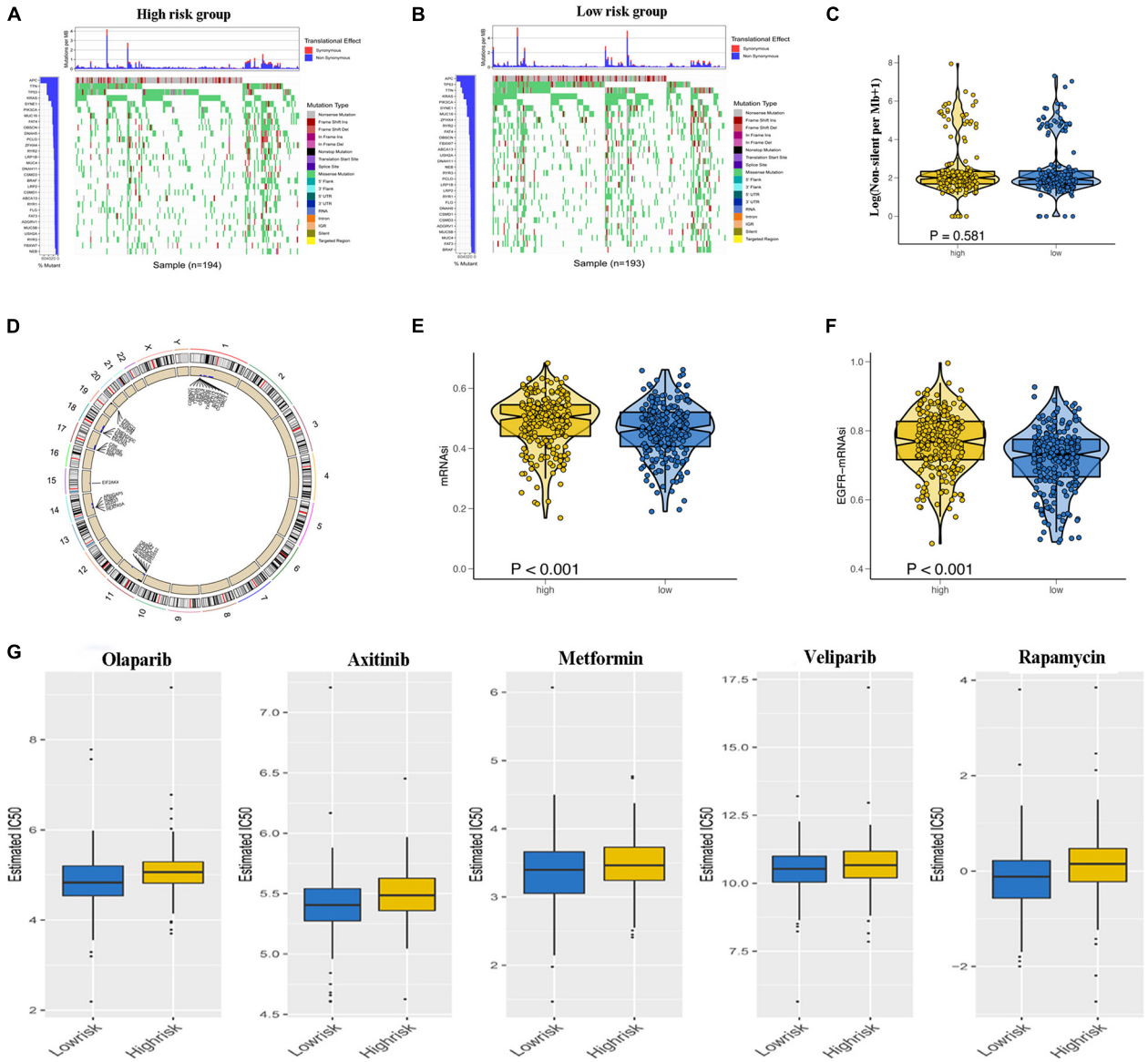


FIGURE 7 | Relationship of glycolysis-related gene (GRG) risk group with tumor mutation burden, copy number variation, tumor stemness, and chemotherapy response. **(A,B)** Waterfall plots demonstrating the higher frequency of mutated genes in the high-risk and low-risk groups. **(C)** Difference in total tumor mutation burden in low- and high-risk colon adenocarcinoma (COAD) patients. **(D)** Distinct genes with different copy number variations due to gains or loss of nucleotides mainly on chromosomes 1, 11, 14, 17, 18, and 19 **(E,F)**. Difference in tumor stemness by the mRNAsi **(E)** and EGFR mRNAsi **(F)**. **(G)** Estimated IC50 indicating the efficiency of chemotherapy to GRGs in low- and high-risk patients of olaparib, veliparib, axitinib, metformin, and rapamycin.

GRGs were identified as the most robust glycolysis-related prognostic signature. This signature was used to categorize the colon cancer patients into the low- and high-risk groups with significant differences in survival outcomes. Similarly, univariate and multivariate Cox regression analyses of eight GRGs demonstrated the independent ability to predict patient prognosis. The results further indicate that the signature GRGs are closely related to many clinical characteristics, such as the increased risk score with tumor development and distant metastasis. This discovery could aid in the dynamic monitoring of colon cancer patients. In addition, we further confirmed the expression of eight GRGs in colon cancer by qRT-PCR and IHC analyses. Surprisingly, the expression of protein increased with increased disease stage. Cell cloning and invasion experiments further confirmed the tumor-promoting function of STC2.

To explore the underlying mechanisms of the GRGs model, GSEA was performed to explore KEGG pathways among the two risk groups. Aberrant activation of distinct signaling pathways, including the notch signaling pathway, mitogen-activated protein kinase signaling pathway, regulation of actin cytoskeleton, phosphatidylinositol, and the transforming growth factor-beta signaling pathway, among the high-risk score patients

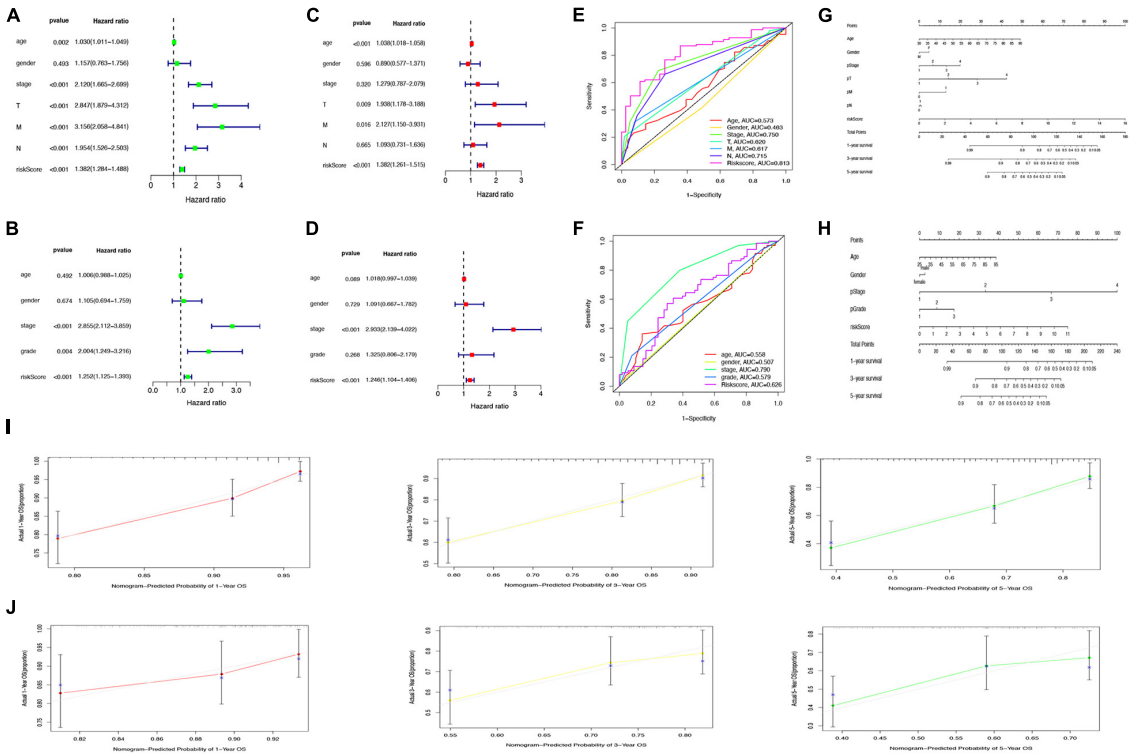


FIGURE 8 | Prediction of overall survival (OS) by risk score and clinical characteristics. **(A,B)** Forest plot of risk scores and other clinical factors based on a univariate Cox regression analysis in The Cancer Genome Atlas (TCGA) **(A)** and GSE17536 cohort **(B)**. **(C,D)** Forest plot of risk scores and other clinical factors based on a multivariate Cox regression analysis in TCGA **(C)** and GSE17536 **(D)** cohort. **(E,F)** Comparison of specificity and sensitivity of 5-year overall survival rate between different clinical traits and GRG risk models in TCGA **(E)** and GSE17536 **(F)** cohort. **(G,H)** Nomogram with a combination of risk score and different clinical traits in TCGA **(G)** and GSE17536 **(H)** cohort. **(I,J)** A calibration plot for predicting the accuracy of the nomogram in TCGA **(I)** and GSE17536 **(J)** cohort.

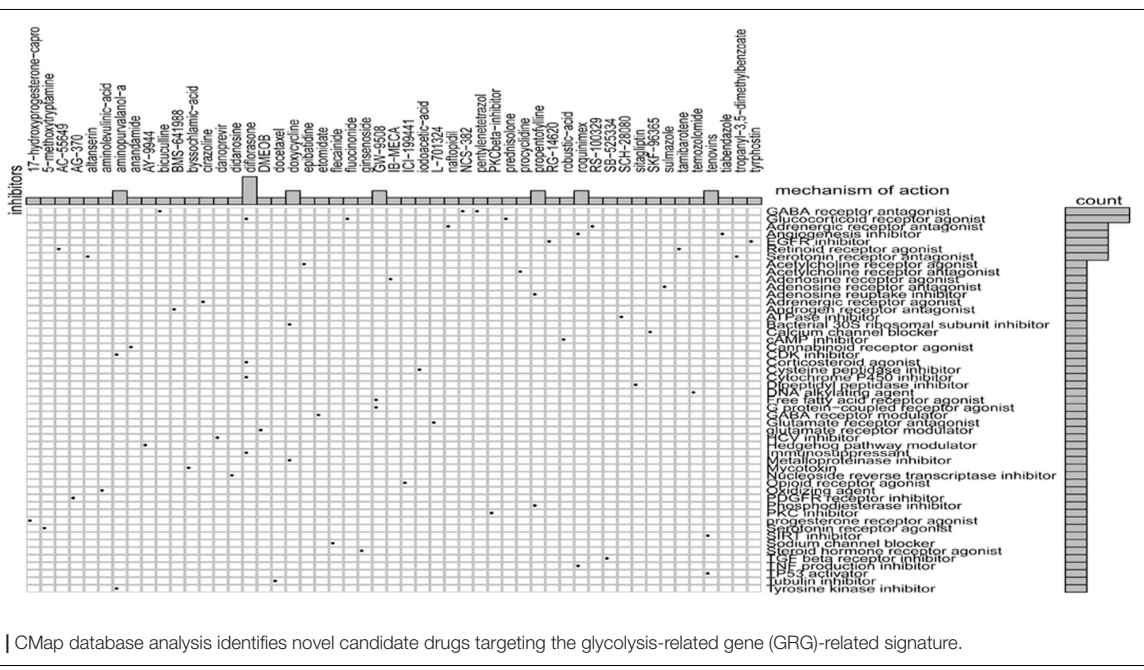


FIGURE 9 | CMap database analysis identifies novel candidate drugs targeting the glycolysis-related gene (GRG)-related signature.

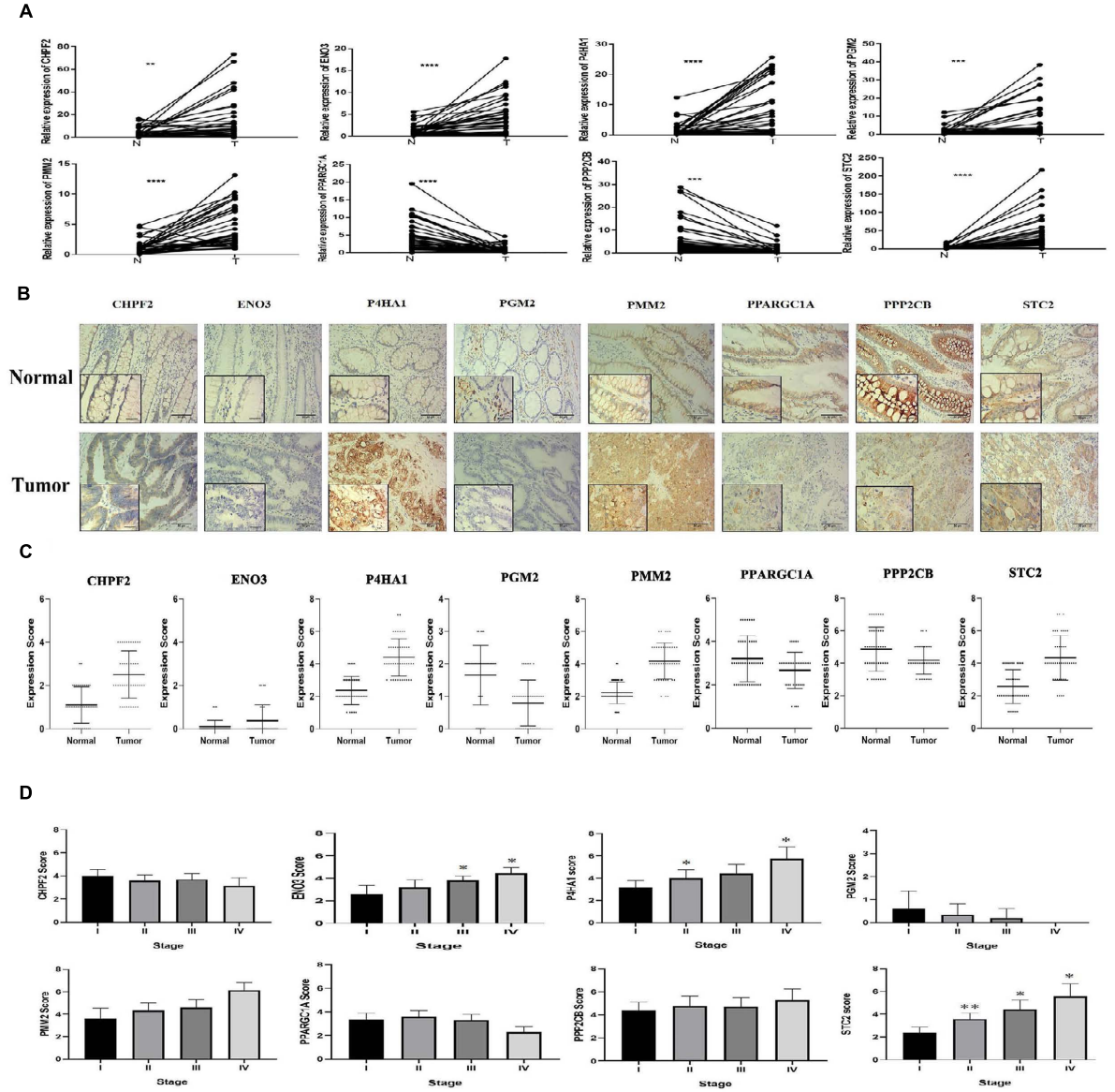


FIGURE 10 | Eight glycolysis-related gene GRG expression in clinical samples. **(A)** mRNA expression of eight GRGs in clinical samples. **(B)** Protein expression analysis of eight GRGs in tumor and normal samples using immunohistochemistry analysis. Scale bar represents × 100, × 400, or 100 μm. **(C)** Eight GRG protein scores in tumor and normal samples. **(D)** GRG protein expression at the distinct stage of tumor. One-way analysis of variance used for statistical analysis, $^{*}p < 0.05$, $^{**}p < 0.01$ versus stage I. All data are expressed as mean ± SEM.

is associated with poor prognosis of COAD (Meng et al., 2009; Lähde et al., 2020).

Considering the heterogeneity and complexity in tumor development, we comprehensively analyzed glycolysis-related models, including the TME, epigenetics, and tumor stemness index, to better understand the potential underlying mechanism among the low- and high-risk groups of COAD patients. Analysis of immune cell infiltration revealed a significant difference between the risk groups, suggesting that abnormal glycolysis metabolism could alter the immune microenvironment, which could affect prognosis and treatment. A prior study (Song and Wu, 2020) reported an association between decreased infiltration of CD4+ memory resting T cells and poor prognosis. Other studies reported a rapid transformation of resting dendritic cells to activated dendritic cells under *in vivo* stimulation and their participation in adaptive immune responses (Joffre et al., 2009). The number of infiltrating dendritic cells has been positively associated with survival in colon cancer and pancreatic ductal adenocarcinoma (Nagorsen et al., 2007; Wu et al., 2020). In addition, the increase of M0 macrophages can lead to decreased immune activity and poor prognosis in digestive system tumors (Nagorsen et al., 2007; Nie et al., 2020). Considering the

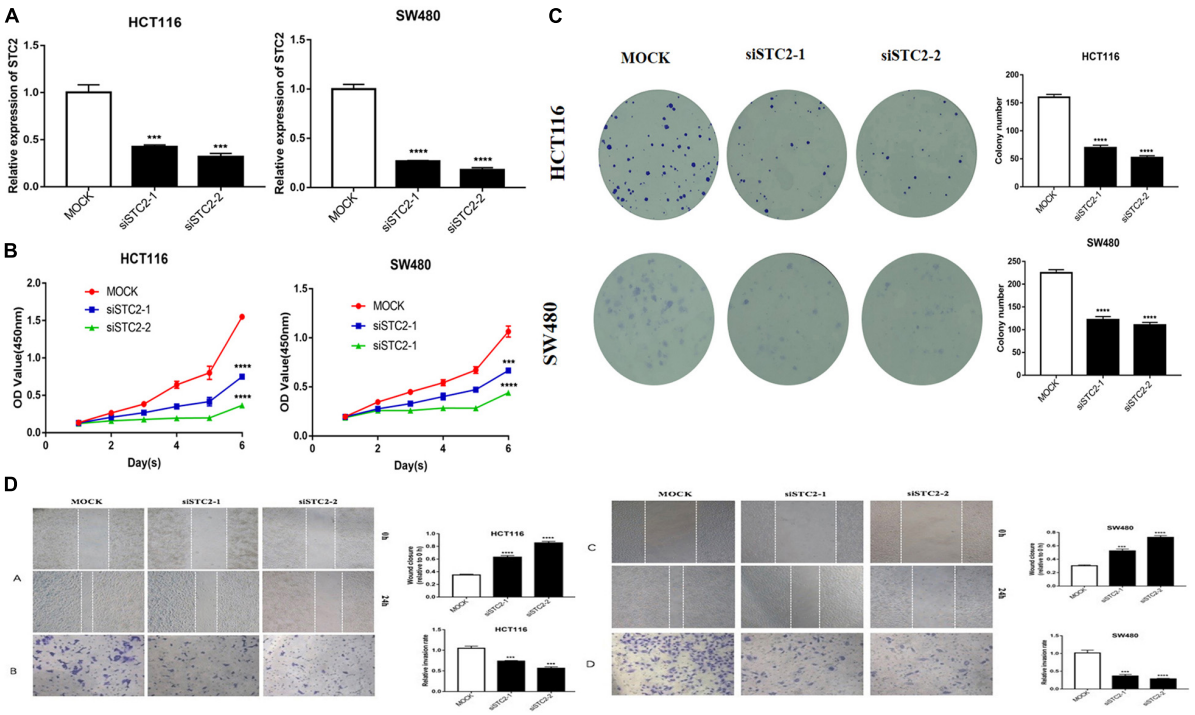


FIGURE 11 | STC2 knockdown inhibits cell proliferation, migration, and invasion *in vitro*. **(A)** STC2 expression in STC2-siRNA HCT-116 and SW-480 cells. **(B)** Growth curves for HCT-116 and SW-480 cells after transfection with STC2 siRNA determined by Cell Counting Kit-8 (CCK-8) analysis. **(C)** Clonogenic assay on STC2-siRNA knockdown HCT-116 and SW-480 cell lines showing decreased colony formation compared with the negative control. **(D)** Decreased cell migration and Transwell invasion activity in STC2-siRNA knockdown HCT-116 and SW-480 cells. Results are expressed as mean ± SEM. Significance difference was tested by one-way ANOVA: ***$p < 0.001$ versus mock, ****$p < 0.0001$ versus mock.

significant role of GRGs in the development and prognostic of colon cancer, we analyzed the differences between TMB, CNV, and mRNAsi in the low- and high-risk groups of COAD patients. Patients with high-risk score had higher TMB; however, there was no significant difference in total TMB. Higher tumor stem cell index and abnormal CNV were observed in the high-risk group, strongly implicating GRGs as the genetic alteration.

Similarly, scrutiny of the GDSC database determined the effectiveness of different chemotherapeutic drugs in the treatment of low- and high-risk COAD patients. Based on the IC50 values, olaparib, veliparib, axitinib, metformin, and rapamycin displayed a better response in treating the low-risk score COAD patients. These drugs have been reported to be important in the treatment of colon cancer patients (Faller et al., 2015; Arena et al., 2020; Guo et al., 2020). Additionally, using the CMap database, we identified 60 small molecule drugs with therapeutic potential. These drugs may have pronounced efficacy in COAD treatment. The drugs include glucocorticoid receptor agonist (prednisolone, diflorasone, and fluocinonide), GABA receptor antagonists (bicuculline, NCS-382, and pentylenetetrazol), inhibitors of angiogenesis (roquinimex and tiabendazole), and EGFR inhibitors (tyrphostin and RG-14620). Identification of distinct candidate drugs might be useful in the treatment of COAD patients by targeting GRGs. Together, the findings clearly indicate the significant importance of the eight GRG genes in COAD prognosis and treatment.

CONCLUSION

This comprehensive multi-database study is the first to investigate the expression profile of GRGs and their clinical significance in patients with colon cancer. The formulated risk model may provide an effective clinical tool in the prognosis and optimization of treatment for patients with colon cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Shanghai Tenth people's Hospital of Tongji University of Medicine Shanghai. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

ZC contributed to the research design and data analysis. GS performed the experiments and drafted the manuscript. RmB performed the experiments and assisted in editing. RjB assisted in editing. JL and MZ collected the clinical samples. FS assisted in modifying the manuscript and collating references. ZL and SZ revised the manuscript and provided writing guidance. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

This work was supported by the research fund of Shanghai Municipal Health Bureau (No. 2019cxjq03), the National Natural Science Foundation of China (Nos. 81930066 and 81772941), and Shanghai Municipal Health Commission (No. 20204Y0088).

ACKNOWLEDGMENTS

We acknowledge the Ethical Institutions and Colorectal Cancer Center of Shanghai Tenth People's Hospital for their cooperation, as well as the financial support institutions and all the authors who made an effort to this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.684322/full#supplementary-material

Supplementary Figure 1 | Identification of GRGs. GSEA identification of five gene sets that are significantly enriched, including Biocarta Glycolysis Pathway, Glycolytic Process, Hallmark Glycolysis, KEGG Glycolysis Gluconeogenesis, and Reactome Glycolysis.

Supplementary Figure 2 | Relationship of GRGs signature with clinicopathological characteristics in GSE17536. **(A)** Box plots showing risk score distribution of different clinical characteristics of COAD tumors in GSE17536 cohort. **(B)** GRGs risk model is of good value in predicting the OS rate of patients with colon cancer in multiple clinical subgroups.

Supplementary Figure 3 | Expression of the eight signature GRGs in online databases. **(A)** UALCAN online database (http://ualcan.path.uab.edu/) expression of eight glycolysis related genes in TCGA. **(B)** GEPIA online database (http://gepia.cancer-pku.cn/) expression of eight glycolysis related genes in TCGA and GTEx.

Supplementary Figure 4 | Evaluation of eight glycolysis genes on the prognosis of colon cancer.

Supplementary Table 1 | Primer sequences of eight glycolysis related genes.

Supplementary Table 2 | GO analysis of genes involved in differential glycolysis.

Supplementary Table 3 | Analysis of related KEEG pathways involved in differential glycolysis genes.

Supplementary Table 4 | Glycolysis related genes are differentially expressed in colon cancer tissues.

REFERENCES

Arena, S., Corti, G., Durinikova, E., Montone, M., Reilly, N. M., Russo, M., et al. (2020). A subset of colorectal cancers with cross-sensitivity to olaparib and oxaliplatin. *Clin. Cancer Res.* 26, 1372–1384. doi: 10.1158/1078-0432.Ccr-19-2409

Arnold, M., Abnet, C. C., Neale, R. E., Vignat, J., Giovannucci, E. L., McGlynn, K. A., et al. (2020). Global burden of 5 major types of gastrointestinal cancer. *Gastroenterology* 159, 335–349. doi: 10.1053/j.gastro.2020.02.068

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Cairns, R. A., Harris, I. S., and Mak, T. W. (2011). Regulation of cancer cell metabolism. *Nat. Rev. Cancer* 11, 85–95. doi: 10.1038/nrc2981

Corbet, C., and Feron, O. (2017). Tumour acidosis: from the passenger to the driver's seat. *Nat. Rev. Cancer* 17, 577–593. doi: 10.1038/nrc.2017.77

DeBerardinis, R. J., Lum, J. J., Hatzivassiliou, G., and Thompson, C. B. (2008). The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab.* 7, 11–20. doi: 10.1016/j.cmet.2007.10.002

Faller, W. J., Jackson, T. J., Knight, J. R., Ridgway, R. A., Jamieson, T., Karim, S. A., et al. (2015). mTORC1-mediated translational elongation limits intestinal tumour initiation and growth. *Nature* 517, 497–500. doi: 10.1038/nature13896

Feng, B., Shen, Y., Pastor Hostench, X., Bieg, M., Plath, M., Ishaque, N., et al. (2020). Integrative analysis of multi-omics data identified EGFR and PTGS2 as key nodes in a gene regulatory network related to immune phenotypes in head and neck cancer. *Clin. Cancer Res.* 26, 3616–3628. doi: 10.1158/1078-0432.Ccr-19-3997

Gatenby, R. A., and Gillies, R. J. (2004). Why do cancers have high aerobic glycolysis? *Nat. Rev. Cancer* 4, 891–899. doi: 10.1038/nrc1478

Gill, K. S., Fernandes, P., O'Donovan, T. R., McKenna, S. L., Doddakula, K. K., Power, D. G., et al. (2016). Glycolysis inhibition as a cancer treatment and its role in an anti-tumour immune response. *Biochim. Biophys. Acta* 1866, 87–105. doi: 10.1016/j.bbcan.2016.06.005

Guo, J., Li, Y., Duan, H., and Yuan, L. (2020). Metformin suppresses the proliferation and promotes the apoptosis of colon cancer cells through inhibiting the expression of long noncoding RNA-UCA1. *Onco Targets Ther.* 13, 4169–4181. doi: 10.2147/ott.S245091

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Joffre, O., Nolte, M. A., Spörri, R., and Reis e Sousa, C. (2009). Inflammatory signals in dendritic cell activation and the induction of adaptive immunity. *Immunol. Rev.* 227, 234–247. doi: 10.1111/j.1600-065X.2008.00718.x

Kalyan, A., Kircher, S., Shah, H., Mulcahy, M., and Benson, A. (2018). Updates on immunotherapy for colorectal cancer. *J. Gastrointest. Oncol.* 9, 160–169. doi: 10.21037/jgo.2018.01.17

Ladabaum, U., Dominitz, J. A., Kahi, C., and Schoen, R. E. (2020). Strategies for colorectal cancer screening. *Gastroenterology* 158, 418–432. doi: 10.1053/j.gastro.2019.06.043

Lähde, M., Heino, S., Högström, J., Kaijalainen, S., Anisimov, A., Flanagan, D., et al. (2020). Expression of R-Spondin 1 in apc(Min/+) mice reduces growth of intestinal adenomas by altering Wnt and transforming growth factor beta signaling. *Gastroenterology* 160, 245–259. doi: 10.1053/j.gastro.2020.09.011

Li, C., Shen, Z., Zhou, Y., and Yu, W. (2018). Independent prognostic genes and mechanism investigation for colon cancer. *Biol. Res.* 51:10. doi: 10.1186/s40659-018-0158-7

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* 47, W199–W205. doi: 10.1093/nar/gkz401

Liu, L., Wang, J., Sun, G., Wu, Q., Ma, J., Zhang, X., et al. (2019). m(6)A mRNA methylation regulates CTNNB1 to promote the proliferation of hepatoblastoma. *Mol. Cancer* 18:188. doi: 10.1186/s12943-019-1119-7

Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., et al. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* 173, 338–354. doi: 10.1016/j.cell.2018.03.034

Martinez-Outschoorn, U. E., Peiris-Pagés, M., Pestell, R. G., Sotgia, F., and Lisanti, M. P. (2017). Cancer metabolism: a therapeutic perspective. *Nat. Rev. Clin. Oncol.* 14, 11–31. doi: 10.1038/nrclinonc.2016.60

Meng, R. D., Shelton, C. C., Li, Y. M., Qin, L. X., Notterman, D., Paty, P. B., et al. (2009). Gamma-secretase inhibitors abrogate oxaliplatin-induced activation of the notch-1 signaling pathway in colon cancer cells resulting in enhanced chemosensitivity. *Cancer Res.* 69, 573–582. doi: 10.1158/0008-5472.Can-08-2088

Nagorsen, D., Voigt, S., Berg, E., Stein, H., Thiel, E., and Loddenkemper, C. (2007). Tumor-infiltrating macrophages and dendritic cells in human colorectal cancer: relation to local regulatory T cells, systemic T-cell response against tumor-associated antigens and survival. *J. Transl. Med.* 5:62. doi: 10.1186/1479-5876-5-62

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Nie, K., Zheng, Z., Wen, Y., Shi, L., Xu, S., Wang, X., et al. (2020). Construction and validation of a TP53-associated immune prognostic model for gastric cancer. *Genomics* 112, 4788–4795. doi: 10.1016/j.ygeno.2020.08.026

Pilleron, S., Sarfati, D., Janssen-Heijnen, M., Vignat, J., Ferlay, J., Bray, F., et al. (2019). Global cancer incidence in older adults, 2012 and 2035: a population-based study. *Int. J. Cancer* 144, 49–58. doi: 10.1002/ijc.31664

Skidmore, Z. L., Wagner, A. H., Lesurf, R., Campbell, K. M., Kunisaki, J., Griffith, O. L., et al. (2016). GenVisR: genomic visualizations in R. *Bioinformatics* 32, 3012–3014. doi: 10.1093/bioinformatics/btw325

Song, J., and Wu, L. (2020). Friend or foe: prognostic and immunotherapy roles of BTLA in colorectal cancer. *Front. Mol. Biosci.* 7:148. doi: 10.3389/fmolb.2020.00148

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: l1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452. doi: 10.1016/j.cell.2017.10.049

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Vander Heiden, M. G., Cantley, L. C., and Thompson, C. B. (2009). Understanding the warburg effect: the metabolic requirements of cell proliferation. *Science* 324, 1029–1033. doi: 10.1126/science.1160809

Vuik, F. E., Nieuwenburg, S. A., Bardou, M., Lansdorp-Vogelaar, I., Dinis-Ribeiro, M., Bento, M. J., et al. (2019). Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. *Gut* 68, 1820–1826. doi: 10.1136/gutjnl-2018-317592

Wu, M., Li, X., Liu, R., Yuan, H., Liu, W., and Liu, Z. (2020). Development and validation of a metastasis-related gene signature for predicting the overall survival in patients with pancreatic ductal adenocarcinoma. *J. Cancer* 11, 6299–6318. doi: 10.7150/jca.47629

Zhang, H., Meltzer, P., and Davis, S. (2013). RCircos: an R package for circos 2D track plots. *BMC Bioinformatics* 14:244. doi: 10.1186/1471-2105-14-244

Zhen, N., Gu, S., Ma, J., Zhu, J., Yin, M., Xu, M., et al. (2019). CircHMGCS1 promotes hepatoblastoma cell proliferation by regulating the IGF signaling pathway and glutaminolysis. *Theranostics* 9, 900–919. doi: 10.7150/thno.29515

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Cui, Sun, Bhandari, Lu, Zhang, Bhandari, Sun, Liu and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 10 August 2021
doi: 10.3389/fcell.2021.691161

Identification of Novel Metabolism-Associated Subtypes for Pancreatic Cancer to Establish an Eighteen-Gene Risk Prediction Model

*Yang Gao[1], Enchong Zhang[2], Xiang Fei[1], Lingming Kong[1], Peng Liu[1] and Xiaodong Tan[1]**

[1] Department of General Surgery, Shengjing Hospital of China Medical University, Shenyang, China, [2] Department of Urology, Shengjing Hospital of China Medical University, Shenyang, China

Pancreatic cancer (PanC) is an intractable malignancy with a high mortality. Metabolic processes contribute to cancer progression and therapeutic responses, and histopathological subtypes are insufficient for determining prognosis and treatment strategies. In this study, PanC subtypes based on metabolism-related genes were identified and further utilized to construct a prognostic model. Using a cohort of 171 patients from The Cancer Genome Atlas (TCGA) database, transcriptome data, simple nucleotide variants (SNV), and clinical information were analyzed. We divided patients with PanC into metabolic gene-enriched and metabolic gene-desert subtypes. The metabolic gene-enriched subgroup is a high-risk subtype with worse outcomes and a higher frequency of SNVs, especially in *KRAS*. After further characterizing the subtypes, we constructed a risk score algorithm involving multiple genes (i.e., *NEU2*, *GMPS*, *PRIM2*, *PNPT1*, *LDHA*, *INPP4B*, *DPYD*, *PYGL*, *CA12*, *DHRS9*, *SULT1E1*, *ENPP2*, *PDE1C*, *TPH1*, *CHST12*, *POLR3GL*, *DNMT3A,* and *PGS1*). We verified the reproducibility and reliability of the risk score using three validation cohorts (i.e., independent datasets from TCGA, Gene Expression Omnibus, and Ensemble databases). Finally, drug prediction was completed using a ridge regression model, yielding nine candidate drugs for high-risk patients. These findings support the classification of PanC into two metabolic subtypes and further suggest that the metabolic gene-enriched subgroup is associated with worse outcomes. The newly established risk model for prognosis and therapeutic responses may improve outcomes in patients with PanC.

Keywords: pancreatic cancer, transcriptome, metabolic genes, subtype, risk model

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Qian Chen,
Guangxi Medical University Cancer Hospital, China
Musalula Sinkala,
University of Cape Town, South Africa

****Correspondence:***
Xiaodong Tan
tanxdcmu@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *05 April 2021*
Accepted: *12 July 2021*
Published: *10 August 2021*

Citation:
Gao Y, Zhang E, Fei X, Kong L, Liu P and Tan X (2021) Identification of Novel Metabolism-Associated Subtypes for Pancreatic Cancer to Establish an Eighteen-Gene Risk Prediction Model. Front. Cell Dev. Biol. 9:691161. doi: 10.3389/fcell.2021.691161

INTRODUCTION

Pancreatic cancer (PanC) is the fourth most common cause of cancer-related deaths in the United States and accounts for over 227,000 deaths each year globally (Raimondi et al., 2009). The prevalence and incidence have increased in the past few years and this trend is expected to continue (Oshi et al., 2020). The etiology of PanC is complex and multiple risk factors have been reported,

including genetic factors, lifestyle factors, prior diseases, infections, and occupational exposure (Vincent et al., 2011). Universal treatment strategies fail to manage PanC owing to its heterogeneity at the level of molecular phenotypes and pathological and clinical properties (Makohon-Moore and Iacobuzio-Donahue, 2016). The complicated molecular characteristics render histopathology insufficient for clinical decisions and prognostic analyses (Collisson et al., 2019). The Cancer Genome Atlas (TCGA) Research Network completed integrated genomic, transcriptomic, and proteomic profiling of 150 pancreatic ductal adenocarcinoma specimens. They revealed a complex molecular landscape of pancreatic ductal adenocarcinoma and provided a roadmap for precision medicine (Cancer Genome Atlas Research Network, 2017). Two large researches using PanC samples recently reported gene expression subtypes of PanC, extending the subtypes previously described by Collisson et al. (2011), Bailey et al. (2016), Martinez-Useros et al. (2021). Sinkala et al. employed transcriptomic, copy number alteration and mutation profiling datasets from PanC patients together with data on clinical outcomes to show that the three PanC subtypes each display distinctive aberrations in cell signaling and metabolic pathways (Sinkala et al., 2018). Sinkala et al. also identified another two PanC subtypes and biomarker sets that can be used to accurately and sensitively classify novel pancreatic tumors (Sinkala et al., 2020).

Cancer metabolism plays an important role in oncogenesis and progression, and metabolic activity could be considered a cancer hallmark. There is compelling evidence that tumors reprogram cellular pathways for nutrient retrieval and consumption, allowing cancer cells to meet the demands of bioenergy generation, biosynthesis, and redox reactions (DeBerardinis and Chandel, 2016; Reina-Campos et al., 2017). Orchestrating adaptations of malignant cells in the hypoxic environment, cancer metabolism also presents heterogeneity and plasticity, even within the same malignancy, with differences among subpopulations, such as cell clusters with distinct rates of division (Birsoy et al., 2014; LeBleu et al., 2014; Yoshida, 2015). The dominant genetic alterations in pancreatic ductal adenocarcinoma are *KRAS* oncogene mutations (Buscail et al., 2020). In PanC cells, *KRAS* promotes extracellular glucose uptake by upregulating the glucose transporter GLUT1 and hexokinase (Ying et al., 2012). Furthermore, the hypoxic environment in PanC induces the stabilization of the transcription factor hypoxia-inducible factor 1, which promotes cancer cell adaptation to hypoxic conditions and is associated with an unfavorable prognosis (Jin et al., 2020). Metabolic mechanisms, involving micropinocytosis, in PanC can predict therapeutic efficacies (Commisso et al., 2013). Accordingly, studies of metabolic diversity and the identification of reliable prognostic markers for PanC are urgently needed.

Machine learning approaches can provide insights into disease diagnosis and treatment based on large amounts of data (Rajkomar et al., 2019). In PanC, Zhu et al. used a bioinformatics approach to identify *TMPRSS4*, *SERPINB5*, *SLC6A14*, *SCEL*, and *TNS4* as diagnostic biomarkers in PanC (Cheng et al., 2019). Machine learning provides depth and width in PanC investigations and may facilitate precision medicine (Herbst and Zheng, 2019). Another interesting study identified pancreatic ductal adenocarcinoma morphological subtypes using machine learning, providing new ideas for PanC management (Kalimuthu et al., 2020). In our study, we hypothesized that PanC could be divided into subtypes with distinct metabolic involvement, and genes related to these subtypes could be utilized to predict outcomes. In particular, we identified and characterized two subgroups with metabolic-enriched and metabolic-desert characteristics based on genes from Molecular Signatures Database v7.2. Finally, we trained and validated a risk model based on these metabolic classifications.

MATERIALS AND METHODS

Data Acquisition and Processing

Transcriptome profiles, simple nucleotide variant (SNV), and clinical data for 171 patients with PanC were downloaded from The Cancer Genome Atlas (TCGA)[1] (Blum et al., 2018). All analyses of TCGA data were performed using R 3.6.3. Pathological and clinical information is shown in **Table 1**. A risk prediction model was developed. Two additional datasets were

[1]https://portal.gdc.cancer.gov/

TABLE 1 | The disease-related clinical information of patients with pancreatic cancer included in the study.

Characteristics	Value
Patients(n)	171
Age(year), median(IQR)	65.7(57.2–73.3)
Gender, n(%)	
female	78(45.6%)
male	93(54.4%)
Pathological M, n(%)	
M0	77(45.0%)
M1	4(2.3%)
NA	90(52.7%)
Pathological T, n(%)	
T1 + T2	28(16.4%)
T3 + T4	141(82.5%)
NA	2(1.1%)
Pathological N, n(%)	
N0	47(27.5%)
N1 + N1b	119(69.6%)
NA	5(2.9%)
Pathological stage, n(%)	
Stage I	19(11.1%)
Stage II	142(83.0%)
Stage III + IV	7(4.1%)
NA	3(1.8%)
Outcome, n(%)	
Dead	91(46.8%)
Alive	80(53.2%)

IQR, interquantile range; NA, not analyzed.

TABLE 2 | Information of the two publicly available independent validation datasets.

Dataset	Sample size	Transcriptome platform	Tissue
GSE729	125	Agilent-014850 Whole Human Genome Microarray	Fresh frozen
E-MTAB-6134	288	Affymetrix Human Genome U219 Array	Fresh frozen and formalin-fixed paraffin-embedded

obtained from ArrayExpress[2] and Gene Expression Omnibus[3] to validate the effectiveness of the model. The details of these two datasets are provided in **Table 2**. One flow chart was displayed in **Figure 1** to summarize the process of this study.

Identification of Metabolism-Based Subtypes

To filter out genes not associated with prognosis in PanC, a global survival analysis for each gene was performed via log-rank test and Cox regression using the survival R package. Thereafter, metabolic genes were extracted from Molecular Signatures Database v7.2[4] (Liberzon et al., 2015). The overlap between survival-associated genes and metabolism-related genes was obtained. Finally, the expression levels of intersecting genes were used for a consensus clustering analysis to identify novel subtypes using the ConsensusClusterPlus R package. Default parameters were used (Wilkerson and Hayes, 2010). As determined using the ConsensusClusterPlus R package, the curve of the cumulative distribution function (CDF) and area under the CDF were used to choose the best k-value for the number of clusters. To evaluate differences between novel subtypes, a principal component analysis (PCA) was used. Next, survival curves were generated to evaluate the prognostic value of subtypes using the survminer and survival R packages. Expression levels of ten randomly selected genes used as inputs in the clustering analysis were examined to further characterize the different subtypes. The relationship (possible overlaps in patient samples) between the transcription subtypes that have been previously defined and the subtypes identified by us was shown via circle plot (Collisson et al., 2011; Moffitt et al., 2015; Bailey et al., 2016). Finally, tumor purity information acquired from TCGA Research Network was compared between two subtypes by *t*-test (Cancer Genome Atlas Research Network, 2017).

SNVs in the Subtypes

Using SNV data obtained from TCGA, the mutation frequencies of all genes were calculated. Thereafter, the ten genes with the highest mutation rates were utilized to evaluate the difference between subtypes using the GenVisR R package. Differences were visualized using waterfall plots. Next, genes with significant differences in SNV statuses were identified and the relationships between SNV statuses and transcript levels were determined. Finally, survival curves based on SNV statuses of these genes were generated using the survival and survminer R packages.

[2]https://www.ebi.ac.uk/arrayexpress/

[3]https://www.ncbi.nlm.nih.gov/geo/

[4]https://www.gsea-msigdb.org/gsea/msigdb

Weighted Correlation Network Analysis of Genes Related to Subtypes

Although novel subtypes were identified using the ConsensusClusterPlus package based on the expression levels of selected genes, this analysis did not reveal the precise genes related to each subtype and the sample size of genes was not sufficient for a network analysis. Therefore, a weighted correlation network analysis (WGCNA) was employed to identify gene co-expression modules related to the characteristics of the subtypes (Langfelder and Horvath, 2008) using the R package WGCNA. As recommended in the package instructions, the soft threshold was set to 7. For gene module fusion, the cutoff value was set to 0.25, as described previously (Kong et al., 2020).

Identification of Hub Genes Based on Subtypes

Based on the WGCNA, genes that were highly related to the high-risk subtype were obtained. These genes were used as inputs for an enrichment analysis using Metascape[5] (Kong et al., 2020) using default parameters. According to the annotations in Metascape, for each given gene list, protein-protein interaction enrichment analysis has been carried out, and if the network contains between 3 and 500 proteins, the Molecular Complex Detection (MCODE) algorithm would be applied to identify densely connected network components. Pathway and process enrichment analysis has been applied to each MCODE component independently, and the five best-scoring terms by p-value have been selected.

Enrichment Analyses of Subtypes

To further clarify the crucial processes activated in the high-risk subtype, a GSEA was performed. Hallmark gene sets were used as the background set (Liberzon et al., 2011). Fold change values were obtained for the comparison between high-risk and low-risk subtypes using the DESeq2 R package (Love et al., 2014). Next, the clusterProfiler R package was used to perform the GSEA (Yu et al., 2012). Furthermore, Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses directly based on the genes found up-regulated in C1 and also on the genes up-regulated in C2 were conducted via clusterProfiler R package (Yu et al., 2012).

Development of a Risk Prediction Model Based on Metabolism-Related Subtypes

A risk prediction model was developed based on the metabolism-associated subtypes. A group of genes associated with the high-risk subgroup was used to develop a risk model by the least

[5]https://metascape.org/

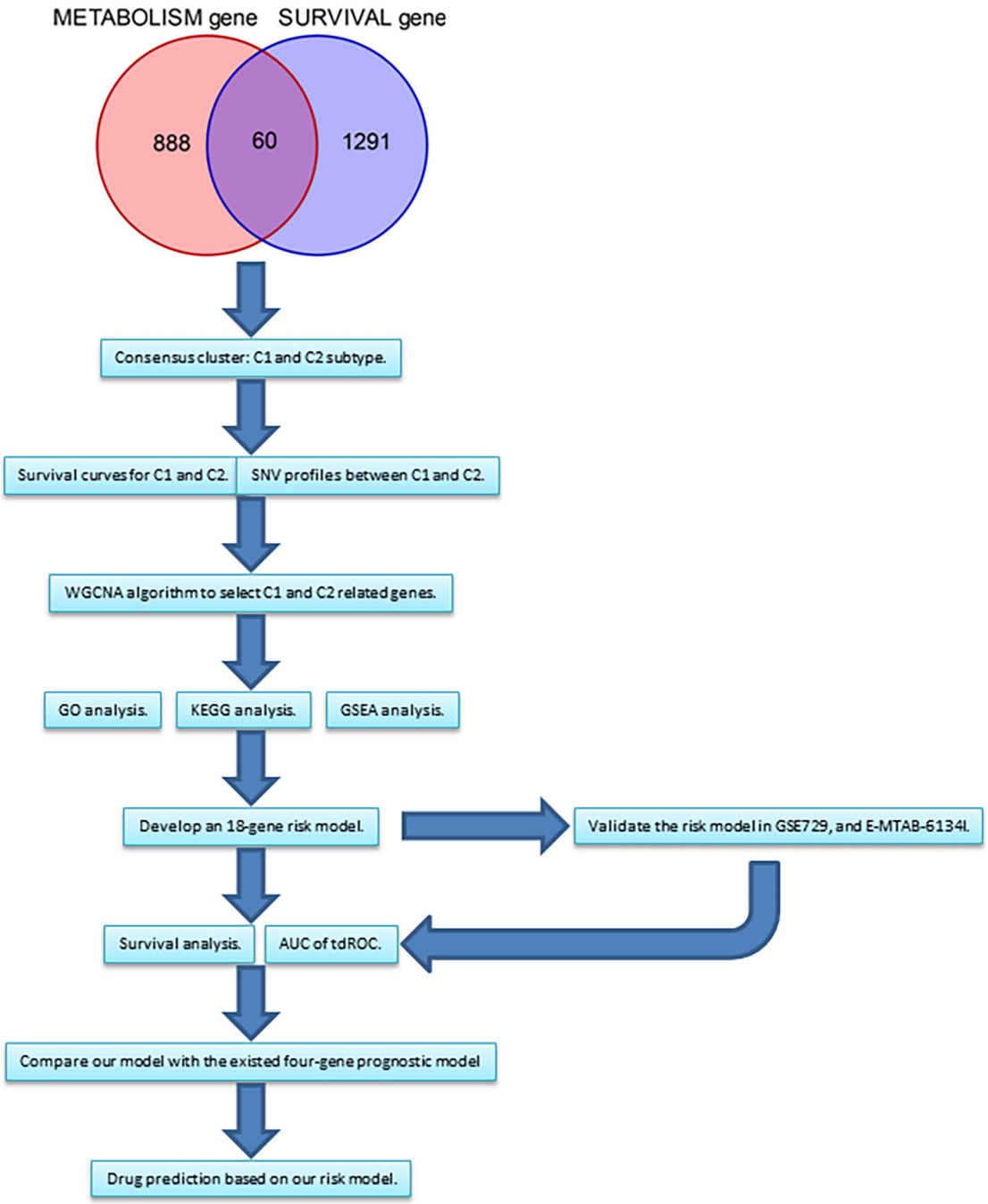


FIGURE 1 | The flow chart summarizing the process of this study. WGCNA, weighted correlation network analysis. GO, gene ontology. KEGG, Kyoto Encyclopedia of Genes and Genomes.

absolute shrinkage and selection operator (Lasso), a frequently adopted method (Zhang E. et al., 2020; Zhang H. et al., 2020; Zhang M. et al., 2020). The set of 171 patients from TCGA was divided into a training group (n = 120) and internal validation group (n = 51) using the caret R package. This package ensures that patients are equally divided into two distinct groups. Subsequently, data for patients in the training group were used to construct a risk model by Lasso. The regression analysis was performed using the glmnet R package with default parameter settings. The model was validated using the training group, internal validation group, and two additional external validation data sets. A survival analysis and time-dependent receiver operating characteristic (tdROC) curves were used as indicators for model effectiveness. Finally, the risk score of our

model was compared to that of an existed four-gene prognostic model in term of tdROC curves (Yan et al., 2020).

Univariate and Multivariate Cox Regression Analyses

To determine the independent prognostic value of the model, univariate and multivariate Cox regression analyses of the model and other clinical or pathological variables were performed. A univariate Cox regression was performed first. Statistically significant variables in the univariate analysis were included in the multivariate analysis, as described previously (Kong et al., 2020). These analyses were performed using the survival R package.

Drug Target Prediction

Using the CTRP2.0 and PRISM databases, the sensitivities of various agents for patients with high risk scores based on gene expression levels were predicted via ridge regression. The pRRophetic R package was used for prediction (Geeleher et al., 2014). Components with significantly lower areas under the dose–response curve (dr-AUC) in high-risk patients than in low-risk patients were selected. Thereafter, Spearman's correlation coefficients for the relationship between dr-AUC and risk score were determined. Components with significantly negative rho values (i.e., less than −0.3) were obtained.

Statistical Analysis

All statistical analyses were performed using R version 3.6.3. In a global survival of all genes, log-rank test and Cox regression were used, setting $P < 0.01$ as the cut-off value for significance. In all other analyses, $P < 0.05$ was the threshold for statistical significance.

RESULTS

Identification of Two Metabolism-Associated Subtypes

In total, 60 genes were included in a consensus clustering analysis (**Figure 2A**). As shown in **Figure 2B**, $k = 2$ was the optimal value. As shown in **Figure 2C**, the area under the CDF did not increase significantly for $k > 2$, further indicating that the optimal k-value was 2. **Figure 2D** shows the consensus matrix for two subtypes (referred to as C1 and C2). A PCA further supported the difference between the two subtypes (**Figure 2E**). A survival analysis indicated that patients with the C1 subtype had a worse prognosis than that of patients with the C2 subtype (**Figure 2F**). As shown in **Figures 2G–P**, the C1 subtype was highly enriched for metabolism-related genes. We found that the classical subtype identified by Collisson et al. was mainly a subset of C1 subtype (**Supplementary Figure 1A**) (Collisson et al., 2011). Furthermore, progenitor and squamous subtypes via Bailey et al. were mainly subsets of C1 subtype (**Supplementary Figure 1A**) (Bailey et al., 2016). Finally, high tumor purity might be the driver factor for the characteristics of C1 subtype (**Supplementary Figure 1B**, *t*-test $P < 0.001$).

Difference in SNVs Between Subtypes

We compared SNVs in the two subgroups. As shown in **Figures 3A,B**, the proportion of SNVs was higher in C1 than in C2. *KRAS*, *TP53*, *SMAD3*, and *CDKN2A* had high SNV frequencies. There were 69 patients harbored *KRAS* mutation in C1 subtype, and 10 patients harbored *KRAS* mutation in C2 subtype. There were 65 patients harbored *TP53* mutation in C1 subtype, and 17 patients harbored *TP53* mutation in C2 subtype. We found that there were lots of patients harbored *TTN* mutations in C1 and C2 subtypes. However, *TTN* was not significantly mutated based on the MutSigCV algorithm on Firebrowse database[6] (Broad Institute TCGA Genome Data Analysis Center, 2016). We found that patients with SNVs of *KRAS*, *SMAD3*, and *CDKN2A* exhibited a worse overall survival (**Figures 3C–F**). Finally, we found that patients with SNVs in *KRAS* and *CDKN2A* exhibited higher gene expression levels than those in patients without SNVs. However, the SNV in *SMAD4* was correlated with a lower level of gene expression. SNVs in *TP53* did not show a significant relationship with the level of gene expression (**Figures 3G–J**). Furthermore, *KRAS* expression was significantly higher in the C1 subtype than in the C2 subtype (**Figure 3K**, Wilcoxon test $P < 0.001$). *SMAD4* expression was significantly lower in the C1 subtype than in the C2 subtype (**Figure 3M**, Wilcoxon test $P < 0.001$). However, there was no significant difference for levels of TP53 and CDKN2A between C1 and C2 subtypes (**Figures 3L,N**). Generally, alterations in gene expression levels in the C1 subtype could be explained by the SNV status.

Identification of Gene Networks Related to the C1 and C2 Subtypes by WGCNA

As shown in **Figure 4A**, when the soft threshold was set to 7, the gene networks satisfied both a high degree of internal connectivity and high gene similarity. The top 5000 genes with a high degree of variation were divided into gene networks. Gene networks whose genes showed no significant differences were merged and finally 13 networks were identified (**Figure 4B**). Next, correlations between networks and phenotypes were evaluated. The blue network shown in **Figure 4C** was most closely related to the C1 subtype and the turquoise network was most related to the C2 subtype. The significance (i.e., the extent to which it represents the corresponding phenotype) and module membership of genes in the blue network are shown in **Figure 4D**, and these parameter values exhibited a strong positive correlation.

Enrichment Analysis and Hub Gene Selection

Owing to the worse prognosis of the C1 subtype, we focused on the network that was most strongly correlated with this subtype (i.e., the blue network in **Figure 4**). We extracted genes in this network and used them as inputs for an analysis using the Metascape database. As determined via an enrichment analysis, the downregulation

[6]http://firebrowse.org

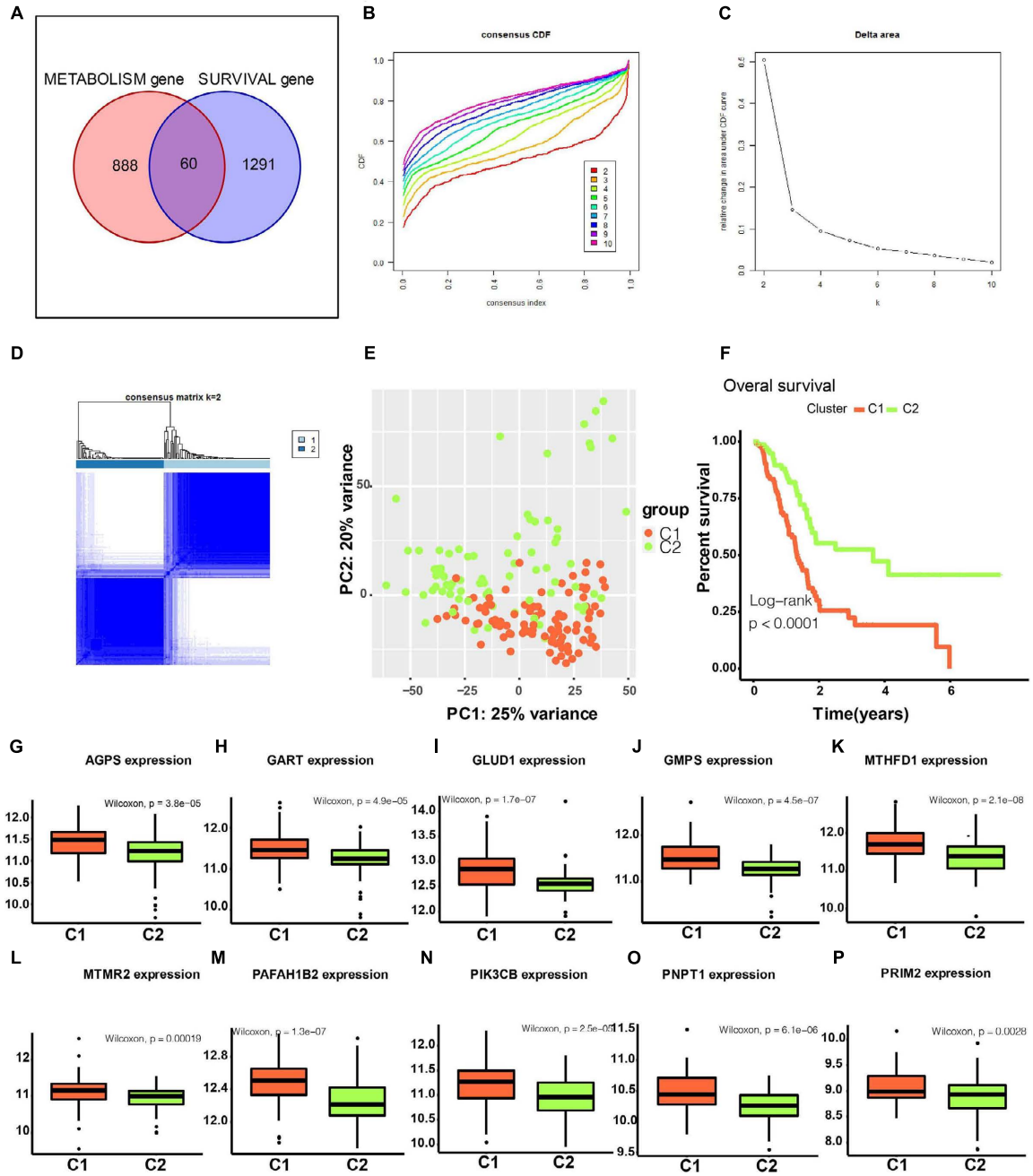


FIGURE 2 | Identification of two metabolism-associated subtypes. **(A)** In total, 60 genes were identified in the intersection between survival-associated genes and metabolism-related genes. **(B)** CDF curve for different k-values (representing the number of clusters in the consensus clustering analysis). When the optimal k-value is reached, the area under the CDF curve does not increase significantly as the k-value increases. **(C)** Relative change in the area under the CDF curve for different values of k. **(D)** Consensus matrix obtained when $k = 2$. Consistency values range from 0 to 1, where 0 indicates that genes never cluster together (white) and 1 indicates that they always cluster together (dark blue). **(E)** PCA showed a significant difference between subtypes. **(F)** Survival curves for patients with different subtypes. **(G–P)** Boxplot depicting expression levels of some metabolic genes in the C1 and C2 subtypes. CDF, cumulative distribution function. PCA, principal components analysis. $P < 0.05$ is defined as statistically significant.

of P53 signaling and activation of KRAS, AKT, and MEK signaling were enriched in the blue network, suggesting that these biological activities were likely associated with the C1 subtype (**Figures 5A,B**). Thereafter, hub genes were selected using the MCODE algorithm. As shown in **Figure 5C**, five sub-networks were selected. *KRT18* was identified as a hub gene with the potential to affect the development of PanC (**Figure 5D**).

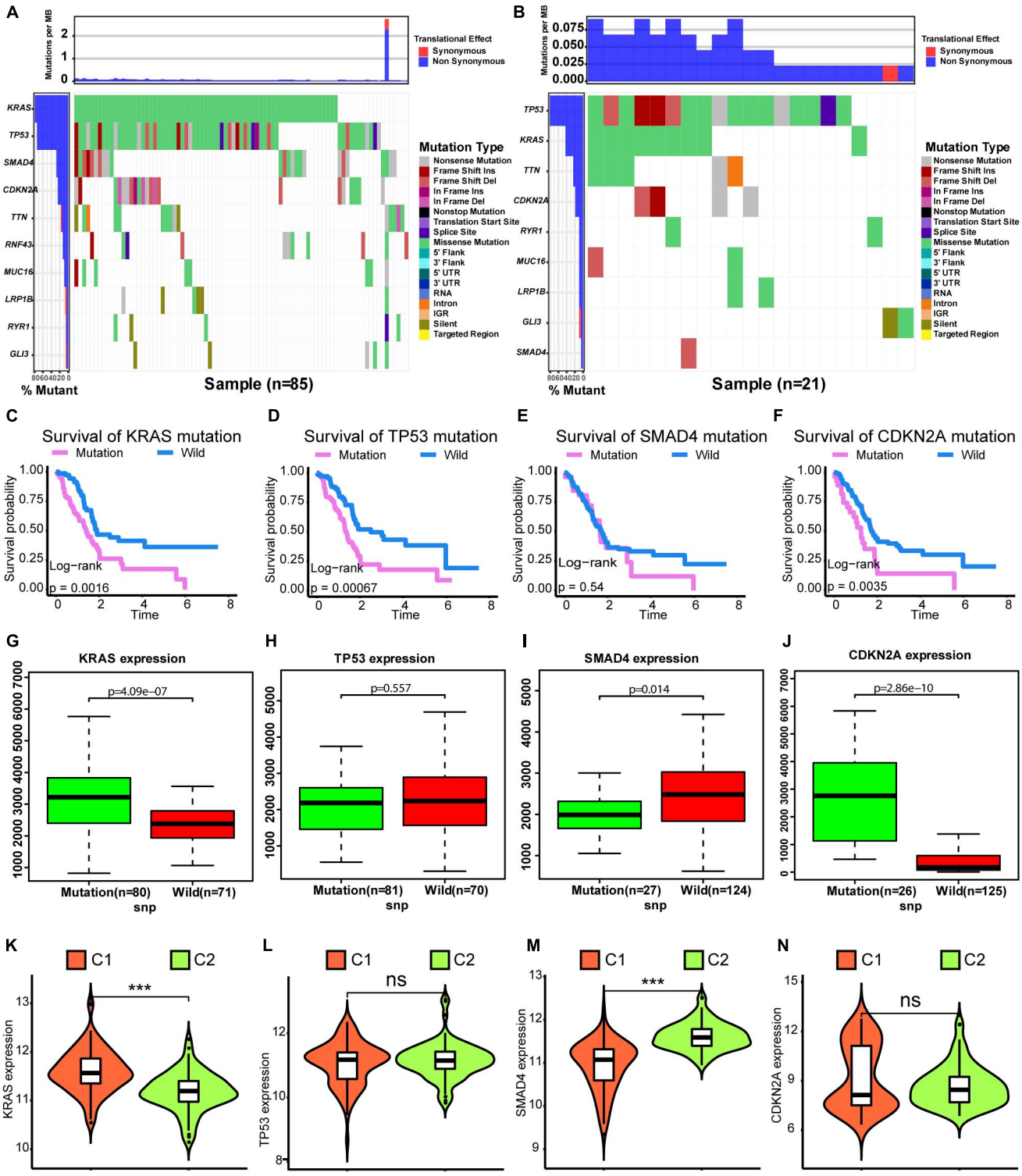


FIGURE 3 | Differences in SNVs between C1 and C2 subtypes. **(A)** Waterfall plot for the C1 subtype. **(B)** Waterfall plot for the C2 subtype. **(C–F)** Kaplan–Meier survival analysis of patients with mutations in *KRAS*, *TP53*, *SMAD4*, and *CDKN2A*. *P*-values obtained via log-rank tests are shown. **(G–J)** Boxplots illustrating correlations between gene expression levels and SNVs for *KRAS*, *TP53*, *SMAD4*, and *CDKN2A*. **(K–N)** Violin plots exhibit differences in expression levels of *KRAS*, *TP53*, *SMAD4*, and *CDKN2A* between C1 and C2 subtypes. Wilcoxon test *P*-values are shown. SNV: simple nucleotide variation. $P < 0.05$ was defined as statistically significant.

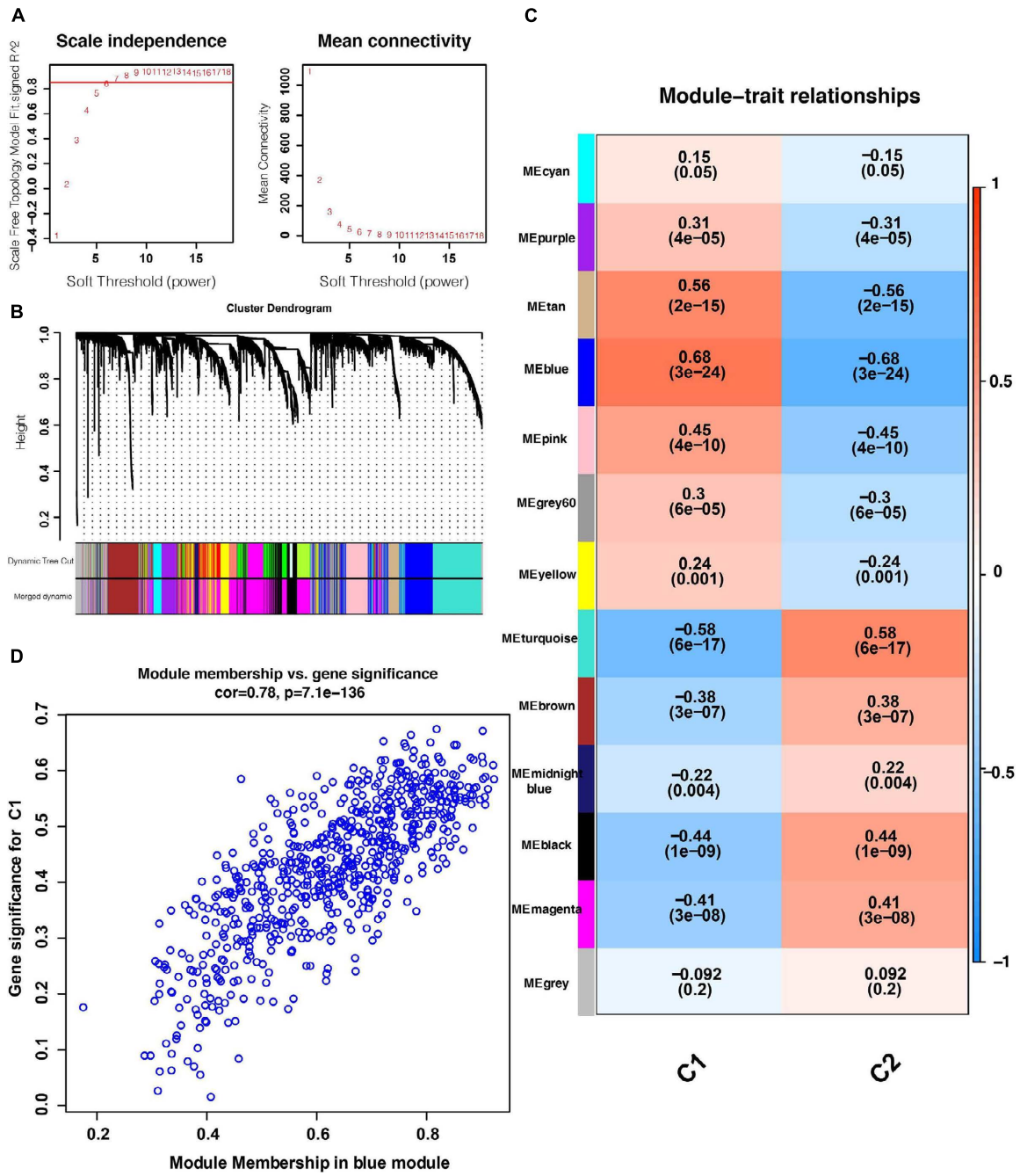


FIGURE 4 | WGCNA results. **(A)** Relationship between the soft threshold and TOM-based dissimilarity (left). Relationship between the soft threshold and mean connectivity (right). **(B)** After cutting and merging, 13 gene modules were finally generated. **(C)** Heat map of the correlations between gene modules and phenotypes. **(D)** Scatter plot depicting the correlations between gene significance and module membership of genes in the blue network. WGCNA: weighted correlation network analysis. TOM, topological overlap matrix. $P < 0.05$ is defined as statistically significant.

GSEA of the High-Risk Subtype

We selected statistically significant gene sets in the GSEA and ranked these sets according to normalized enrichment scores (NES). The top five results are displayed in **Figures 6A–F**. We found that HALLMARK_E2F_TARGETS, HALLMARK_MYC_TARGETS_V1, HALLMARK_GLYCOLYSIS, HALLMARK_MTORC1_SIGNALING, and HALLMARK_NOTCH_SIGNALING were all activated in the C1 subtype. These biological activities were all cancer-promoting processes, consistent with the high-risk feature of the C1 subtype (Sancho et al., 2015; Di Malta et al., 2017; Qin et al., 2019; Katoh and Katoh, 2020; Oshi et al., 2020). Furthermore,

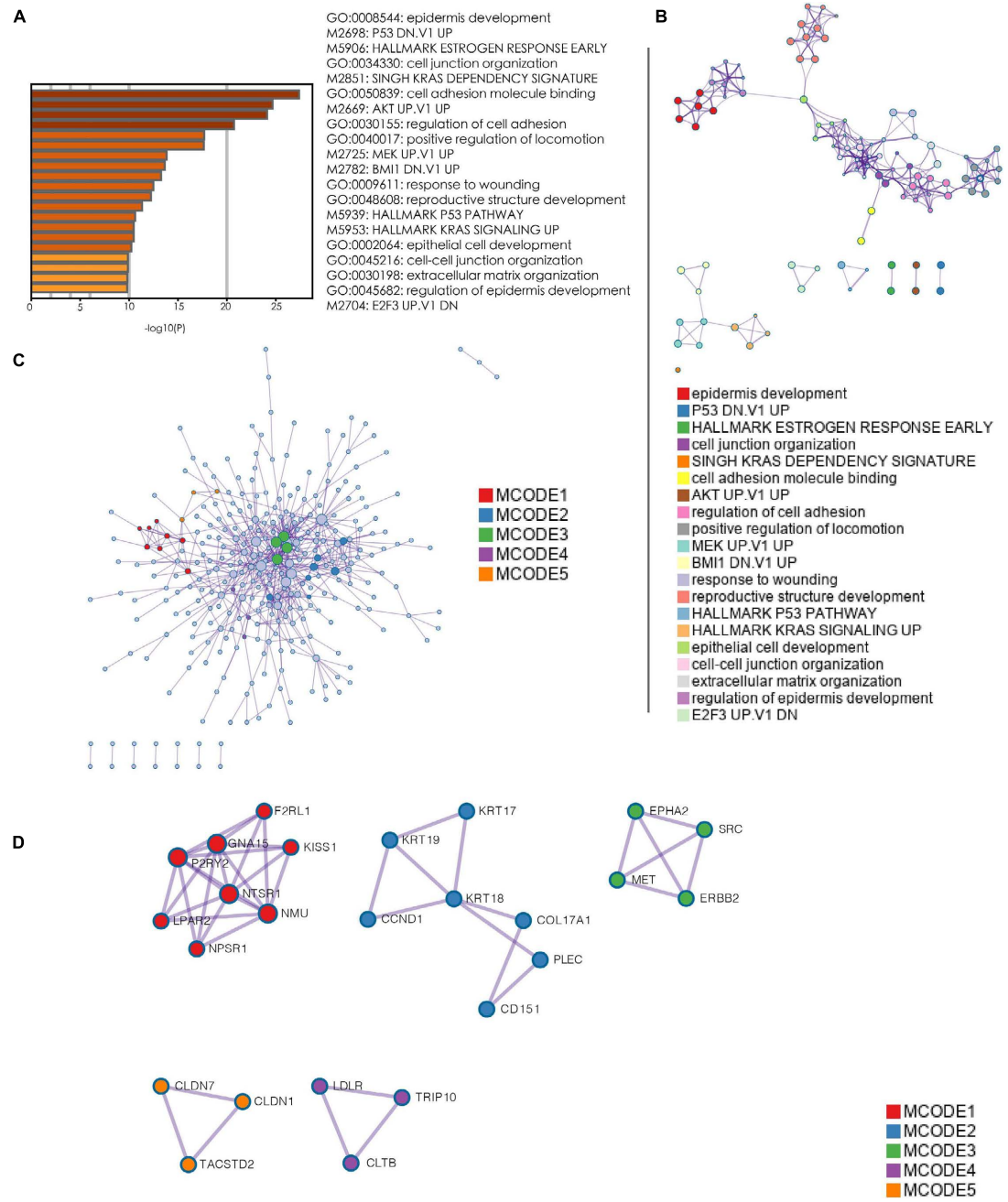


FIGURE 5 | Enrichment analysis and hub gene selection. **(A)** Bar plot depicting the results of an enrichment analysis of genes in the blue module. **(B)** Interaction network with enrichment terms. **(C)** Protein–protein interaction networks for genes generated using MCODE. **(D)** Five key protein–protein interaction networks identified using MCODE. (MCODE: molecular complex detection).

the GO and KEGG enrichment results were showed in **Supplementary Figure 2**.

Construction and Validation of an 18-Gene Risk Model

Due to patients in C1 subtype had the worst prognosis, we used the genes in blue network, which is most related to C1 subtype, to train the risk model via Lasso regression using the training group. During the selection of genes for the model, the C-index obtained by cross-validation was used for feature selection. **Figure 7A** showed the corresponding C-index values for models with different gene combinations. The X-axis is the number of genes in the corresponding model. As shown in **Figure 7A**, a combination of 18 genes showed the highest C-index. **Figure 7B** shows the coefficients for each gene and combination.

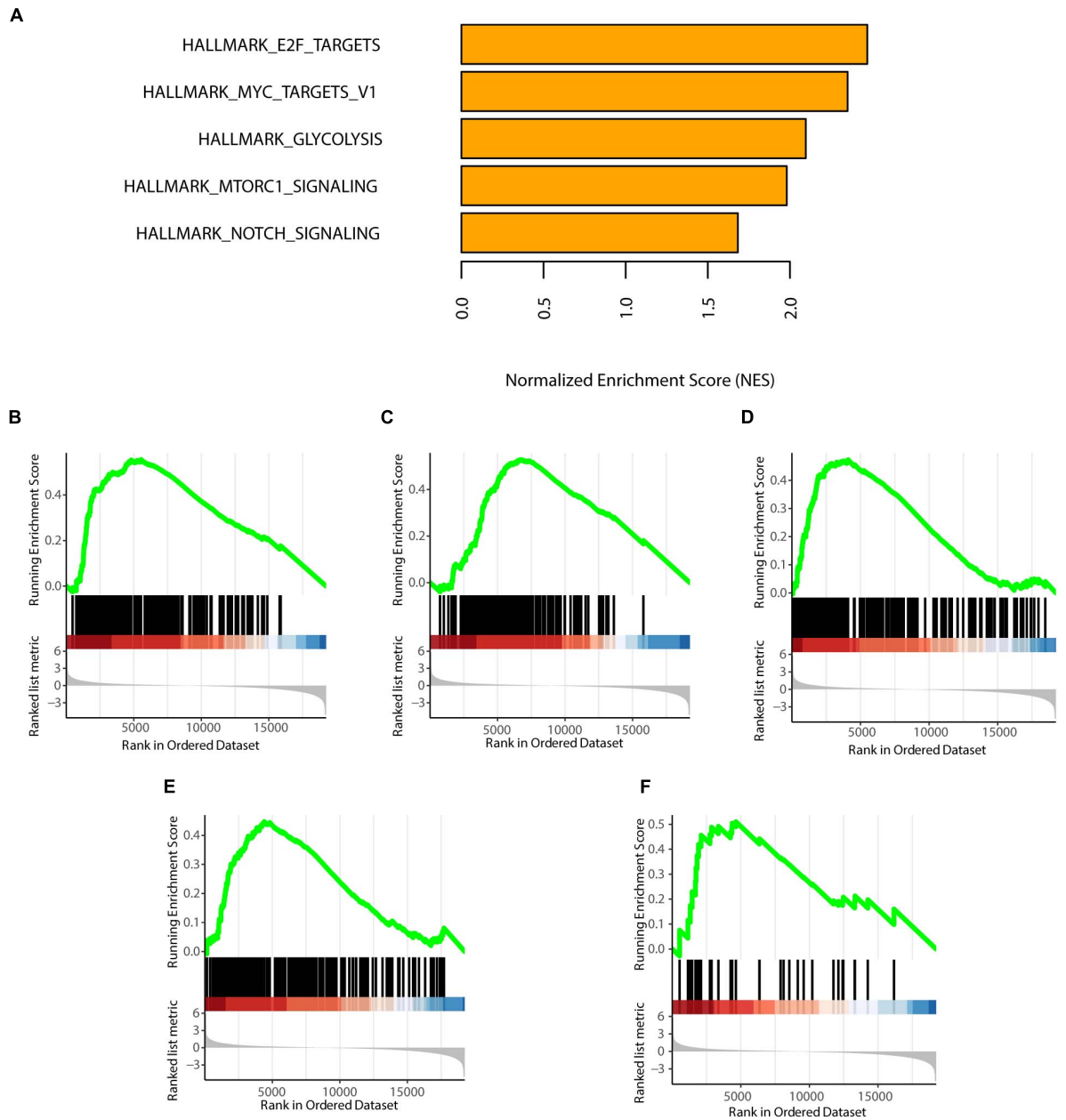


FIGURE 6 | GSEA of the two subtypes. **(A)** Top five enrichment terms (ranked in descending order of NES). **(B)** HALLMARK_E2F_TARGETS. **(C)** HALLMARK_MYC_TARGETS_V1. **(D)** HALLMARK_GLYCOLYSIS. **(E)** HALLMARK_MTORC1_SIGNALING. **(F)** HALLMARK_NOTCH_SIGNALING. NES: normalized enrichment scores. $P < 0.05$ was defined as statistically significant.

Finally, we developed an 18-gene risk model, providing a basis for determining risk levels of patients as follows: Risk score = NEU2 × 0.390 + GMPS × 0.350 + PRIM2 × 0.144 + PNP T1 × 0.154 + LDHA × 0.0004 + INPP4B × 0.011 + DPYD × 0.187 + PYGL × 0.363 + CA12 × 0.098 + DHRS9 × 0.080 + SULT1E1 × 0.158 - ENPP2 × 0.039 - PDE1C × 0.297 - TPH1 × 0.042 - CHST12 × 0.145 - POLR3GL × 0.421 - DNMT3A × 0.456 - PGS1 × 0.267. Next, patients in the training group (n = 120), internal validation group (n = 51), GSE729 data set (n = 125), and E-MTAB-6134 data set (n = 288) were ranked according to risk scores (**Figures 7C–N**). The expression levels of these 18 genes were globally evaluated in the four groups and the stability of the results were robust. Thereafter, survival and tdROC analyses were used to evaluate the effectiveness of the risk model in four data sets. As shown in **Figures 8A–D**, patients with higher risk scores had a worse overall survival. In the tdROC, the AUC values all exceeded 0.65, indicating good model performance (**Figures 8E–H**). Finally as shown in **Figure 8I**, we found that our

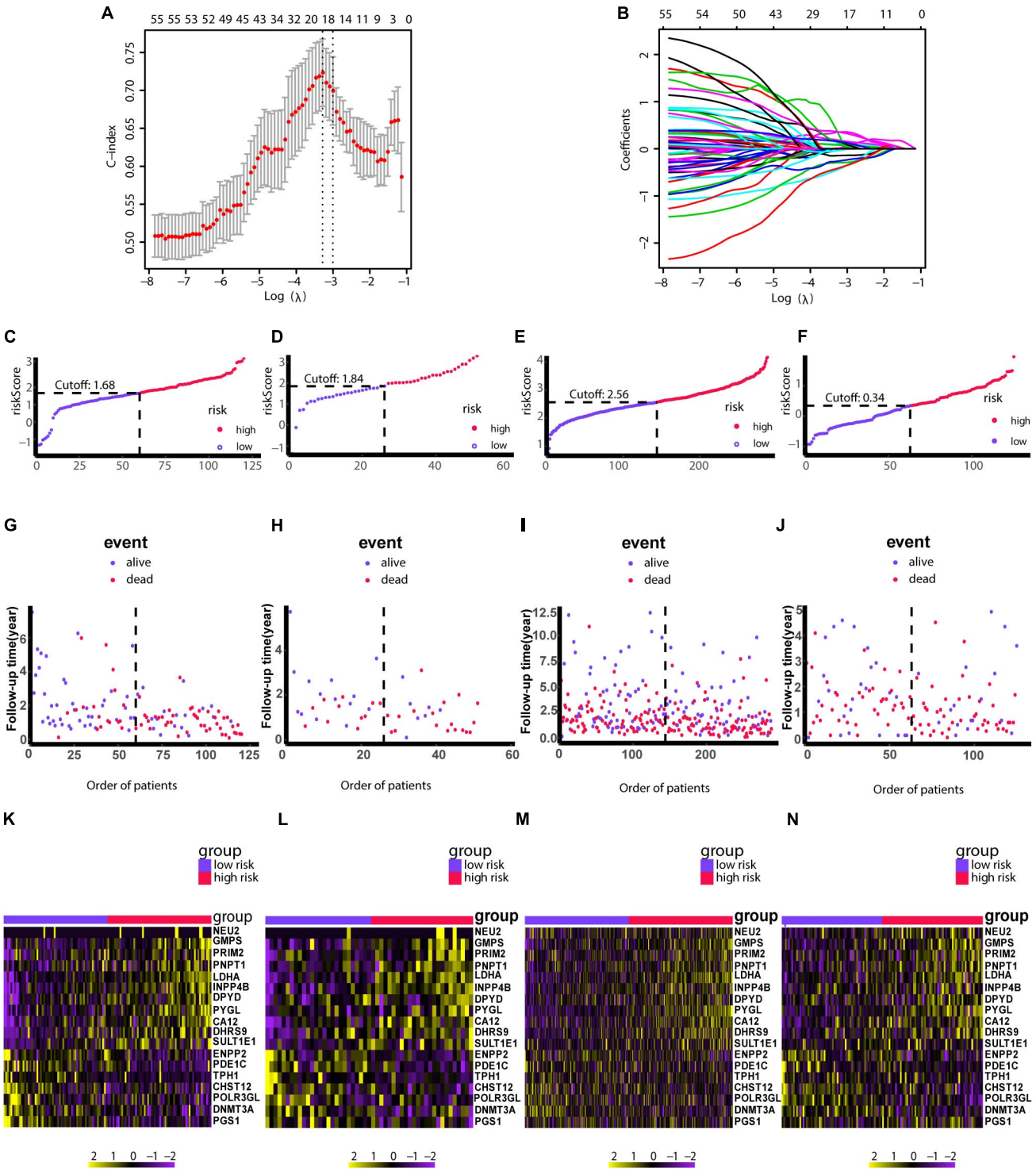


FIGURE 7 | Construction of the risk model using LASSO. **(A)** Cross-validation based on the C-index to determine the optimal number of genes in the model. **(B)** Genes in different combinations and corresponding coefficients. **(C,G,K)** Patients in the training set (n = 120) were arranged in ascending order of risk scores. **(D,H,L)** Patients in the internal validation set (n = 51) were arranged in ascending order of risk scores. **(E,I,M)** Patients in the GSE729 data set (n = 125) were arranged in ascending order of risk scores. **(F,J,N)** Patients in the E-MTAB-6134 data set (n = 288) were arranged in ascending order of risk scores. **(C–F)** Patients were divided into different risk levels according to median risk scores in their respective data sets. **(G–J)** Relationships between survival outcomes and risk levels. Low-risk patients are shown on the left side of the dotted line and high-risk patients are shown on the right side. **(K–N)** Heat maps for genes in the model. LASSO, least absolute shrinkage and selection operator. $P < 0.05$ was defined as statistically significant.

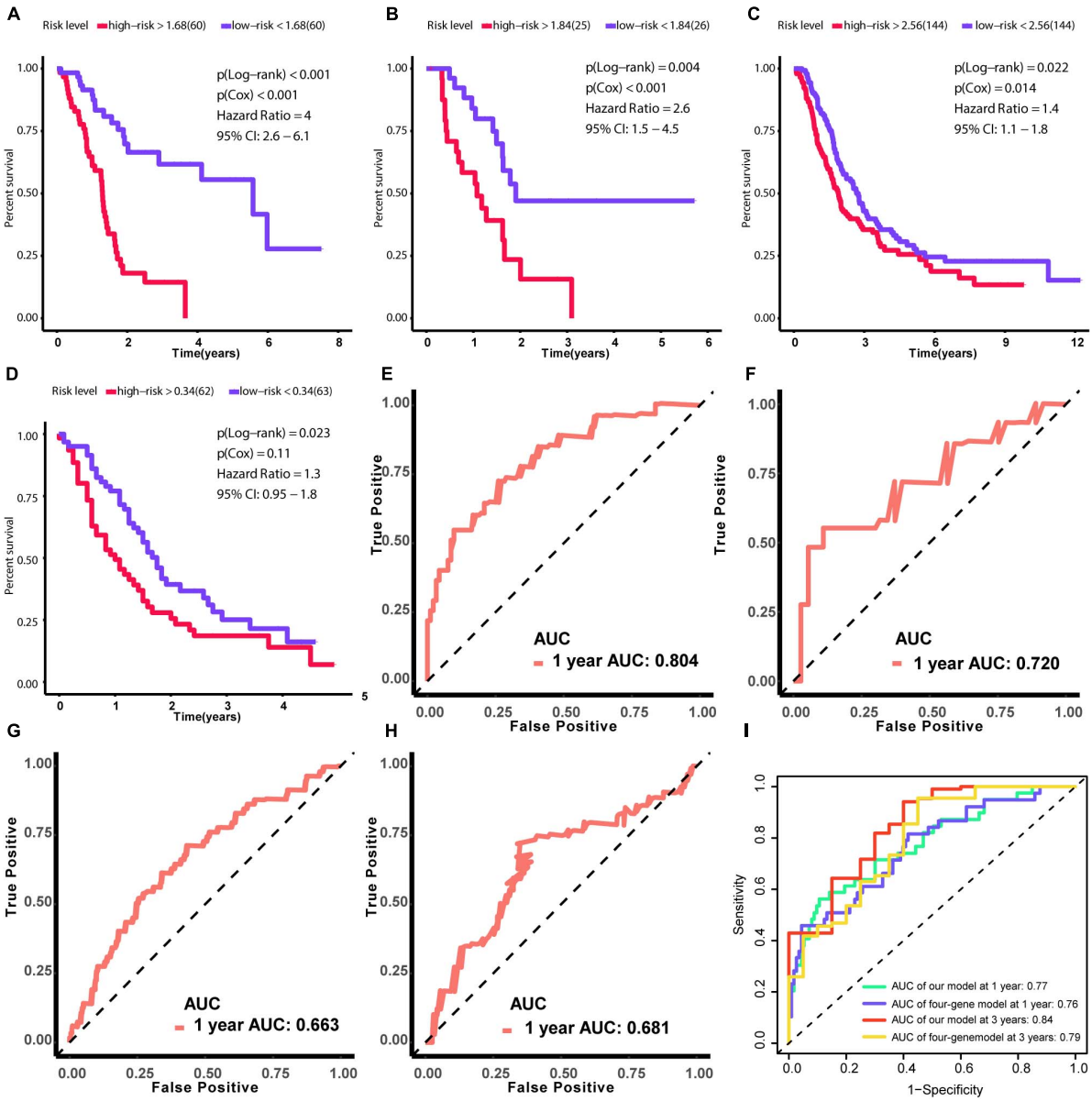


FIGURE 8 | Verification of the effectiveness of the prognostic signature. **(A–D)** Survival analysis based on Kaplan–Meier curves. **(E–H)** ROC curve for 1-year follow-up. **(A,E)** Results obtained using the training set. **(B,F)** Results obtained using the internal validation set. **(C,G)** Results obtained using GSE729. **(D,H)** Results obtained using E-MTAB-6134. **(I)** The comparison between our model and the existed four-gene model in term of tdROC curves. (AUC, area under curve; tdROC, time-dependent receiver operating characteristic). The clinical outcome endpoint was overall survival. $P < 0.05$ was defined as statistically significant.

model performed better than the existed four-gene prognostic model in TCGA-PAAD cohort ($n = 171$) (AUC of our model at 1 year: 0.77; AUC of four-gene model at 1 year: 0.76; AUC of our model at 3 years: 0.84; AUC of four-gene model at 3 years: 0.79).

Prognostic Value of the Model

To determine whether the model is an independent predictor of prognosis, we performed univariate and multivariate Cox regression analyses. In the univariate Cox regression analysis, metabolism-associated subtype, pathological T, pathological N, age, and risk score from the model were risk factors for prognosis (**Figure 9A**). Next, these variables were included in a multivariate Cox regression analysis, which showed that age and risk score are independent predictors of prognosis (**Figure 9B**).

Target Drug Prediction for High-Risk Patients

As shown in **Figure 10A**, afatinib, dasatinib, paclitaxel, pluripotin, and saracatinib were predicted with high sensitivity in patients with high risk scores (Spearman correlation test

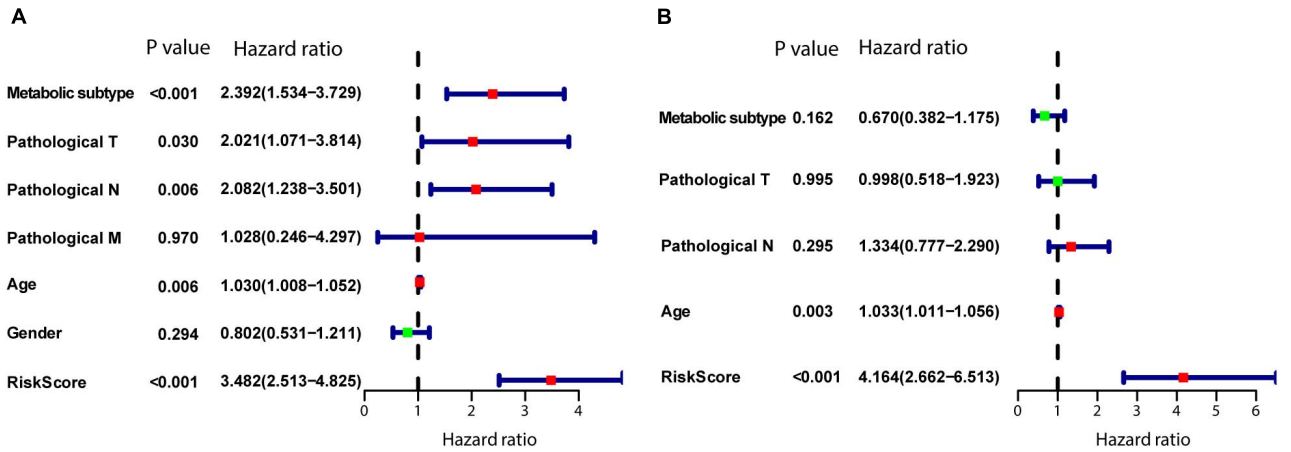


FIGURE 9 | Establishment of the model as an independent prognostic factor. **(A)** Univariate Cox regression. **(B)** Multivariate Cox regression. $P < 0.05$ was defined as statistically significant.

rho < -0.3, Spearman correlation test $P < 0.001$, and Wilcoxon test $P < 0.001$). As shown in **Figure 10B**, AZD8330, ispinesib, LY2606368, and trametinib showed high sensitivity for patients with high risk scores (Spearman correlation test rho < –0.3, Spearman correlation test $P < 0.001$, and Wilcoxon test $P < 0.01$). Collectively, we predicted nine target drugs for high-risk patients; paclitaxel, AZD8330, and LY2606368 showed the highest effectiveness.

DISCUSSION

Recent studies have offered ample insight into multi-omics landscapes to define PanC subtypes (Noll et al., 2016; Mishra and Guda, 2017; Puleo et al., 2018; Lomberk et al., 2018). Lomberk et al. performed multi-parametric integrative analyses of chromatin immunoprecipitation-sequencing on multiple histone modifications, RNA-seq, and DNA methylation to define epigenomic landscapes for PanC subtypes, which can predict their relative aggressiveness and survival. Puleo et al. identified five PanC subtypes, based on features of cancer cells and the tumor microenvironment. Kumar Mishra et al. performed genome-scale methylome analysis of PanC data from TCGA and provided a strong basis for future work on the molecular subtyping of epigenetic regulation in PanC. Using data obtained from multiple platforms (TCGA, GSE729, and E-MTAB-6134), we identified novel PanC subtypes and prognostic signatures based on cancer metabolism. Patients with PanC from the TCGA dataset were divided into metabolic-enriched and metabolic-desert groups, C1 and C2, respectively. Multiple metabolic genes (*AGPS*, *GART*, *GLUD1*, *GMPS*, *MTHFD1*, *MTMR2*, *PAFAH1B2*, *PIK3CB*, *PNPT1*, and *PRIM2*) related to aggressiveness and poor outcomes in cancer showed higher expression levels in the C1 group than in the C2 group. Alkylglyceronephosphate synthase (AGPS), an essential metabolic enzyme in cancer, is involved in the synthesis of ether lipids and promotes the aggressive features of multiple malignances (Benjamin et al., 2013). Widely recognized as a prognostic factor in many cancers, glucose transporter-1 (GLUT1) is associated with both a poor prognosis and drug resistance in cancers, such as PanC, cervical cancer, and breast cancer (Nagarajan et al., 2017; He et al., 2019; Kim and Chang, 2019). Other metabolic genes with notably higher expression in C1 than in C2 are also related to tumorigenesis and shown prognostic value (Knox et al., 2009; Sdelci et al., 2019; Wang et al., 2019). Consistent with these findings, the survival rate in the metabolic-enriched group was significantly lower than that in the metabolic-desert group. The evaluation of the SNV profile of the novel two subtypes further demonstrated that *KRAS*, *TP53*, *SMAD4*, and *CDKN2A* are frequently mutated in the C1 subtype. According to Terumi et al., *KRAS*, *TP53*, *SMAD3*, and *CDKN2A* are the key driver genes in PanC (Kamisawa et al., 2016). *KRAS* mutations resulting in the persistent activation of the protein contribute to PanC (Buscail et al., 2020). The KRAS protein could activate several signaling pathways to promote the proliferation, invasion, and migration of cancer cells (Buscail et al., 2020). As a hallmark oncogene, *KRAS* upregulates glycolysis-related enzymes and promotes energy supply in cancer cells, indicating a worse survival (Buscail et al., 2020). Furthermore, *TP53*, *SMAD4,* and *CDKN2A* are frequently mutated genes in malignancies (Alhejaily et al., 2014; Bykov et al., 2018; Ritterhouse et al., 2019). The *KRAS* mutation frequency was significantly higher in the C1 group than in the C2 group, and *KRAS* mutations were associated with increased expression. These findings support the reliability of the novel metabolic subtypes and the association between the metabolism-related gene-enriched group (C1) and a poor prognosis.

We explored the core differences between the subtypes. Using a GSEA, we identified five hallmark gene sets, HALLMARK_E2F_TARGETS, HALLMARK_MYC_TARGETS_V1, HALLMARK_GLYCOLYSIS, HALLMARK_MTORC1_SIGNALING, and HALLMARK_NOTCH_SIGNALING, activated in the C1 (high risk) group. Repressed by the tumor suppressor gene *RB1* (retinoblastoma susceptibility gene) in RB-E2F complexes, E2F1 induces cell cycle entry and could be a cancer inducer (Fischer and Müller, 2017). Amplification of the *MYC* gene is related to poor outcomes in PanC

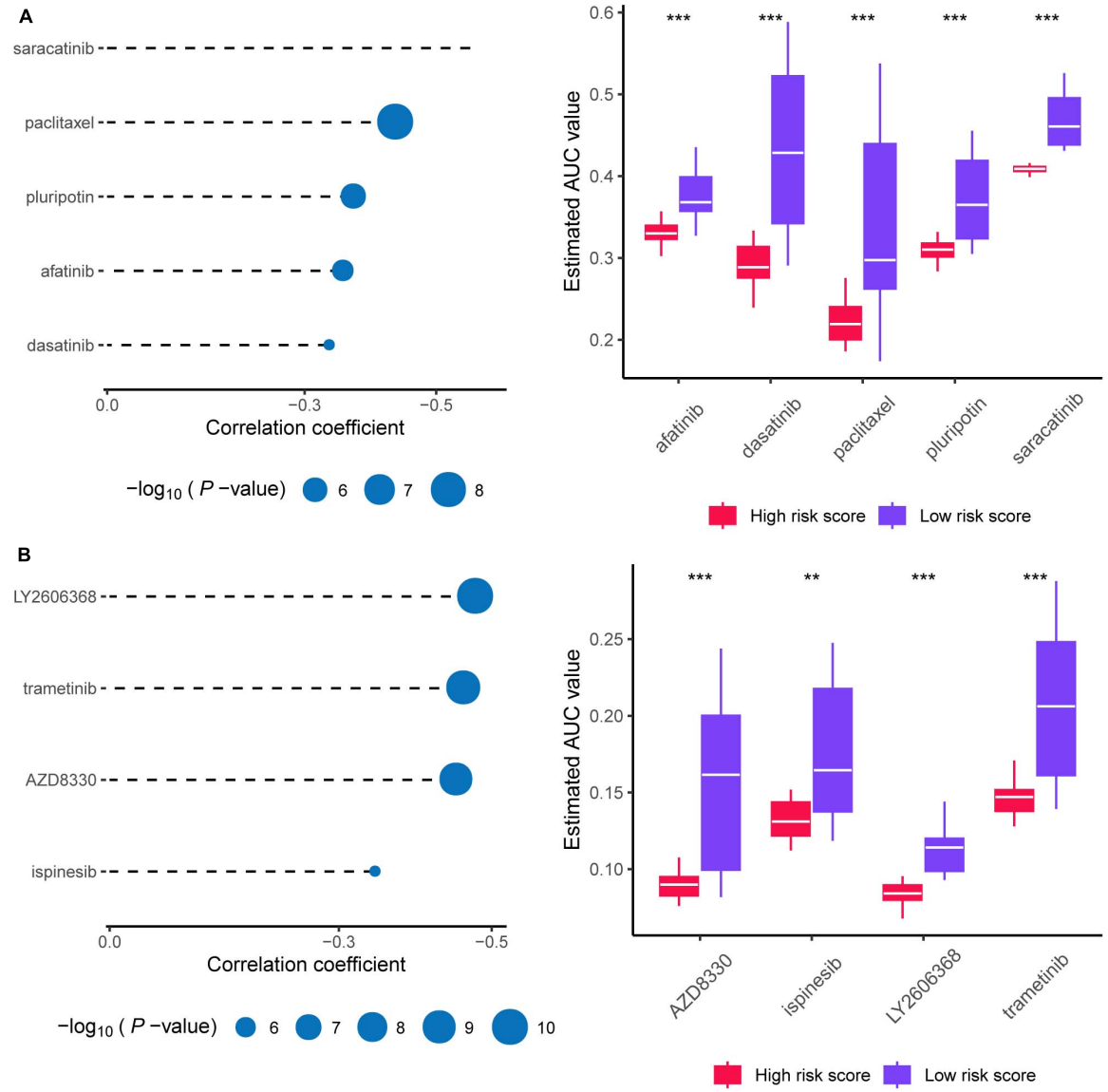


FIGURE 10 | Prediction of candidate agents with higher drug sensitivity in patients with high risk score. **(A)** The results of Spearman's correlation analysis and differential drug response analysis of three CTRP-derived compounds. **(B)** The results of Spearman's correlation analysis and differential drug response analysis of three PRISM-derived compounds. Note that lower values on the y-axis of boxplots imply greater drug sensitivity. * means $P < 0.05$, ** means $P < 0.01$, *** means $P < 0.001$, ns means $P > 0.05$, and $P < 0.05$ is defined as statistically significant.

(Witkiewicz et al., 2015). Additionally, mTORC signaling, glycolysis-related genes, and NOTCH signaling all contribute to the pathogenesis of PanC (Mann et al., 2016; Gao et al., 2017; Yang et al., 2020). And based on network system biology we select one target KRT18 associated with poor prognosis in PanC. KRT18 could induce proliferation, migration, and invasion via MAKP signaling in gastric cancer (Wang et al., 2020). However, the mechanism by which *KRT18* contributes to PanC has not been determined and should be a focus of further research.

Pancreatic cancer (PanC) is a highly lethal malignancy with high complexity and heterogeneity (Yao et al., 2020). The Cancer Genome Atlas Research Network completed integrated genomic, transcriptomic, and proteomic profiling of 150 pancreatic ductal adenocarcinoma specimens. They revealed a complex molecular landscape of pancreatic ductal adenocarcinoma and provided a roadmap for precision medicine (Cancer Genome Atlas Research Network, 2017). Two large researches using PanC samples recently reported gene expression subtypes of PanC, extending the subtypes previously described by Collisson et al. (Collisson et al., 2011; Moffitt et al., 2015; Bailey et al., 2016; Martinez-Useros et al., 2021). Sinkala et al. employed transcriptomic, copy number alteration and mutation profiling datasets from PanC patients together with data on clinical outcomes to show that the three PanC subtypes each display distinctive

aberrations in cell signaling and metabolic pathways (Sinkala et al., 2018). Sinkala et al. also identified another two PanC subtypes and biomarker sets that can be used to accurately and sensitively classify novel pancreatic tumors (Sinkala et al., 2020). Inspired by previous researches, we developed novel risk model and metabolic subtypes to deepen our understanding of PanC and to facilitate individualized diagnosis and treatment in clinical settings.

This study had several limitations. First, despite our use of data for large patient cohorts obtained from multiple platforms supporting the reliability and reproducibility of our results, prospective studies are needed before applying the risk model in clinical settings. Additionally, our risk score formula is based on 18 metabolic genes; however, further studies on interactions among these genes and their biological mechanisms are needed.

CONCLUSION

In this study, we identified two subtypes, a metabolic gene-enriched subtype (C1) and metabolic gene-desert subtype (C2), in PanC based on the degree of metabolic gene enrichment. The C1 subtype was related to a worse prognosis based a survival analysis and distinct SNVs. We then developed a risk model based on genes associated with the high-risk subgroup (C1) and validated the reliability of the predictive model using an internal validation dataset of TCGA and two external validation datasets. Finally, nine drugs were identified for high-risk patients.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YG and XT were responsible for the design and conception of the research. XF, EZ, LK, PL, and YG contributed to data acquisition or data analysis, and data cleaning. XT and YG participated in the drafting of the manuscript and the rigorous modification of the manuscript to clearly convey the research contents. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Natural Science Foundation of China (grant number 81902953), Natural Science Foundation of Liaoning Province (grant number 180530068), and 345 Talent Project of Shengjing Hospital of China Medical University.

ACKNOWLEDGMENTS

We want to thank the Medical Research Center of Shengjing Hospital for providing us with laboratories. We would like to thank Editage for providing the English writing assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.691161/full#supplementary-material

REFERENCES

Alhejaily, A., Day, A. G., Feilotter, H. E., Baetz, T., and Lebrun, D. P. (2014). Inactivation of the CDKN2A tumor-suppressor gene by deletion or methylation is common at diagnosis in follicular lymphoma and associated with poor clinical outcome. *Clin. Cancer Res.* 20, 1676–1686. doi: 10.1158/1078-0432.Ccr-13-2175

Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M., Gingras, M.-C., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 531, 47–52. doi: 10.1038/nature16965

Benjamin, D. I., Cozzo, A., Ji, X., Roberts, L. S., Louie, S. M., Mulvihill, M. M., et al. (2013). Ether lipid generating enzyme AGPS alters the balance of structural and signaling lipids to fuel cancer pathogenicity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 14912–14917. doi: 10.1073/pnas.1310894110

Birsoy, K., Possemato, R., Lorbeer, F. K., Bayraktar, E. C., Thiru, P., Yucel, B., et al. (2014). Metabolic determinants of cancer cell sensitivity to glucose limitation and biguanides. *Nature* 508, 108–112. doi: 10.1038/nature13110

Blum, A., Wang, P., and Zenklusen, J. C. (2018). SnapShot: TCGA-analyzed tumors. *Cell* 173:530. doi: 10.1016/j.cell.2018.03.059

Broad Institute TCGA Genome Data Analysis Center (2016). *Mutation Analysis (MutSigCV v0.9): Broad Institute of MIT and Harvard*. Available online at: http://firebrowse.org.

Buscail, L., Bournet, B., and Cordelier, P. (2020). Role of oncogenic KRAS in the diagnosis, prognosis and treatment of pancreatic cancer. *Nat. Rev. Gastroenterol. Hepatol.* 17, 153–168. doi: 10.1038/s41575-019-0245-4

Bykov, V. J. N., Eriksson, S. E., Bianchi, J., and Wiman, K. G. (2018). Targeting mutant p53 for efficient cancer therapy. *Nat. Rev. Cancer* 18, 89–102. doi: 10.1038/nrc.2017.109

Cancer Genome Atlas Research Network (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 32, 185–203.e13. doi: 10.1016/j.ccell.2017.07.007

Cheng, Y., Wang, K., Geng, L., Sun, J., Xu, W., Liu, D., et al. (2019). Identification of candidate diagnostic and prognostic biomarkers for pancreatic carcinoma. *EBioMedicine* 40, 382–393. doi: 10.1016/j.ebiom.2019.01.003

Collisson, E. A., Bailey, P., Chang, D. K., and Biankin, A. V. (2019). Molecular subtypes of pancreatic cancer. *Nat. Rev. Gastroenterol. Hepatol.* 16, 207–220. doi: 10.1038/s41575-019-0109-y

Collisson, E. A., Sadanandam, A., Olson, P., Gibb, W. J., Truitt, M., Gu, S., et al. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* 17, 500–503. doi: 10.1038/nm.2344

Commisso, C., Davidson, S. M., Soydaner-Azeloglu, R. G., Parker, S. J., Kamphorst, J. J., Hackett, S., et al. (2013). Macropinocytosis of protein is an amino acid supply route in Ras-transformed cells. *Nature* 497, 633–637. doi: 10.1038/nature12138

DeBerardinis, R. J., and Chandel, N. S. (2016). Fundamentals of cancer metabolism. *Sci. Adv.* 2:e1600200. doi: 10.1126/sciadv.1600200

Di Malta, C., Siciliano, D., Calcagni, A., Monfregola, J., Punzi, S., Pastore, N., et al. (2017). Transcriptional activation of RagD GTPase controls

mTORC1 and promotes cancer growth. *Science* 356, 1188–1192. doi: 10.1126/science.aag2553

Fischer, M., and Müller, G. A. (2017). Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. *Crit. Rev. Biochem. Mol. Biol.* 52, 638–662. doi: 10.1080/10409238.2017.1360836

Gao, J., Long, B., and Wang, Z. (2017). Role of notch signaling pathway in pancreatic cancer. *Am. J. Cancer Res.* 7, 173–186.

Geeleher, P., Cox, N., and Huang, R. S. (2014). pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS One* 9:e107468. doi: 10.1371/journal.pone.0107468

He, Y., Deng, F., Zhao, S., Zhong, S., Zhao, J., Wang, D., et al. (2019). Analysis of miRNA-mRNA network reveals miR-140-5p as a suppressor of breast cancer glycolysis via targeting GLUT1. *Epigenomics* 11, 1021–1036. doi: 10.2217/epi-2019-0072

Herbst, B., and Zheng, L. (2019). Precision medicine in pancreatic cancer: treating every patient as an exception. *Lancet Gastroenterol. Hepatol.* 4, 805–810. doi: 10.1016/s2468-1253(19)30175-x

Jin, X., Dai, L., Ma, Y., Wang, J., and Liu, Z. (2020). Implications of HIF-1α in the tumorigenesis and progression of pancreatic cancer. *Cancer Cell Int.* 20:273. doi: 10.1186/s12935-020-01370-0

Kalimuthu, S. N., Wilson, G. W., Grant, R. C., Seto, M., O'Kane, G., Vajpeyi, R., et al. (2020). Morphological classification of pancreatic ductal adenocarcinoma that predicts molecular subtypes and correlates with clinical outcome. *Gut* 69, 317–328. doi: 10.1136/gutjnl-2019-318217

Kamisawa, T., Wood, L. D., Itoi, T., and Takaori, K. (2016). Pancreatic cancer. *Lancet* 388, 73–85. doi: 10.1016/S0140-6736(16)00141-0

Katoh, M., and Katoh, M. (2020). Precision medicine for human cancers with Notch signaling dysregulation (Review). *Int. J. Mol. Med.* 45, 279–297. doi: 10.3892/ijmm.2019.4418

Kim, B. H., and Chang, J. H. (2019). Differential effect of GLUT1 overexpression on survival and tumor immune microenvironment of human papilloma virus type 16-positive and -negative cervical cancer. *Sci. Rep.* 9:13301. doi: 10.1038/s41598-019-49928-x

Knox, A. J., Graham, C., Bleskan, J., Brodsky, G., and Patterson, D. (2009). Mutations in the Chinese hamster ovary cell GART gene of de novo purine synthesis. *Gene* 429, 23–30. doi: 10.1016/j.gene.2008.10.007

Kong, L., Liu, P., Fei, X., Wu, T., Wang, Z., Zhang, B., et al. (2020). A prognostic prediction model developed based on four CpG sites and weighted correlation network analysis identified as a novel biomarker for pancreatic cancer. *Front. Oncol.* 10:1716. doi: 10.3389/fonc.2020.01716

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

LeBleu, V. S., O'Connell, J. T., Gonzalez Herrera, K. N., Wikman, H., Pantel, K., Haigis, M. C., et al. (2014). PGC-1α mediates mitochondrial biogenesis and oxidative phosphorylation in cancer cells to promote metastasis. *Nat. Cell Biol.* 16, 992–1003,1–15. doi: 10.1038/ncb3039

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Lomberk, G., Blum, Y., Nicolle, R., Nair, A., Gaonkar, K. S., Marisa, L., et al. (2018). Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes. *Nat. Commun.* 9:1978. doi: 10.1038/s41467-018-04383-6

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Makohon-Moore, A., and Iacobuzio-Donahue, C. A. (2016). Pancreatic cancer biology and genetics from an evolutionary perspective. *Nat. Rev. Cancer* 16, 553–565. doi: 10.1038/nrc.2016.66

Mann, K. M., Ying, H., Juan, J., Jenkins, N. A., and Copeland, N. G. (2016). KRAS-related proteins in pancreatic cancer. *Pharmacol. Ther.* 168, 29–42. doi: 10.1016/j.pharmthera.2016.09.003

Martinez-Useros, J., Martin-Galan, M., and Garcia-Foncillas, J. (2021). The match between molecular subtypes, histology and microenvironment of pancreatic cancer and its relevance for chemoresistance. *Cancers (Basel)* 13:322. doi: 10.3390/cancers13020322

Mishra, N. K., and Guda, C. (2017). Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget* 8, 28990–29012. doi: 10.18632/oncotarget.15993

Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G. H., Hoadley, K. A., et al. (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* 47, 1168–1178. doi: 10.1038/ng.3398

Nagarajan, A., Dogra, S. K., Sun, L., Gandotra, N., Ho, T., Cai, G., et al. (2017). Paraoxonase 2 facilitates pancreatic cancer growth and metastasis by stimulating GLUT1-mediated glucose transport. *Mol. Cell* 67, 685–701.e6. doi: 10.1016/j.molcel.2017.07.014

Noll, E. M., Eisen, C., Stenzinger, A., Espinet, E., Muckenhuber, A., Klein, C., et al. (2016). CYP3A5 mediates basal and acquired therapy resistance in different subtypes of pancreatic ductal adenocarcinoma. *Nat. Med.* 22, 278–287. doi: 10.1038/nm.4038

Oshi, M., Tokumaru, Y., Patel, A., Yan, L., Matsuyama, R., Endo, I., et al. (2020). A novel four-gene score to predict pathologically complete (R0) resection and survival in pancreatic cancer. *Cancers (Basel)* 12:3635. doi: 10.3390/cancers12123635

Puleo, F., Nicolle, R., Blum, Y., Cros, J., Marisa, L., Demetter, P., et al. (2018). Stratification of pancreatic ductal adenocarcinomas based on tumor and microenvironment features. *Gastroenterology* 155, 1999–2013.e3. doi: 10.1053/j.gastro.2018.08.033

Qin, Y., Hu, Q., Xu, J., Ji, S., Dai, W., Liu, W., et al. (2019). PRMT5 enhances tumorigenicity and glycolysis in pancreatic cancer via the FBW7/cMyc axis. *Cell Commun. Signal.* 17, 30. doi: 10.1186/s12964-019-0344-4

Raimondi, S., Maisonneuve, P., and Lowenfels, A. B. (2009). Epidemiology of pancreatic cancer: an overview. *Nat. Rev. Gastroenterol. Hepatol.* 6, 699–708. doi: 10.1038/nrgastro.2009.177

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *N. Engl. J. Med.* 380, 1347–1358. doi: 10.1056/NEJMra1814259

Reina-Campos, M., Moscat, J., and Diaz-Meco, M. (2017). Metabolism shapes the tumor microenvironment. *Curr. Opin. Cell Biol.* 48, 47–53. doi: 10.1016/j.ceb.2017.05.006

Ritterhouse, L. L., Wu, E. Y., Kim, W. G., Dillon, D. A., Hirsch, M. S., Sholl, L. M., et al. (2019). Loss of SMAD4 protein expression in gastrointestinal and extra-gastrointestinal carcinomas. *Histopathology* 75, 546–551. doi: 10.1111/his.13894

Sancho, P., Burgos-Ramos, E., Tavera, A., Bou Kheir, T., Jagust, P., Schoenhals, M., et al. (2015). MYC/PGC-1α balance determines the metabolic phenotype and plasticity of pancreatic cancer stem cells. *Cell Metab.* 22, 590–605. doi: 10.1016/j.cmet.2015.08.015

Sdelci, S., Rendeiro, A. F., Rathert, P., You, W., Lin, J. G., Ringler, A., et al. (2019). MTHFD1 interaction with BRD4 links folate metabolism to transcriptional regulation. *Nat. Genet.* 51, 990–998. doi: 10.1038/s41588-019-0413-z

Sinkala, M., Mulder, N., and Martin, D. (2020). Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. *Sci. Rep.* 10:1212. doi: 10.1038/s41598-020-58290-2

Sinkala, M., Mulder, N., and Martin, D. P. (2018). Integrative landscape of dysregulated signaling pathways of clinically distinct pancreatic cancer subtypes. *Oncotarget* 9, 29123–29139. doi: 10.18632/oncotarget.25632

Vincent, A., Herman, J., Schulick, R., Hruban, R. H., and Goggins, M. (2011). Pancreatic cancer. *Lancet* 378, 607–620. doi: 10.1016/s0140-6736(10)62307-0

Wang, P., Zhang, Z., Ma, Y., Lu, J., Zhao, H., Wang, S., et al. (2019). Prognostic values of GMPS, PR, CD40, and p21 in ovarian cancer. *PeerJ* 7:e6301. doi: 10.7717/peerj.6301

Wang, P.-B., Chen, Y., Ding, G.-R., Du, H.-W., and Fan, H.-Y. (2020). Keratin 18 induces proliferation, migration, and invasion in gastric cancer via the MAPK signalling pathway. *Clin. Exp. Pharmacol. Physiol.* doi: 10.1111/1440-1681.13401 [Epub ahead of print].

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26, 1572–1573. doi: 10.1093/bioinformatics/btq170

Witkiewicz, A. K., McMillan, E. A., Balaji, U., Baek, G., Lin, W. C., Mansour, J., et al. (2015). Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.* 6:6744. doi: 10.1038/ncomms7744

Yan, J., Wu, L., Jia, C., Yu, S., Lu, Z., Sun, Y., et al. (2020). Development of a four-gene prognostic model for pancreatic cancer based on transcriptome dysregulation. *Aging (Albany NY)* 12, 3747–3770. doi: 10.18632/aging.102844

Yang, J., Ren, B., Yang, G., Wang, H., Chen, G., You, L., et al. (2020). The enhancement of glycolysis regulates pancreatic cancer metastasis. *Cell Mol. Life Sci.* 77, 305–321. doi: 10.1007/s00018-019-03278-z

Yao, W., Maitra, A., and Ying, H. (2020). Recent insights into the biology of pancreatic cancer. *EBioMedicine* 53:102655. doi: 10.1016/j.ebiom.2020.102655

Ying, H., Kimmelman, A. C., Lyssiotis, C. A., Hua, S., Chu, G. C., Fletcher-Sananikone, E., et al. (2012). Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. *Cell* 149, 656–670. doi: 10.1016/j.cell.2012.01.058

Yoshida, G. J. (2015). Metabolic reprogramming: the emerging concept and associated therapeutic strategies. *J. Exp. Clin. Cancer Res.* 34:111. doi: 10.1186/s13046-015-0221-y

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, E., Hou, X., Hou, B., Zhang, M., and Song, Y. (2020). A risk prediction model of DNA methylation improves prognosis evaluation and indicates gene targets in prostate cancer. *Epigenomics* 12, 333–352. doi: 10.2217/epi-2019-0349

Zhang, H., Qin, C., Gan, H., Guo, X., and Zhang, L. (2020). Construction of an immunogenomic risk score for prognostication in colon cancer. *Front. Genet.* 11:499. doi: 10.3389/fgene.2020.00499

Zhang, M., Wang, X., Chen, X., Zhang, Q., and Hong, J. (2020). Novel immune-related gene signature for risk stratification and prognosis of survival in lower-grade glioma. *Front. Genet.* 11:363. doi: 10.3389/fgene.2020.00363

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Gao, Zhang, Fei, Kong, Liu and Tan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 13 August 2021
doi: 10.3389/fcell.2021.730475

An Ovarian Cancer Susceptible Gene Prediction Method Based on Deep Learning Methods

Lu Ye[1], Yi Zhang[1], Xinying Yang[1], Fei Shen[2] and Bo Xu[2]*

[1] Department of Gynecology, Guangdong Second Provincial General Hospital, Guangzhou, China, [2] Department of Thyroid Surgery, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

Ovarian cancer (OC) is one of the most fatal diseases among women all around the world. It is highly lethal because it is usually diagnosed at an advanced stage which may reduce the survival rate greatly. Even though most of the patients are treated timely and effectively, the survival rate is still low due to the high recurrence rate of OC. With a large number of genome-wide association analysis (GWAS)-discovered risk regions of OC, expression quantitative trait locus (eQTL) analyses can explore candidate susceptible genes based on these risk loci. However, a large number of OC-related genes remain unknown. In this study, we proposed a novel gene prediction method based on different omics data and deep learning methods to identify OC causal genes. We first employed graph attention network (GAT) to obtain a compact gene feature representation, then a deep neural network (DNN) is utilized to predict OC-related genes. As a result, our model achieved a high AUC of 0.761 and AUPR of 0.788, which proved the accuracy and effectiveness of our proposed method. At last, we conducted a gene-set enrichment analysis to further explore the mechanism of OC. Finally, we predicted 245 novel OC causal genes and 10 top related KEGG pathways.

Keywords: ovarian cancer, gene prediction, omics data, deep learning method, pathway analysis

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Shihua Zhang,
Wuhan University of Science and Technology, China
Hong Ju,
Heilongjiang Vocational College of Biology Science and Technology, China

****Correspondence:***
Lu Ye
yu2223@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *25 June 2021*
Accepted: *22 July 2021*
Published: *13 August 2021*

Citation:
Ye L, Zhang Y, Yang X, Shen F and Xu B (2021) An Ovarian Cancer Susceptible Gene Prediction Method Based on Deep Learning Methods. Front. Cell Dev. Biol. 9:730475. doi: 10.3389/fcell.2021.730475

INTRODUCTION

Ovarian cancer (OC) is one of the major lethal diseases for women, despite ranking tenth in morbidity rate, it is the fifth leading cause of death among cancers (Siegel et al., 2011). Usually, OC is diagnosed at an advanced stage which induced a high death rate. However, even patients got primary treatment such as surgical resection and adjuvant drug therapy, the high rate of recurrence, and high incidence of advanced stage disease eventually caused a high mortality rate (Armstrong, 2002). In terms of treating OC and reducing the high fatality rate to improve survival in OC patients, studies have been exploring the development of new treatment, and effective chemotherapies (Badgwell and Bast, 2007; Kobayashi et al., 2012). While efforts and contributions have been made to improve the treatment, cure rates have not been raised significantly. Thus, it is more important to explore the mechanism and underlying biological pathogenic factors of OC to better understand the disease, and find a better treatment.

Genome-wide association analysis (GWAS) have identified hundreds of risk genetic variants (SNPs) associated with OC (Song et al., 2009; Bolton et al., 2010; Goode et al., 2010; Permuth-Wey et al., 2013; Pharoah et al., 2013; Shen et al., 2013). However, they can only explain a small fraction of disease risk regions in a functional way (Pomerantz et al., 2010; Grisanzio et al., 2012; Bojesen et al., 2013). It is widely known that most of the risk alleles are located in the nonprotein coding

regions of the genome, indicating that most of them are functional regulators of the expression of target genes (Hazelett et al., 2014). Thus, it is not comprehensive to identify disease-related genes by merely being dependent on GWAS datasets. To provide additional evidence for exploring risk genes, expression quantitative trait locus (eQTL) analysis is a direct method to explore candidate genes at risk loci. Since most transcripts are regulated by genes, eQTL can identify genetic variants related to the expression level of genes. eQTL analyses have identified multiple causal genes for different cancer types such as prostate, breast, colorectal, and kidney cancers (Loo et al., 2017; Yang et al., 2017; Beesley et al., 2020; Bicak et al., 2020). Therefore, it is more creditable to identify OC-related genes based on the combination of GWAS and eQTL data.

Besides, over the past decades, numerous noncoding RNAs (ncRNAs), such as lncRNA, siRNA, piRNA, and miRNA have been detected to execute the regulation function by interacting with target genes (Esteller, 2011; Lee and Young, 2013). In humans, it is estimated that the number of ncRNA genes are more than twice as many as that of protein-coding genes (Bunch et al., 2016). Thus, ncRNAs have been considered key regulators of multiple biological processes and development. Along with the rapid advancement of high-throughput sequence analyses of ncRNAs, more and more transcriptional mechanisms have been illustrated. ncRNAs should also be regarded as a major factor to explore the pathologies of OC due to its regulation function of gene expression. Hence, it is important to take into consideration the role of regulatory ncRNAs to identify OC-related causal genes.

However, along with the rapid development of understanding the mechanisms of complex disease, there are a few computational methods to predict disease genes based on various gene features. In this study, we aimed to identify susceptibility genes associated with OC based on integrated gene features. We first employed graph attention network (GAT) to learn the compact gene feature from a gene interaction network with gene features, then employed a deep neural network (DNN) to predict OC-related susceptibility genes. To further explore the mechanism of OC, we also performed a gene-set enrichment analysis to predict more related pathways in OC process.

MATERIALS AND METHODS

Work Frame

In this study, our method contains four main parts, feature extraction, compact gene feature learning based on GAT, and OC-related susceptibility gene prediction based on DNN and OC-related pathway analysis. In the first section, we extracted gene features based on integrated GWAS data, eQTL data, and published data of gene-related lncRNAs and miRNAs. In total, we extracted a 2,664-dimensional feature representation from four gene features. We then utilized a graph neural network with attention mechanism (GAT) to learn the compact gene feature for a low-dimension feature representation in order to obtain a better classification performance in the prediction process. The low-dimension feature matrix is considered the input of DNN to train the model and conduct the prediction process. After obtaining the predicted causal genes related to OC, we further performed a pathway analysis based on enrichR (Chen et al., 2013; Kuleshov et al., 2016), and a gene-set enrichment tool to find more related kyoto encyclopedia of genes and genomes (KEGG) pathways for a better understanding of the mechanism of OC. The workflow is presented in **Figure 1**.

Data Collection

We first downloaded and verified 3,181 OC-related genes from DisGeNET database (Piñero et al., 2017, 2020) as a positive dataset. To build a gene-gene interaction network, we downloaded gene interaction information from HumanNet database (Hwang et al., 2019). For constructing a balanced training set, we randomly selected 3,171 genes which have interactions with positive genes from HumanNet as a negative set. To extract ncRNA-gene interaction feature, we downloaded gene-lncRNA association and gene-miRNA association information from LncRNA2Target, and miRTarBase, respectively (Hsu et al., 2011; Jiang et al., 2015; Cheng et al., 2019; Huang et al., 2020). MiRTarBase is a database providing comprehensive information based on experimentally verified miRNA-target interactions curated from published articles; it accumulated over 13,404 validated associations. LncRNA2Target is a database storing comprehensive lncRNA-target interactions inferred from published articles and experiments.

The GWAS data providing OC susceptibility loci was downloaded from GWAS catalog database, accession ID GCST90011821 (MacArthur et al., 2017). They sampled from 1,259 European ancestry cases and 410,350 controls providing genetic variant loci related to OC. eQTL data in ovary tissue was downloaded from GTEx database (Lonsdale et al., 2013). Finally, our training set is constructed based on 3,181 positive genes and 3,171 negative genes for further feature extraction.

Feature Extraction

We extracted gene features from four aspects, susceptibility loci derived from GWAS, eQTL data from ovary tissue, interactions between genes, and miRNAs/lncRNAs. We first obtained the detail location information of training genes containing chromosome name, start position, and end position. Then, we mapped the genes to the susceptibility loci and sorted by *p*-value provided by GWAS data. We extracted the top five significant *p*-values as GWAS feature of the gene. Thus, the GWAS feature of a gene can be represented as a 5-D vector:

$$F_{\mathrm{GWAS}} = [D_1, D_2, D_3, D_4, D_5] \quad (1)$$

For those genes that cannot be mapped to five SNPs, the feature value is set to one. To extract eQTL-based gene feature, we mapped the genes to eQTL data based on gene location information, and extracted the top five significant *p*-values as eQTL feature, set feature value to one if a gene cannot map to five SNPs. Thus, the eQTL feature can also be represented as a 5-D vector:

$$F_{\mathrm{eQTL}} = [D_1, D_2, D_3, D_4, D_5] \quad (2)$$

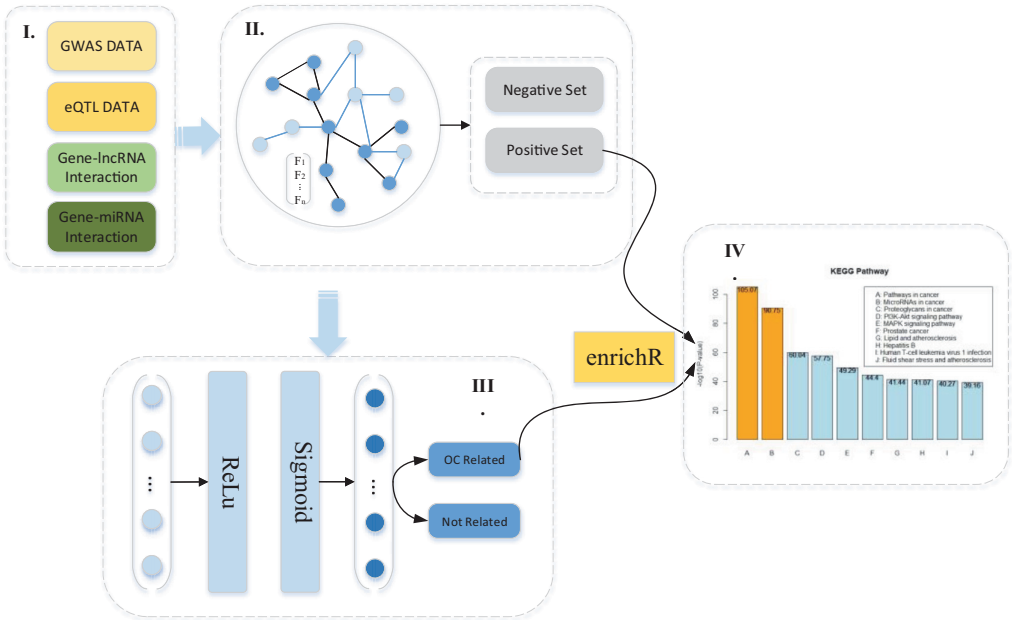


FIGURE 1 | The pipeline of ovarian cancer (OC) causal gene prediction method.

From the gene-lncRNA interaction obtained from lncRNA2Target, we filtered the interactions to make sure each of the training genes is related to at least one lncRNA. As a result, 59 lncRNAs are preserved. Thus, the lncRNA feature of a gene can be denoted as a 59-D vector, where the value is 1 if the gene is related to lncRNA[i], the value is set to 0 vice versa.

$$F_{\text{lncRNA}} = [D_1, D_2, D_3, , D_{59}] \tag{3}$$

$$D_i = \begin{cases} 0, & if\ gene\ is\ not\ related\ to\ lnRNA_i \\ 1. & if\ \ \ gene\ is\ related\ to\ lnRNA_i \end{cases} \tag{4}$$

We then filtered the gene-miRNA interactions to make sure each of the training genes is related to at least one miRNA. As a result, 2,595 miRNAs are preserved. Thus, the miRNA feature of a gene can be denoted as a 2,595-D vector:

$$F_{\text{miRNA}} = [D_1, D_2, D_3, , D_{2595}] \tag{5}$$

$$D_i = \begin{cases} 0, & if\ gene\ is\ not\ related\ to\ miRNA_i \\ 1. & if\ \ \ gene\ is\ related\ to\ miRNA_i \end{cases} \tag{6}$$

Therefore, the feature representation of each gene in training set can be denoted as a 2,664 dimensional vector. Since the feature vector could be very sparse, we need to learn the compact feature representation to obtain a better classification performance.

Compact Feature Learning Based on GAT

Sparse matrix is a matrix composed of mostly zero values, which often induces a poor classification performance in machine learning methods. Thus, we need to reconstruct the gene feature to get a low-dimensional feature representation. Since we can build a gene-gene interaction network based on gene association information obtained from HumanNet. A 6358*6358 dimensional adjacent matrix can be constructed based on the network. Besides, each gene of the Internet also has a feature representation of itself. Thus, the gene network with gene features can be regarded as a graph-structured data; to make the gene prediction method more general, and we utilized a GAT as a feature learning model. GAT addressed the shortcomings of requiring costly matrix operation and dependency on preknowledge of graph structure by stacking layers in which nodes are participant in features of neighborhoods, and arranging attention weights to each nodes. Consider the node features being denoted as: $h = \left\{\vec{h_1}, \vec{h_2}, , \vec{h_N}\right\}, \vec{h_i} \in \mathbb{R}^F$, where N is the size of training set and F is the dimension of gene features. The output of graph attentional layer is a new set of node features (of a low-dimension F'), denoted as: $h' = \left\{\vec{h_1}', \vec{h_2}', , \vec{h_N}'\right\}, \vec{h_i}' \in \mathbb{R}^{F'}$.

We then performed a self-attention on each node with a shared attentional mechanism a:$\mathbb{R}^{F'} \times \mathbb{R}^{F'}$ to compute attention coefficients:

$$e_{ij} = a(W\vec{h_i}, W\vec{h_j}) \tag{7}$$

where, $W \in \mathbb{R}^{F'\ F}$ is a weight matrix applied to each node. e_{ij} denotes the importance of node j to node i. Based on this formulation, the model allows each node to participate with every other node and dropping structural information. For each node j

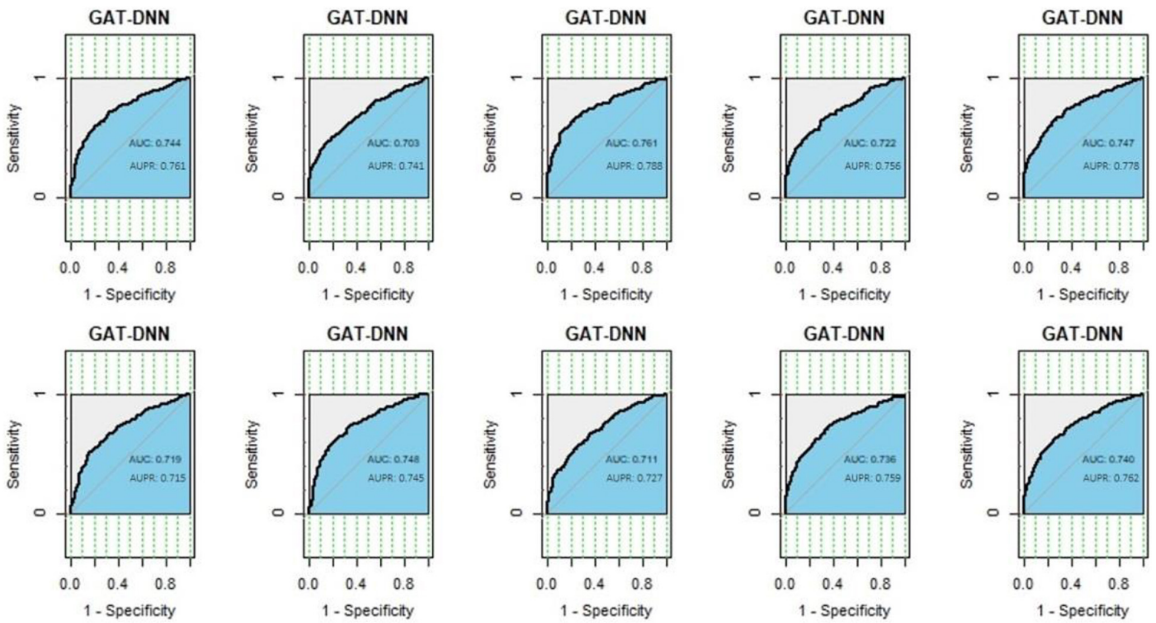


FIGURE 2 | The performance of graph attention network-deep neural network (GAT-DNN) across a 10-fold cross-validations.

in the neighborhood of node i (denoted as $\mathcal{N}_i$), we performed a softmax function to normalize the coefficients e_{ij}:

$$\alpha_{ij} = \text{softmax}\left(e_{ij}\right) = \frac{\exp\left(\text{e}_{\text{ij}}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{e}_{\text{ik}}\right)} \quad (8)$$

After being activated by LeakyReLU function, e_{ij} can be denoted as:

$$e_{ij} = \text{LeakyReLU}(\vec{a}^{T}\left[W\vec{h_i}, ||W\vec{h_j}\right]) \quad (9)$$

where, $\vec{a} \in \mathbb{R}^{2F'}$ is a weight vector; $||$ denotes the concatenation operation. Once obtained, the output feature of each node can be computed as a linear combination of the neighborhood node features with e_{ij}:

$$\vec{h_i}' = \delta(\sum j \in \mathcal{N}_i \alpha_{ij} W\vec{h_j}) \quad (10)$$

where, δ denotes a nonlinear transition. Thus, we obtained a low-dimension feature representation of the genes based on GAT.

DNN Model Construction

Deep neural network has been regarded as a powerful tool in many domains of machine-learning applications. In this part, a binary-classification DNN model is used to predict OC-related genes based on the gene features derived from the output of GAT layer. The gene features were input to the DNN. The DNN model contains one hidden layer with a ReLU activation function and an output layer with a sigmoid activation function and a dropout technique. The sigmoid activation formulation is:

$$\sigma(x) = 1/(1+e^{-x}) \quad (11)$$

We used the Adam optimizer and binary cross-entropy function as the loss function. The loss function is:

$$\text{loss} = -\sum_{i=1}^{n} y'_i \log\left(y_i\right) + (1-y'_i)\log(1-y'_i) \quad (12)$$

Training and Testing

To verify the performance of our GAT-DNN gene prediction model, we conducted a 10-fold cross-validation method on the training dataset containing 3,181 positive samples and 3,171 negative samples. The training set was randomly divided into 10 groups, nine of the 10 are regarded as training samples, and one left group is regarded as the test samples. The training samples were used to train the model and the last one was used to test the classification performance. This process was repeated 10 times to make sure the credibility of the verification.

RESULTS

Measurement of Model Performance

The performance of 10-fold cross-validation was assessed by area under curve (AUC) and the area under precision-recall curve (AUPR), as shown in **Figure 2**. As demonstrated in **Figure 2**, the AUC and AUPR are both over 0.7 across a 10-fold cross-validation, and which is a good performance for a classification model. The best performance is shown in the third validation with the AUC of 0.761 and the AUPR of 0.788, and which is chosen as the final prediction model to identify OC-related genes.

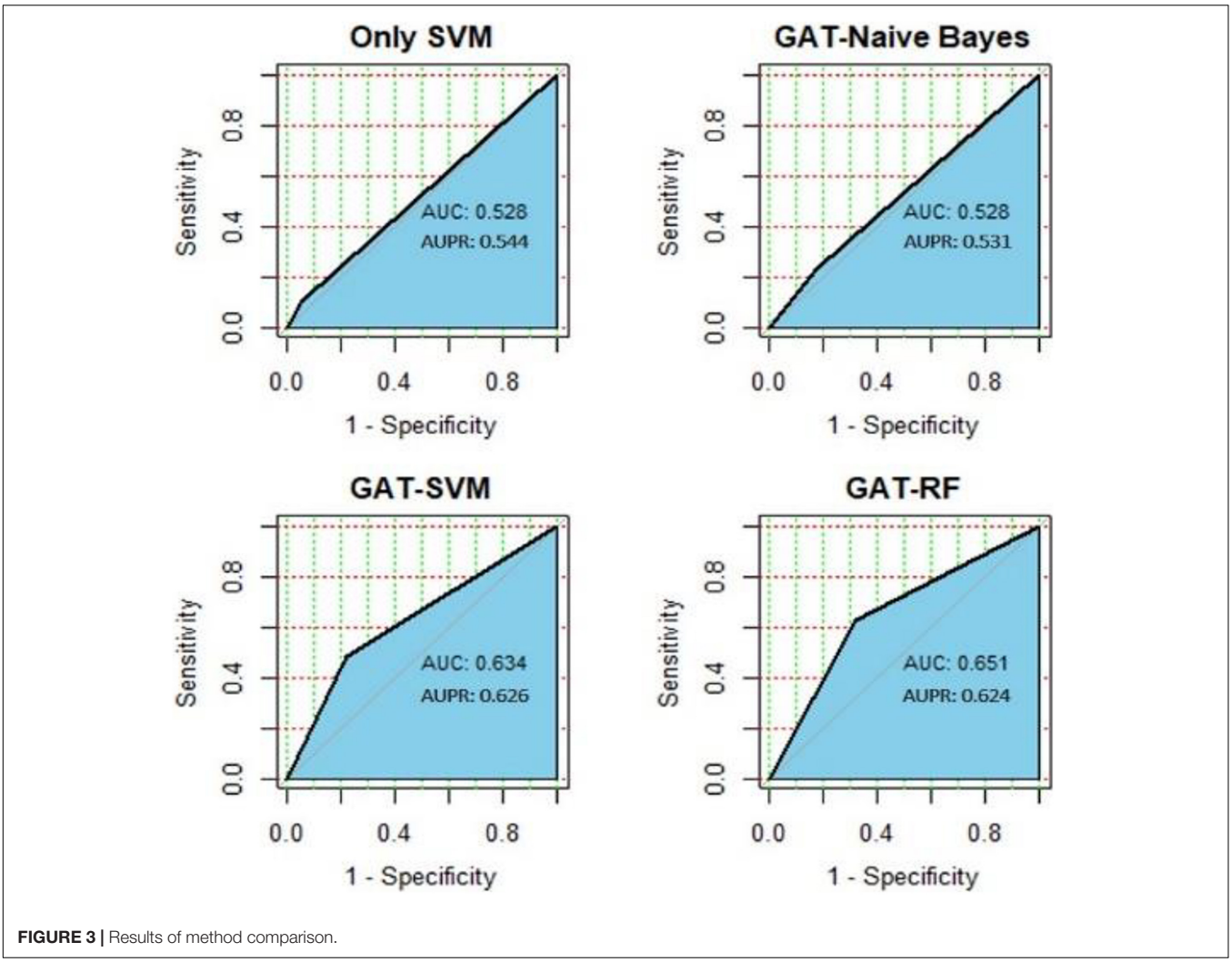


FIGURE 3 | Results of method comparison.

TABLE 1 | Top 20 predicted genes associated with ovarian cancer (OC).

Gene	Score	Gene	Score
HS6ST1	0.98354	TEAD2	0.88879
C2orf83	0.97384	PCSK1	0.88736
TM4SF4	0.95513	MTRR	0.88176
ARTN	0.95246	H6PD	0.87361
WDPCP	0.92387	EIF2B2	0.87341
KCNJ11	0.92073	SOX5	0.85771
TBL2	0.91197	NNMT	0.85176
AGTR2	0.90641	MIR324	0.84883
ATF1	0.90194	MIR33B	0.83557
PKD1	0.89846	TBPL2	0.83491

Performance Comparison

To better illustrate the effectiveness and credibility of our method, we compared our model with other four combinations of machine learning methods with the same training set we used in model training part. We compared our model with GAT-SVM, only SVM (which means the gene features are not operated with GAT), GAT-Random Forest, and GAT-Naïve Bayes. The results are shown in **Figure 3**. As shown in the results, the performance is significantly poorer than our model. The best model is GAT-RF, with an AUC of 0.651 and AUPR of 0.624, which is approximately 0.1 lower than our GAT-DNN model. However, indicated from the performance of only SVM and GAT-SVM, it is obvious that the classification performance has been improved significantly after compact feature learning by GAT layer. Therefore, our model is the best to predict OC-related genes.

OC Gene Prediction Process

Since we have demonstrated the effectiveness of our classification model based on a 10-fold cross-validation and comparison with other classification models, and we then performed the gene prediction process based on our built model on 721 ovary disease-related genes. These 721 candidate genes were downloaded from DisGeNET which is associated with ovary diseases but not OC. We extracted the gene features as mentioned in the section "Feature Extraction". After compact feature learning by GAT layer, we reduced the dimension of gene features from 2,664 to 100. We then input the compact gene features to the DNN

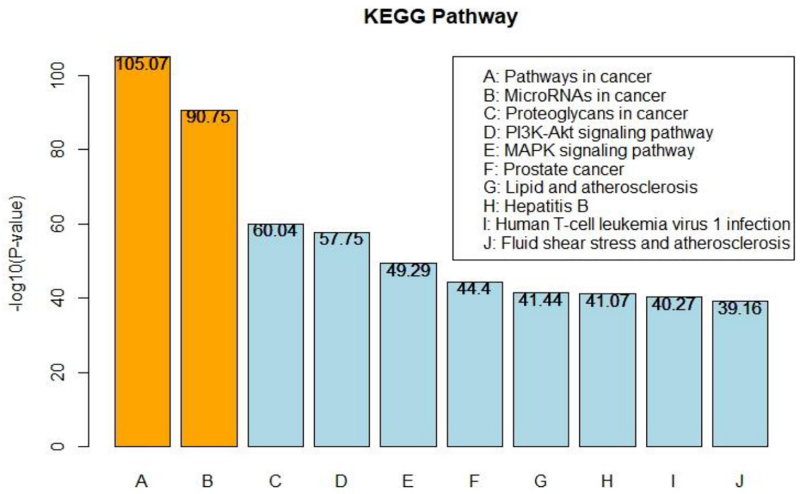


FIGURE 4 | Pathway analysis based on OC-related genes.

model; we finally predicted 245 of 721 candidate genes to be positively related to OC.

Case Study

According to the results we obtained from gene prediction process, 245 of 721 candidate genes were predicted to be associated with OC. We listed the top 20 genes in **Table 1**. Backen et al. (2007) indicated that HS6ST1 are aberrant overexpression in carcinoma of ovary compared with normal ovaries. Natanzon et al. (2018) observed a significant association between methylation WDPCP expression in OC. KCNJ11 could be considered a favorable prognostic factor since they are observed to be expressed in OC according to the investigation of Fukushiro-Lopes et al. (2020). TBL2 was identified by Kim et al. (2012) as a DNA methylation regulated cancer antigen in OC. Park et al. (2014) investigated the expression of AGTR1 and AGTR2 in OC and revealed that the dual regulation of AGTR1/2 may be a therapeutic strategy since AGTR2 could antagonize the cancer cell proliferation induced by AGTR1.

Pathway Analysis

After predicting the causal genes by our proposed model, we combined the published OC-related genes and our predicted genes with a total number of 3,426. We performed a pathway analysis on KEGG pathways using enrichr in order to further understand the mechanism of OC. Enrichr is a gene-set enrichment method to identify pathway enrichment among genes related to OC. The top 10 enriched pathways resulting from enrichr are shown in **Figure 4**. Enriched pathways are ordered by –log(p-value), obtained from a Fisher's exact test.

Consistent with KEGG DISEASE database, top 2 OC-related pathways named Pathways in cancer (hsa05200) and MicroRNAs in cancer (hsa05206) are enriched among the predicted OC-related genes. Pathway proteoglycans (PGs) in cancer are known as a key pathway in understanding cancers since PGs in the tumor microenvironment are indicated to play important roles in contributing to biology of multiple types of cancer. The MAPK and PI3K-AKT pathway have been frequently observed to be important in OC, and both of the pathways are involved in OC cell migration (Jeong et al., 2013; Li et al., 2014). Understanding the pathways related to OC are important in revealing the underlying mechanism of OC.

DISCUSSION

In this article, we proposed an OC causal gene prediction method based on deep learning methods. We first extracted gene features considering ncRNA regulation function of gene expression and integrated GWAS and eQTL data. To learn a compact feature representation, we utilized a GAT, which can learn node features from a graph-structured data format without a preknowledge of the graph structure. After GAT layer, the feature dimension is reduced from 2,664 to 100. The new feature representations were then input to a DNN model which can learn gene features and perform a binary classification task. To demonstrate the performance of our proposed method, we conducted a 10-fold cross-validation on training set and made a comparison with other four integrated machine learning models. As a result, our model is significantly better than other models and achieved a high AUC of 0.761 and the AUPR of 0.788. We then employed the constructed model to predict causal genes and obtained 245 related genes. From the result of KEGG pathway analysis, we identified more OC-related pathways which are potential favorable evidence in understanding the mechanism of OC and provide new ideas for diagnosis and treatment.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and

accession number(s) can be found in the article/supplementary material.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LY, YZ, and BX participated in its design. LY, YZ, XY, and FS analyzed the data. LY, YZ, and FS wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

Armstrong, D. K. (2002). Relapsed ovarian cancer: challenges and management strategies for a chronic disease. *Oncologist* 7, 20–28. doi: 10.1634/theoncologist.7-suppl_5-20

Backen, A. C., Cole, C. L., Lau, S. C., Clamp, A. R., McVey, R., Gallagher, J. T., et al. (2007). Heparan sulphate synthetic and editing enzymes in ovarian cancer. *Br. J. Cancer* 96, 1544–1548. doi: 10.1038/sj.bjc.6603747

Badgwell, D., and Bast, R. C. JR. (2007). Early detection of ovarian cancer. *Dis. Mark.* 23, 397–410. doi: 10.1155/2007/309382

Beesley, J., Sivakumaran, H., Marjaneh, M. M., Shi, W., Hillman, K. M., Kaufmann, S., et al. (2020). eQTL colocalization analyses identify NTN4 as a candidate breast cancer risk gene. *Am. J. Hum. Genet.* 107, 778–787. doi: 10.1016/j.ajhg.2020.08.006

Bicak, M., Wang, X., Gao, X., Xu, X., Väänänen, R.-M., Taimen, P., et al. (2020). Prostate cancer risk SNP rs10993994 is a trans-eQTL for SNHG11 mediated through MSMB. *Hum. Mol. Genet.* 29, 1581–1591. doi: 10.1093/hmg/ddaa026

Bojesen, S. E., Pooley, K. A., Johnatty, S. E., Beesley, J., Michailidou, K., Tyrer, J. P., et al. (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* 45, 371–384. doi: 10.1038/ng.2566

Bolton, K. L., Tyrer, J., Song, H., Ramus, S. J., Notaridou, M., Jones, C., et al. (2010). Common variants at 19p13 are associated with susceptibility to ovarian cancer. *Nat. Genet.* 42, 880–884. doi: 10.1038/ng.666

Bunch, H., Lawney, B. P., Burkholder, A., Ma, D., Zheng, X., Motola, S., et al. (2016). RNA polymerase II promoter-proximal pausing in mammalian long non-coding genes. *Genomics* 108, 64–77. doi: 10.1016/j.ygeno.2016.07.003

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* 14:128. doi: 10.1186/1471-2105-14-128

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144. doi: 10.1093/nar/gky1051

Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–874. doi: 10.1038/nrg3074

Fukushiro-Lopes, D., Hegel, A. D., Russo, A., Senyuk, V., Liotta, M., Beeson, G. C., et al. (2020). Repurposing Kir6/SUR2 channel activator minoxidil to arrests growth of gynecologic cancers. *Front. Pharmacol.* 11:577. doi: 10.3389/fphar.2020.00577

Goode, E. L., Chenevix-Trench, G., Song, H., Ramus, S. J., Notaridou, M., Lawrenson, K., et al. (2010). A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat. Genet.* 42, 874–879. doi: 10.1038/ng.668

Grisanzio, C., Werner, L., Takeda, D., Awoyemi, B. C., Pomerantz, M. M., Yamada, H., et al. (2012). Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11252–11257. doi: 10.1073/pnas.1200853109

Hazelett, D. J., Rhie, S. K., Gaddis, M., Yan, C., Lakeland, D. L., Coetzee, S. G., et al. (2014). Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.* 10:e1004102. doi: 10.1371/journal.pgen.1004102

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., et al. (2011). miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Res.* 39, D163–D169. doi: 10.1093/nar/gkq1107

Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.* 48, D148–D154. doi: 10.1093/nar/gkz896

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2019). HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* 47, D573–D580. doi: 10.1093/nar/gky1126

Jeong, G. O., Shin, S. H., Seo, E. J., Kwon, Y. W., Heo, S. C., Kim, K.-H., et al. (2013). TAZ mediates lysophosphatidic acid-induced migration and proliferation of epithelial ovarian cancer cells. *Cell. Physiol. Biochem.* 32, 253–263. doi: 10.1159/000354434

Jiang, Q., Wang, J., Wu, X., Ma, R., Zhang, T., Jin, S., et al. (2015). LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic Acids Res.* 43, D193–D196. doi: 10.1093/nar/gku1173

Kim, K.-M., Song, M.-H., Kim, M.-J., Daudi, S., Miliotto, A., Old, L., et al. (2012). A novel cancer/testis antigen KP-OVA-52 identified by SEREX in human ovarian cancer is regulated by DNA methylation. *Int. J. Oncol.* 41, 1139–1147. doi: 10.3892/ijo.2012.1508

Kobayashi, E., Ueda, Y., Matsuzaki, S., Yokoyama, T., Kimura, T., Yoshino, K., et al. (2012). Biomarkers for screening, diagnosis, and monitoring of ovarian cancer. *Cancer Epidemiol. Prevent. Biomark.* 21, 1902–1912. doi: 10.1158/1055-9965.epi-12-0646

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Lee, T. I., and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237–1251. doi: 10.1016/j.cell.2013.02.014

Li, H., Zeng, J., and Shen, K. (2014). PI3K/AKT/mTOR signaling pathway as a therapeutic target for ovarian cancer. *Arch. Gynecol. Obstetr.* 290, 1067–1078. doi: 10.1007/s00404-014-3377-3

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi: 10.1038/ng.2653

Loo, L. W., Lemire, M., and Le Marchand, L. (2017). In silico pathway analysis and tissue specific cis-eQTL for colorectal cancer GWAS risk variants. *BMC Genom.* 18:381. doi: 10.1186/s12864-017-3750-2

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901. doi: 10.1093/nar/gkw1133

Natanzon, Y., Earp, M., Cunningham, J. M., Kalli, K. R., Wang, C., Armasu, S. M., et al. (2018). Genomic analysis using regularized regression in high-grade serous ovarian cancer. *Cancer Inform.* 17:1176935118755341. doi: 10.1177/1176935118755341

Park, Y.-A., Choi, C. H., Do, I.-G., Song, S. Y., Lee, J. K., Cho, Y. J., et al. (2014). Dual targeting of angiotensin receptors (AGTR1 and AGTR2) in epithelial ovarian carcinoma. *Gynecol. Oncol.* 135, 108–117. doi: 10.1016/j.ygyno.2014.06.031

Permuth-Wey, J., Lawrenson, K., Shen, H. C., Velkova, A., Tyrer, J. P., Chen, Z., et al. (2013). Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21. 31. *Nat. Commun.* 4:1627. doi: 10.1038/ncomms2613

Pharoah, P. D., Tsai, Y.-Y., Ramus, S. J., Phelan, C. M., Goode, E. L., Lawrenson, K., et al. (2013). GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat. Genet.* 45, 362–370. doi: 10.1038/ng.2564

Piñero, J., Bravo, À, Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855. doi: 10.1093/nar/gkz1021

Pomerantz, M. M., Shrestha, Y., Flavin, R. J., Regan, M. M., Penney, K. L., Mucci, L. A., et al. (2010). Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. *PLoS Genet.* 6:e1001204. doi: 10.1371/journal.pgen.1001204

Shen, H., Fridley, B. L., Song, H., Lawrenson, K., Cunningham, J. M., Ramus, S. J., et al. (2013). Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. *Nat. Commun.* 4:1628. doi: 10.1038/ncomms2629

Siegel, R., Ward, E., Brawley, O., and Jemal, A. (2011). Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths. *Cancer J. Clin.* 61, 212–236. doi: 10.3322/caac.20121

Song, H., Ramus, S. J., Tyrer, J., Bolton, K. L., Gentry-Maharaj, A., Wozniak, E., et al. (2009). A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22. 2. *Nat. Genet.* 41, 996–1000. doi: 10.1038/ng.424

Yang, M. Q., Li, D., Yang, W., Zhang, Y., Liu, J., and Tong, W. (2017). A gene module-based eQTL analysis prioritizing disease genes and pathways in kidney cancer. *Comput. Struct. Biotechnol. J.* 15, 463–470. doi: 10.1016/j.csbj.2017.09.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Ye, Zhang, Yang, Shen and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

METHODS
published: 13 August 2021
doi: 10.3389/fcell.2021.731424

Psi-Caller: A Lightweight Short Read-Based Variant Caller With High Speed and Accuracy

Yadong Liu†, Tao Jiang†, Yan Gao, Bo Liu, Tianyi Zang and Yadong Wang**

Center for Bioinformatics, Faculty of Computing, Harbin Institute of Technology, Harbin, China

With the rapid development of short-read sequencing technologies, many population-scale resequencing studies have been carried out to study the associations between human genome variants and various phenotypes in recent years. Variant calling is one of the core bioinformatics tasks in such studies to comprehensively discover genomic variants in sequenced samples. Many efforts have been made to develop short read-based variant calling approaches; however, state-of-the-art tools are still computationally expensive. Meanwhile, cutting-edge genomics studies also have higher requirements on the yields of variant calling. Herein, we propose Partial-Order Alignment-based single nucleotide polymorphism (SNV) and Indel caller (Psi-caller), a lightweight variant calling algorithm that simultaneously achieves high performance and yield. Mainly, Psi-caller recognizes and divides the candidate variant site into three categories according to the complexity and location of the signatures and employs various methods including binomial model, partial-order alignment, and de Bruijn graph-based local assembly to handle various categories of candidate variant sites to call and genotype SNVs/Indels, respectively. Benchmarks on simulated and real short-read sequencing data sets demonstrate that Psi-caller is times faster than state-of-the-art tools with higher or equal sensitivity and accuracy. It has the potential to well handle large-scale data sets in cutting-edge genomics studies.

Keywords: variant calling, partial order alignment, short read sequencing, SNV/Indel detection, local assembly

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Le Zhang,
Sichuan University, China
Xin Gao,
King Abdullah University of Science and Technology, Saudi Arabia

****Correspondence:***
Tianyi Zang
tianyi.zang@hit.edu.cn
Yadong Wang
ydwang@hit.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *27 June 2021*
Accepted: *15 July 2021*
Published: *13 August 2021*

Citation:
Liu Y, Jiang T, Gao Y, Liu B, Zang T and Wang Y (2021) Psi-Caller: A Lightweight Short Read-Based Variant Caller With High Speed and Accuracy. Front. Cell Dev. Biol. 9:731424. doi: 10.3389/fcell.2021.731424

INTRODUCTION

High-throughput sequencing (HTS) has become a fundamental approach to characterize human genomes (Lander et al., 2001; Shendure et al., 2019). Especially, the discovery of genomic variants from HTS data, i.e., variant calling, is one of the most important HTS applications that is fundamental in many genomics studies to discover the associations between genome variations and important phenotypes and diseases (Shastry, 2002; Weischenfeldt et al., 2013; Kosugi et al., 2019), as well as the diversity and evolution of human genomes at both individual and population levels (Auton et al., 2015; Wu et al., 2019). Single nucleotide polymorphisms (SNVs) and short insertions/deletions (Indels) are the genomic alteration that usually refers to the change of less than 50-base pair (bp) nucleotide fragments compared to structural variants. Variant calling is the detection of the nucleotide differences between donor and reference genomes from millions to billions of HTS reads. With the rapid development of HTS technologies such as Illumina platform (Caporaso et al., 2012) and long-read sequencing technologies, such as Pacific Biosciences (PacBio)

(Roberts et al., 2013) and Oxford Nanopore Technologies (ONT) (Jain et al., 2016), the calling of variants are promising with high yield and precision. Long reads greatly improve the detection of structural variants because of the long-range spanning information (Goodwin et al., 2016). However, long reads often suffer from high error rates including substitution, small insertions, and deletion (Roberts et al., 2013; Jain et al., 2016), which is still non-trivial for long read-based caller to distinguish genuine variants and sequencing errors. For most short-read sequencing platforms, Indel errors are rare and simultaneously achieve high base accuracy (>99%), which have been proven more helpful for SNV/Indel calling in several large-population genome projects (Auton et al., 2015; Wu et al., 2019).

However, existing variant calling approaches are computationally expensive and/or have relatively low sensitivity and accuracy, which becomes the bottleneck of variant calling for large-scale studies, such as population sequencing projects, which requires analyzing tens of thousands of data sets. It is on demand to develop more advanced bioinformatics approaches with higher speed, sensitivity, accuracy, and scalability. In recent years, many efforts have been made to develop alignment-based approaches such as GATK (McKenna et al., 2010), FreeBayes (Garrison and Marth, 2012), and Clair (Luo et al., 2020), which has been mainstream approaches in short read-based variant calling. Such approaches extract signatures of SNVs and Indels from the pileups of the reads in relatively small genomic regions, and various kinds of methods, such as local assembly (GATK), Bayesian statistics (FreeBayes, GAKT), and deep neural network (Clair), are used to analyze the features to detect potential variants. In detail, GATK uses logistic regression to model base errors, hidden Markov models to compute read likelihoods, and naive Bayes classification to identify variants, which are then filtered to remove likely false positives using a Gaussian mixture model with hand-crafted features capturing common error modes. FreeBayes enables direct detection of haplotypes from short reads using a Bayesian statistical framework, which is capable of modeling multiallelic loci in sets of individuals with non-uniform copy number. Clair supports variant calling for short reads and long reads using a deep learning-based method, where the summary of aligned reads around putative candidate variant sites was used as input of the deep learning framework.

Herein, we propose Partial-Order Alignment-based SNV and Indel caller (Psi-caller), a novel ultrafast and versatile alignment-based variant calling approach with two key features. (1) Psi-caller recognizes candidate variant sites from pileup alignments and then divides the sites into three different categories according to the complexity and position of signatures. (2) Psi-caller applies various approaches, i.e., binomial model, partial-order alignment (POA) and de Bruijn graph-based local assembly, to detect and genotype SNVs and Indels in the above three different categories. The speed of Psi-caller is faster than that of state-of-the-art variant callers, especially one order of magnitude faster than some of the most popular approaches such as GATK and FreeBayes. Moreover, it also has higher or equal yields as well. We believe that it has the potential to play an important role in many forthcoming genomic studies.

METHODS

Overview of the Psi-Caller Approach

Psi-caller uses sorted BAM or CRAM files as inputs. The input files can be provided by commonly used read aligners such as BWA (Li and Durbin, 2009). Mainly, Psi-caller extracts the divergences between reads and reference from the detailed information of read alignments (i.e., CIGARs) and uses them as signatures to detect and genotype SNVs and short Indels. The approach has four major steps as follows, and a schematic illustration is shown in **Figure 1**.

(1) Task splitting: Psi-caller splits reference genome into fixed-size blocks to make a number of variant calling subtasks. The subtasks are further operated in a parallel way with multiple CPU threads.
(2) Candidate recognition: Psi-caller analyzes the divergences between the reads and reference to recognize candidate SNV and Indel sites. Local genomic regions with variant signatures are recognized as candidates and categorized into three classes according to their positions and the ratio of supporting reads, i.e., high confidence candidates (termed as HCs), low confidence candidates (termed as LCs), and candidates in tandem repeat regions and low-complexity regions (termed as TRCs).
(3) Variants detection and genotyping: Psi-caller employs various approaches, i.e., direct binomial model, POA, and de Bruijn graph-based local assembly, to call and genotype SNVs and Indels in HCs, LCs, and TRCs.
(4) Post-processing: Psi-caller integrates the variants from multiple subtasks and filters the calls with relatively low scores.

Task Splitting

Psi-caller divides the input data into blocks to implement parallel computing. Benchmarks on several data sets in various sizes suggest that this strategy greatly helps achieve highly scaling performance (refer to "Results" section for details). In detail, a user-defined size parameter is employed (default: 20,000,000 bp) to split the reference genome non-overlapping blocks. For each block, a Psi-caller instance (termed as a task) is run and output the corresponding results to a separate intermediate file. Thus, all the tasks are run in a parallel way using GNU parallel (Tange, 2020) with a user-defined number of CPU cores.

Candidate Variant Sites Generation

For each task, Psi-caller uses Samtools (Li et al., 2009) to extract detailed information of read alignment, including the chromosome, starting position, and CIGARs for each of the reads. All the alignments with low quality (default: MAPQ < 10, can be tuned with "-q" option) and supplementary alignments (flag = 2048), secondary alignments (flag = 256), mate unmapped reads (flag = 8), and unmapped reads (flag = 4) are filtered out. The remaining high-quality alignments are then traversed from upstream to downstream to make a triple consisting of three variables, i.e., (V, S, L), which can be seen as a summary of

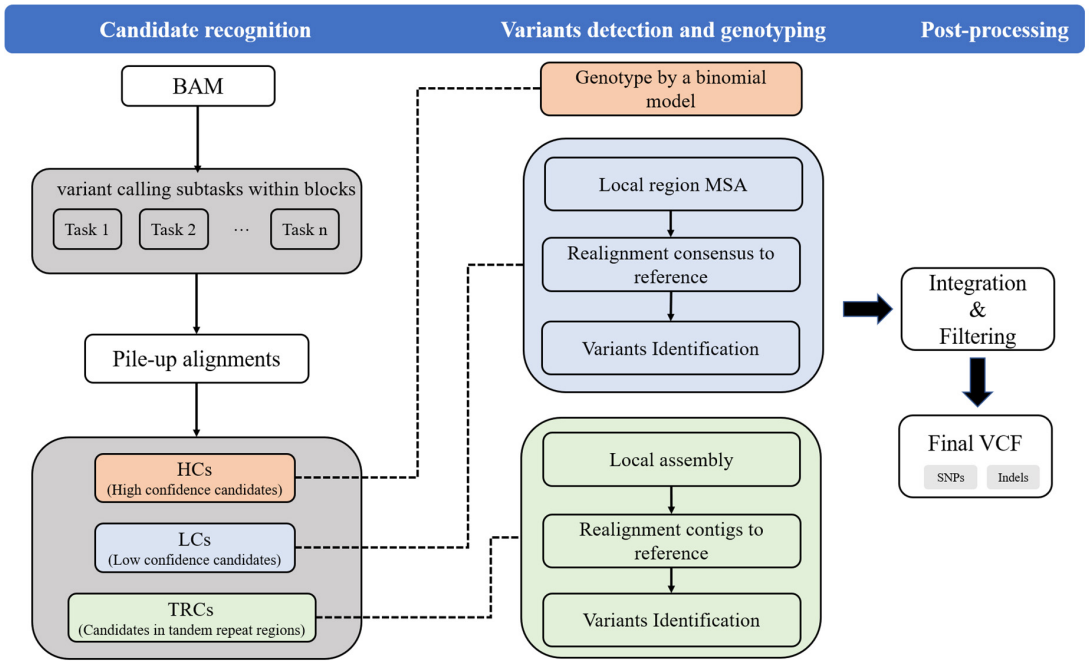


FIGURE 1 | The workflow of Psi-caller approach. Psi-caller detects SNV/Indels mainly in three steps. In step 1 ("Candidate recognition"), Psi-caller first generates multiple tasks according to chromosome and block size for parallelization and then extract candidate variant site separately for each task. In step 2 ("Variants detection and genotyping"), Psi-caller implements three difference approaches to generate variant calls and assigns genotypes. In step 3 ("Post-processing"), Psi-caller concatenates the resulting output from various tasks and removes duplicated variants.

the pileup read alignments. A schematic illustration is shown in **Supplementary Figure 1**.

(1) V (Dictionary) lists the number of alternate bases at a SNV, INDEL, or the number of reference base otherwise. For example,

$$V = \{\text{"A"} : C_1, \text{"C"} : C_2, \text{"G"} : C_3, \text{"T"} : C_4, \text{"N"} : C_5, \text{"I"} : C_6, \text{"D"} : C_7 \}.$$

(2) S (Integer) is the accumulated base quality for alternative bases.

(3) L (Dictionary) lists the inserted or deleted sequence. For example,

$$L = \{\text{"I"} : [\text{"AC"}, \text{"AC"}], \text{"D"} : [\text{"TTA"}, \text{"TA"}] \}.$$

Psi-caller sorts V according the number of base occurrences and sums all the occurrences as T. Psi-caller calculates the frequencies of various alternative bases at each genomic position by the following equation:

$$P_{alt} = c_{alt}/T \quad (1)$$

where c_{alt} is the number of a specific kind of alternative base listed in V. A position is recognized as a candidate variant site if there is at least one kind of alternative base whose frequency is higher than a threshold T_{alt}. It is also worth noting that there could be multiple alternative bases having frequencies higher than T_{alt}. In this situation, Psi-caller records all of them and recognizes this site as a multi-allele site. In order to reduce the false positives caused by sequencing errors or mistakes in alignments, Psi-caller calculates the average base quality of all the alternative bases at this genomic position by $BQ_{ave} = S/N$, where N is the total number of alternative bases at this genomic position. The position is discarded if $BQ_{ave} < 20$ empirically.

The simple rules mentioned above is able to detect a large proportion of variants since they have homogeneous and evident signatures, an example is shown in **Supplementary Figure 2**. However, there are a small proportion of variants with more complex signatures, which are usually longer Indels and/or located in repeat-rich regions; an example is provided in **Supplementary Figure 3**. There are usually relatively poor read alignments around the variant sites, and more advance methods are needed. To address these issues, Psi-caller divides all the candidate variant sites into HCs, LCs, and TRCs beforehand according to the complexity of the signatures as well as genomic locations to handle them in various approaches.

Herein, a candidate variant site is regarded as a HC if it meets the following conditions:

(1) It has high supporting evidences, in practice, $P_{alt} > 0.5$.
(2) The location of the site is out of tandem repeat and low-complexity regions (LCRs).
(3) There is no other candidate variant site within flanking sites (100 bp upstream and downstream empirically).

A candidate variant site is regarded as an LC if it satisfies the following conditions:

(1) The highest frequency of all the alternative bases is less than 0.5, or it is a multi-allele site.
(2) The location of the site is out of tandem repeat and low-complexity regions (LCRs).
(3) There is another candidate variant site within 100 bp upstream or downstream. It is worth noting that a candidate variant site is categorized as an LC if this condition is met, even if it has a >0.5 frequency for any alternative base.

Sequencing-based methods of variant detection can be confounded by repetitive and low-complexity regions (LCRs) (Trost et al., 2018). The LCRs contains (1) assembly gaps that includes centromeres, telomeres, and constitutive heterochromatin domains; (2) segmental duplications; (3) the pseudo-autosomal regions of the sex chromosomes. The short tandem repeats are defined by RepeatMasker (Smit et al., 2015).

A candidate variant site is regarded as a TRCs if it is located in short tandem repeat regions or LCRs. Herein, Psi-caller supports the use of an annotation (in BED format) on short tandem repeat and LCRs for reference genome to mark such regions in advance. In this study, the employed repeat annotation on human reference genome is available at https://github.com/PacificBiosciences/pbsv/tree/master/annotations, and the CLR annotation on human reference genome is available at http://tcag.ca/documents/projects/RLCRs_no_Repeat_Masker.zip.

For Indels, L records the allele sequence for all supporting alignments. If the context and length of allele sequence are divergent, the Indel candidate variant site is also regarded as LC.

Variant Detection and Genotyping

Psi-caller handles various types of candidate sites with three different approaches: (1) for HCs, Psi-caller employs a binomial model with a genotype prior probability to call and genotype variants; (2) for LCs, Psi-caller uses a SIMD-based fast POA method proposed in our previous study [abPOA (Gao et al., 2020)] to generate the consensus sequence of the reads of the potential variant sites; and (3) for TRCs, Psi-caller uses a de Bruijn graph-based local assembly on a larger local region to generate contigs of the reads. The POA-based and local assembly based methods enable high-quality consensus sequences of the reads around variant sites to be generated, and the variants are detected by realigning the consensus sequences against local reference. Moreover, the assembly based method has higher ability to handle local short repeats.

Variant Calling With Binomial Model for HCs

For an HC site, Psi-caller simply uses the alternative base as alternative allele if it is a single-allele site. The corresponding genotype is inferred by a maximum likelihood strategy where the likelihoods of three possible genotypes are computed by a binomial model, assuming a diploid individual

$$\begin{cases} L\left(\frac{0}{0}\right) = \frac{1-prior}{2} \times (1-\varepsilon)^{SR_{Ref}} \times \varepsilon^{SR_{Alt}} \\ L\left(\frac{0}{1}\right) = prior \times \left(\frac{1}{2}\right)^{SR_{Ref}+SR_{Alt}} \\ L\left(\frac{1}{1}\right) = \frac{1-prior}{2} \times (1-\varepsilon)^{SR_{Alt}} \times \varepsilon^{SR_{Ref}} \end{cases} \quad (2)$$

where ε is the probability of a read being mapped to a given allele mistakenly (default: 0.03), assuming it is constant and independent between all observations; *prior* is the prior probability for a heterozygous genotype. Herein, the prior distribution is configured as equal for all the three genotypes, i.e., $prior = 1/3$, which was determined by empirical experiments; SR_{Ref} and SR_{Alt} are the numbers of the read alignments carrying reference and alternative alleles, respectively.

Variant Calling With SIMD-Based POA Approach for LCs

For an LC site, a more advanced POA-based approach is used since the binomial model produces more false positives, mainly caused by the more complicated read alignments. Variant calling is implemented in the following four steps:

(1) Psi-caller traverses and merges the low confidence sites from upstream to downstream. An LC site is merged by an upstream site if their distance is shorter than a predefined threshold T_{lc} (default value: 10 bp). Psi-caller considers a set of merged LC sites as a new low confidence region and extends it both upstream and downstream with f bp (default value: 25 bp).
(2) For the extended region, Psi-caller extracts all the spanning and overlapping alignments. Alignments with overlapped ratio less than 0.5 are discarded. Further, Psi-caller employs abPOA to generate one or more consensus sequences of the involved read parts.
(3) Psi-caller aligns the generated consensus sequence(s) against the reference sequence of the candidate region (by KSW2 (Li, 2018; Suzuki and Kasahara, 2018)) and recognizes alternative allele(s) from the detailed alignment information. Moreover, it also records the numbers of supporting reads to the reference and alternative allele(s).
(4) If the base quality of the reads is not available, Psi-caller uses the binomial model same to that of high confidence candidate variant site. Otherwise, Psi-caller implements an all-to-all alignment for all the read parts and the reference and alternative alleles to calculate the likelihoods of various alleles. Further, a Bayesian model (refer to "The Bayesian approach") is used for computing the probability of various genotypes.

To improve the accuracy of consensus sequence(s) in the second step, Psi-caller clips the whole alignments at the starting and ending positions to get homogeneous breakpoint.

Variant Calling With Local Reassembly Approach for TRCs

For repeat-rich regions like tandem repeats and segmental duplications, read mapping-based methods usually fail to give reliable variant calls because of the difficulties of the accurate alignment between short reads and local repeats. The de Bruijn graph has been widely used and efficient for representation of genomic repeats as a repeat graph by short-read assembly approaches and variant callers. Thus, for a TRC site, Psi-caller implements a four-step local assembly based method to handle

the repeats and reference bias to make more accurate calls as follows:

(1) Psi-caller traverses all the TRCs and combines candidate variant sites into roughly 300-bp reassembly windows.
(2) For each reassembly window, all the overlapped reads are used to construct a de Bruijn graph. Increasing length of the k-mers (from 41 to 75, increased by 5) are flagged and treated separately to avoid loops in the de Bruijn graph. The generated contigs from previous iteration are set as pseudoreads for next-round contig generation. One or more haplotypes (default value: 2) are derived from the de Bruijn-like graph using depth-first search by prioritizing variants with the highest degree of read support and scored.
(3) KSW2 is then used to realign the candidate haplotypes to local reference sequence and recognize variants from the alignment.
(4) The inferring of genotype is the same to that of LCs.

The Bayesian Approach

To genotype a specific variant site in the donor genome, we apply a simple Bayesian statistic relating $Pr\,(G_i|R_i)$ to the likelihood of sequencing errors in the reads used for POA or local assembly and the prior likelihood of specific genotypes. For a given genotype, we use simple Bayes' rule to relate the probability of a specific genotype quality of sequencing observations and with prior expectations as follows:

$$Pr\,\{G|R\} = \frac{Pr\,\{G\} \cdot Pr\,\{R|G\}}{Pr\,\{R\}} = \frac{Pr\,\{G\} \cdot Pr\,\{R|G\}}{\sum_i Pr\,\{G_i\} \cdot Pr\,\{R|G_i\}} \quad (3)$$

where, $Pr\,(R_i|G_i)$ is the probability of a sample genotype given sequencing observations, which is represented by the following.

$$\Pr\,\{R|G\} = \prod_j \left(\frac{Pr\left\{R_j|H_1\right\}}{2} + \frac{Pr\left\{R_j|H_2\right\}}{2} \right) \quad (4)$$

Assuming a diploid sample, H_1 and H_2 are two alternative alleles on haplotypes, respectively. $Pr\,(R|H)$ is the haploid likelihood function, which can be related to a specific combination of errors of sequencing observations. In practice, Psi-caller aligns local reads to the consensus sequences and contigs to collect the mismatches, Indels nearby, and then $Pr\,(R|H)$ can be calculated by the local mapping quality scores.

$$Pr\,(R|H) = 10^{-\sum_e \frac{Q_e}{10.0}} \quad (5)$$

assuming there are e errors and they are independent at different sites of the read. Q_e is the Phred-scale base quality.

The Bayesian model is applied to compute the probability of various genotypes after POA of local assembly by implementing an all-to-all alignment for all the read parts and the reference and alternative alleles when the base quality of the reads is available.

Variant Integration and Filtering

Psi-caller concatenates all the results (in VCF format) from various tasks using *vcfcat* and then removes duplicated variants. If a variant detected by an HC site can also be recognized from a POA-based or assembly based result from another task, Psi-caller chooses the one with higher variant quality as a final result.

MATERIALS

Simulated Data Sets for Benchmark

Using the curated benchmark sets for single-nucleotide (SNV), small insertion and deletion (IDNEL) of HG001 sample from the Genome in a Bottle (GIAB) consortium as ground truth, 269,554 SNVs and 42,136 Indels in chromosome 2 of human reference genome (version: hs37d5) were selected to generate simulated data sets and used to assess the "baseline" sensitivity and accuracy of Psi-caller. In the simulation study, two donor haplotypes carrying all the "homozygous" and "heterozygous" variants were generated, and then 40× Illumina-like data sets were simulated using the following three steps:

(1) Variant with genotypes equal to "0| 1" or "1| 1" were extracted to generate the VCF file containing all variants in the first *in silico* haplotype (termed as Hap1). Similarly, variants with "1| 0" and "1| 1" genotypes were used to generate the VCF file containing all the variants in the second *in silico* haplotype (termed as Hap2).
(2) The two VCF files were used as inputs to generate two *in silico* donor genomes by *SimuG* (Yue and Liti, 2019).
(3) *ART* simulator (Huang et al., 2012) was employed to generate two 20× coverage data sets with both of the two *in silico* donor genomes, respectively, and they were merged as a 40× coverage data set for diploid genome. Two such data sets with various read lengths (2 × 150 bp and 2 × 250 bp) and same insert size (i.e., 500 bp) were generated. Refer to the **Supplementary Materials** for used command lines for the simulation.

Real Data Sets for Benchmark

Two data sets produced by the Illumina platforms from the well-studied individual HG002, the son of the so-called Ashkenazi trio in GIAB, were employed to assess the ability of Psi-caller on real data sets. The data sets are paired end with different read lengths (2 × 148 bp and 2 × 250 reads, respectively), downloaded from the FTP server of GIAB (refer to **Supplementary Materials**). Moreover, hs37d5 was used as reference in the benchmark, and the high confidence SNV/INDEL benchmark set of HG002 provided by GIAB as ground truth callset, which has 3,452,896 SNVs and 585,700 Indels (with PASS filter tag). The download links are in the **Supplementary Materials**.

Evaluation Metrics

The submodule *vcfeval* of RTG tool (v3.11) was used to evaluate the results of variant callers by three metrics, i.e., Precision, Recall, and F1 score, which are defined as follows:

$$Precision = \frac{TPs}{TPs + FPs} \quad (6)$$

$$Recall = \frac{TPs}{TPs + FNs} \quad (7)$$

$$\mathrm{F1} = \frac{2 \times Precision * Recall}{Precision + Recall} \quad (8)$$

where TP, FP, and FN are true positives, false positives, and false negatives, respectively. FP are defined as variants existing in the GIAB data set that are also identified as a variant by callers, but with discrepant variant type, alternative allele. FN are defined as the variants existing in the GIAB data set but identified as non-variant by callers. F1 score is the harmonic mean of the precision and recall.

RESULTS

Simulation Benchmarks

We assessed the baseline sensitivity and accuracy of Psi-caller with the two 40× simulated data sets at first. The reads were aligned to human reference genome (version: hs37d5) by BWA-MEM (v0.7.15) and sorted by Samtools beforehand. Three state-of-the-art variant callers, i.e., GATK HaplotypeCaller (v4.1.9.0), FreeBayes (v1.3.5), and Clair (v2.1.1), were implemented on the simulated data sets for comparison. The yields of Psi-caller and the three state-of-the-art callers are shown in **Table 1**. SNVs and Indels were assessed separately.

All the four callers achieved high and close precisions in SNV calling. The sensitivities and F1 scores of Psi-caller and GATK are close to each other for SNVs. For FreeBayes, its sensitivity is about 3% lower than that of Psi-caller and GATK. This indicates that there could be some intrinsic drawbacks in the design and implementation of the Bayesian statistics method employed by FreeBayes so that some SNVs were missed. Moreover, the yields of Clair are also slightly lower than that of Psi-caller, especially caused by lower precisions. This could be due to potential overfitting as deep learning-based methods are used.

A similar trend was observed in Indel calling. The yields of Psi-caller and GATK are also higher than that of Clair and FreeBayes. More precisely, GATK achieved the highest F1 scores (95.20 and 95.20%) and outperformed Psi-caller by 0.5 and 0.4% for PE150 and PE250, respectively. This is mainly due to the relatively lower sensitivity of Psi-caller, especially the ability to detect Indels in repeat-rich loci (refer to "Discussion" section for more details), although the precisions of Psi-caller and GATK are close. Both Clair and FreeBayes had lower F1 scores; however, they had different shortcomings. Clair had obviously lower precisions, although its sensitivities were also slightly lower than that of Psi-caller and GATK, suggesting that Indel calling is still a difficult task to this deep learning-based method. The precision of FreeBayes is high; however, its sensitivity is much lower, similar to that of SNV calling.

Real Sequencing Data Benchmark

Psi-caller, GATK, FreeBayes, and Clair were further implemented on two real sequencing data sets produced by Illumina platforms (2 × 148 bp and 2 × 250 reads, refer to "Materials" section for more details) to assess their ability. The total numbers of variant calls are in the **Supplementary Table 1**. We assessed the number of calls produced by various approaches as well as their consistency at first. Psi-caller on average detected 3.8 million SNVs and 0.88 million Indels for the two data sets ("PASS" filter tag on auto chromosomes only). Meanwhile, the numbers of SNV calls produced by GATK and FreeBayes are close to that of Psi-caller. However, there are higher numbers of SNVs in the callset of Clair, i.e., more than 4 million for both of the two data sets. For Indels, the four approaches had more divergent numbers of calls. Clair had much higher numbers of calls (>900K), while it is much lower for that of FreeBayes (around 700K). Psi-caller and GATK had similar numbers of calls for the 2 × 148 bp data set; however, Psi-caller detected nearly 70K more Indels than GATK on the 2 × 250 bp data set (879K vs. 811K).

Venn diagrams of the callsets (along the whole genome) are shown in **Figures 2A,B**, which helps in investigating the consistency of the detected variants. Mainly, 4,315,869 and 4,254,729 SNV and Indel calls were commonly identified by all the callers on the two real sequencing data, respectively. Moreover, Psi-caller had the least number of unique calls than other tools (57,841 for 148PE and 47,840 for 250PE), meanwhile, GATK also had comparable unique calls. We also investigated the intersect with the ground truth callset in the high confidence region (**Figures 2C,D**). Psi-caller is still the caller having least unique calls (3,903 and 2,414 for the two data sets, respectively), indicating that there are less false positives in its callsets.

The callsets of the four approaches were then compared with the ground truth callset by *vcfeval* module of the RTG tool (**Figure 3**). Moreover, Venn diagrams of the callsets produced by the four callers (for high confidence regions only) as well as the ground truth callsets are in **Supplementary Figure 4** (SNVs and Indels are provided separately) for more detailed information. For SNV calling, Psi-caller achieved the highest precision and F1 score, 99.47 and 99.51% on the two data sets in absolute terms, respectively. Meanwhile, the F1 scores of GATK and Clair are close to that of Psi-caller (i.e., GATK: 99.52 and 99.46%, Clair: 99.49 and 99.48% for the PE148 and 250PE data sets, respectively). The F1 scores of FreeBayes are lower (97.98 and 97.95% for the PE148 and 250PE data sets, respectively), mainly due to its much lower recall rates, although its precision rates are higher or comparable to other callers. This trend is similar to that of the simulated data sets. It is also observed from the Venn diagrams (**Supplementary Figures 4A,C**) that Psi-caller still has the least numbers of unique calls.

For Indel calling, Psi-caller and GATK are the best two callers with relatively higher F1 scores (all of them are >99%), outperforming Clair and FreeBayes by 1 to 2%. For PE150 data, GATK is the best caller whose F1 score is slightly higher (by 0.44%) than that of Psi-caller; meanwhile, Psi-caller outperformed GATK by 0.17% for the PE250 data set. The performance of Indel is similar to SNVs for FreeBayes, i.e., low recalls with comparable precisions. The F1 scores of Clair are lower than those of Psi-caller and GATK, mainly due to its lowest precisions, while its recalls were only higher than that of FreeBayes as well. We further investigated the callsets of Psi-caller and GATK and found that GATK has better ability to handle clipped reads with its local assembly based approach, so that it enables to

TABLE 1 | Results of simulation benchmarks for various callers.

Data set	Method	SNV			INDEL		
		Precision	Recall	F1	Precision	Recall	F1
148PE	GATK	0.9998	0.9980	**0.9988**	0.9971	0.9107	**0.9520**
	FreeBayes	0.9995	0.9674	0.9832	0.9979	0.8624	0.9252
	Clair	0.9986	0.9976	0.9981	0.9809	0.8979	0.9376
	Psi-caller	0.9991	0.9981	0.9986	0.9943	0.9042	0.9471
250PE	GATK	0.9998	0.9980	**0.9988**	0.9971	0.9108	**0.9520**
	FreeBayes	0.9992	0.9675	0.9831	0.9988	0.8621	0.9254
	Clair	0.9983	0.9974	0.9978	0.9815	0.8987	0.9383
	Psi-caller	0.9993	0.9981	**0.9988**	0.9953	0.9051	0.9481

The bold values mean the best F1 scores in the two data sets respectively.

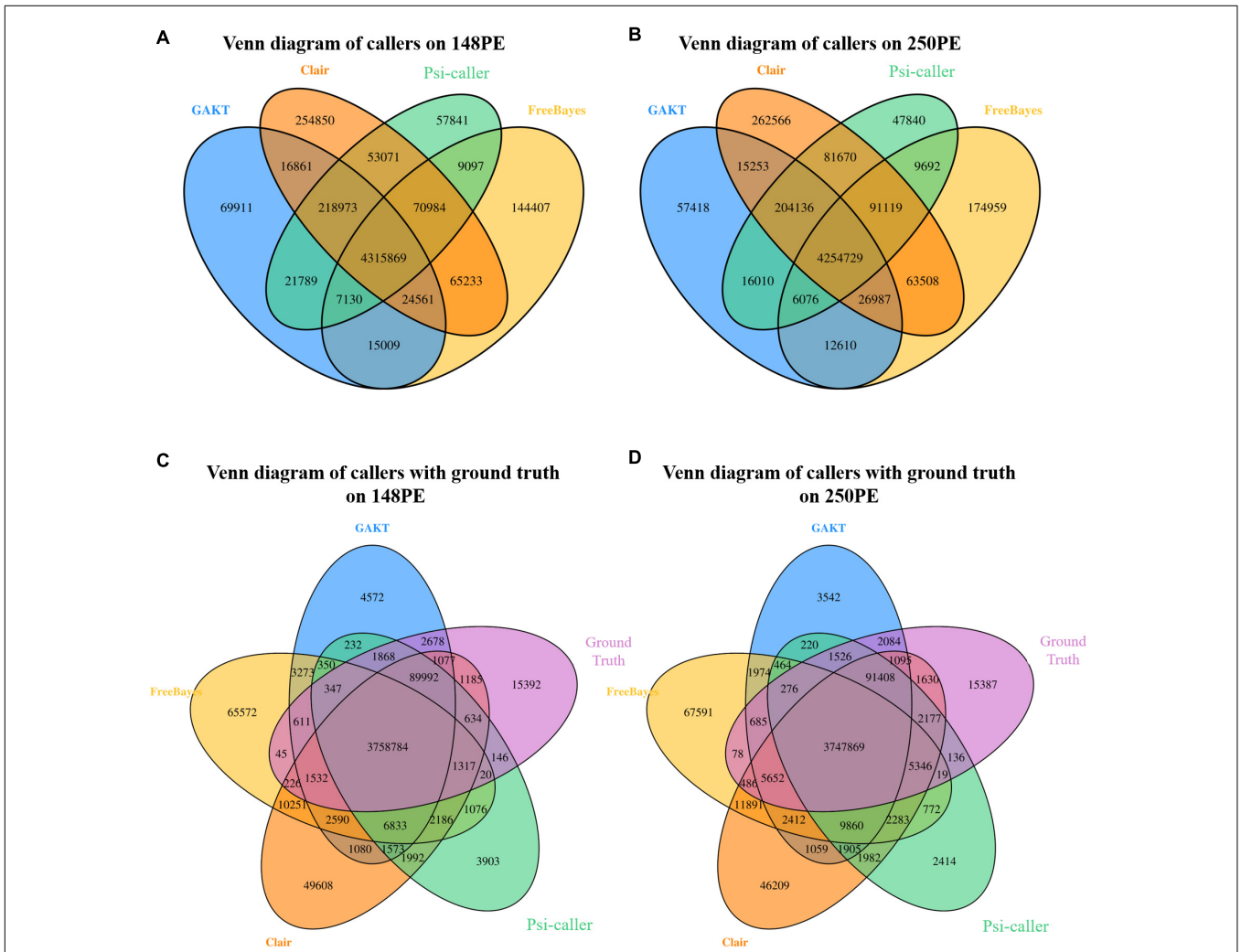


FIGURE 2 | (A,B) The Venn diagram of variant calls produced by different tools on Illumina 148PE and 250PE reads, respectively; **(C,D)** The Venn diagram of variant calls produced by different tools and ground truth on Illumina 148PE and 250PE reads, respectively.

detect more Indels with shorter reads (148PE). Meanwhile, the POA-based method of Psi-caller is more sensitive to the Indel signatures implied by alignment details (CIGARs), so that it achieved higher F1 score for the longer reads (250PE). The Venn diagrams indicate that Psi-caller and GATK had comparable least numbers of unique Indels, respectively,

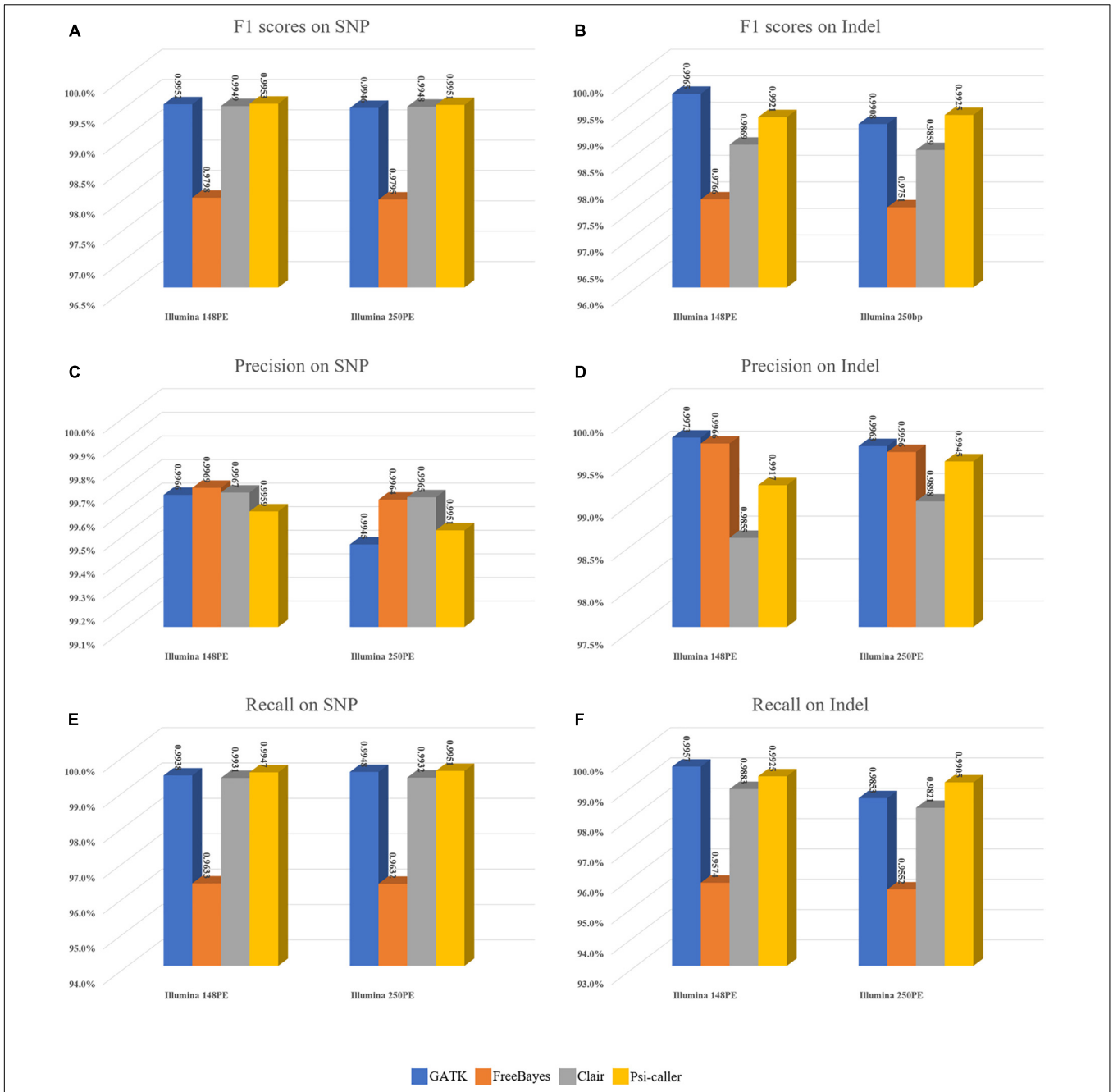


FIGURE 3 | Benchmark results of the variant callers on real sequencing data sets. **(A,B)** F1 scores on SNVs and Indels, respectively. **(C,D)** Precisions on SNVs and Indels, respectively. **(E,F)** Recalls on SNVs and Indels, respectively.

which also partially indicates their similar ability to the detection of Indels.

Performance of Psi-Caller

We assessed the speed of the callers since time cost is also a major concern for the analysis of large-scale data sets or time-sensitive tasks. The result is given in **Table 2** and **Figure 4**. With eight CPU cores, the speed of Psi-caller is tens of times faster than GATK, a few times faster than FreeBayes and slightly outperformed Clair as well. It is also worth noting that Clair is a deep learning-based approach that needs time-consuming training process, although it can be done in advance, which can be downloaded from http://www.bio8.cs.hku.hk/clair_models/illumina/12345.tar. However, Psi-caller does not need training as well as any extra requirement on hardware such as GPUs. We also benchmark the speedup of Psi-caller with different CPU cores (**Figure 4**); the results show that Psi-caller has high scaling performance, i.e., it achieved a quasi-linear speedup with the number CPU cores and its

TABLE 2 | Performance of the benchmarked variant callers.

Running time (min)	GATK	FreeBayes	Clair	Psi-caller
148PE	1244	260	117	**105**
250PE	1293	313	124	**101**

clock time was greatly reduced. Considering its good sensitivity, precision, and high speed, Psi-caller is suited to handle large-scale short-read sequencing data sets, which has potentials to many large-scale genomics studies.

The high-speed feature of Psi-caller derives from both of its designs and implementations. Mainly, SKSV divides the candidate variant sites into three categories, and the vast majority of candidates belong to HCs (details shown in **Supplementary Table 2**), which greatly speed up variant calling using the direct binomial model. In addition, the POA-based method also helps improve the speed compared to local assembly. Moreover, the block division-based subtask implementation also helps Psi-caller to achieve outstanding scaling performance. Psi-caller uses Pypy (Rigo Aea, 2018), a Just-In-Time (JIT) compiler, other than the native python interpreter to implement the operations of subtasks. Because of the existence of Global Interpreter Lock (GIL), the performance cannot be improved by using multiple threading when sample processing. However, the JIT built-in compiler and synchronous programming of Pypy greatly reduce run-time for large programs.

DISCUSSION

The rapid development of HTS technologies is promising to comprehensively discover genomic variants of various populations. However, variant calling is still a computationally intensive task whose cost is non-neglectable. Moreover, cutting-edge genomics studies also have high requirements on the sensitivity and accuracy of variant calling. Herein, we propose Psi-caller, a novel lightweight short read-based variant calling approach, as a solution to this important open problem. Mainly, we show how to achieve higher yield and performance in variant calling with tailored candidate variant sites recognition, consensus sequence reconstruction, and variant calling and genotyping methods for various kinds of candidate variant sites. Benchmarks on simulated and real data sets suggest three features of Psi-caller as follows:

(1) Psi-caller uses simple statistical rules to recognize candidate variant sites from pileup alignments and divides the candidate variant sites into three categories, i.e., high confidence candidates (termed as HCs), low confidence candidates (termed as LCs), and candidates in tandem repeat regions (termed as TRCs), according to the complexity and location of the sites with strict definitions and thresholds. Two examples are shown in **Supplementary Figures 2**, **3**. They show candidate variant sites with homogeneous and evident signatures and complex signatures, respectively.

(2) Psi-caller employs various approaches, i.e., direct binomial model, POA, and de Bruijn graph-based local assembly to call and genotype SNVs and Indels in HCs, LCs, and TRCs, respectively. An example is shown in **Supplementary Figure 5**. The specifically designed POA and local assembly based methods enable complicated signatures of SNV and Indels to be well handled.

(3) Psi-caller has outstanding speed with the help of subtask division strategy and JIT compiler in data processing. This is suited to modern high-performance computing cluster and also scalable to large-scale data analysis tasks.

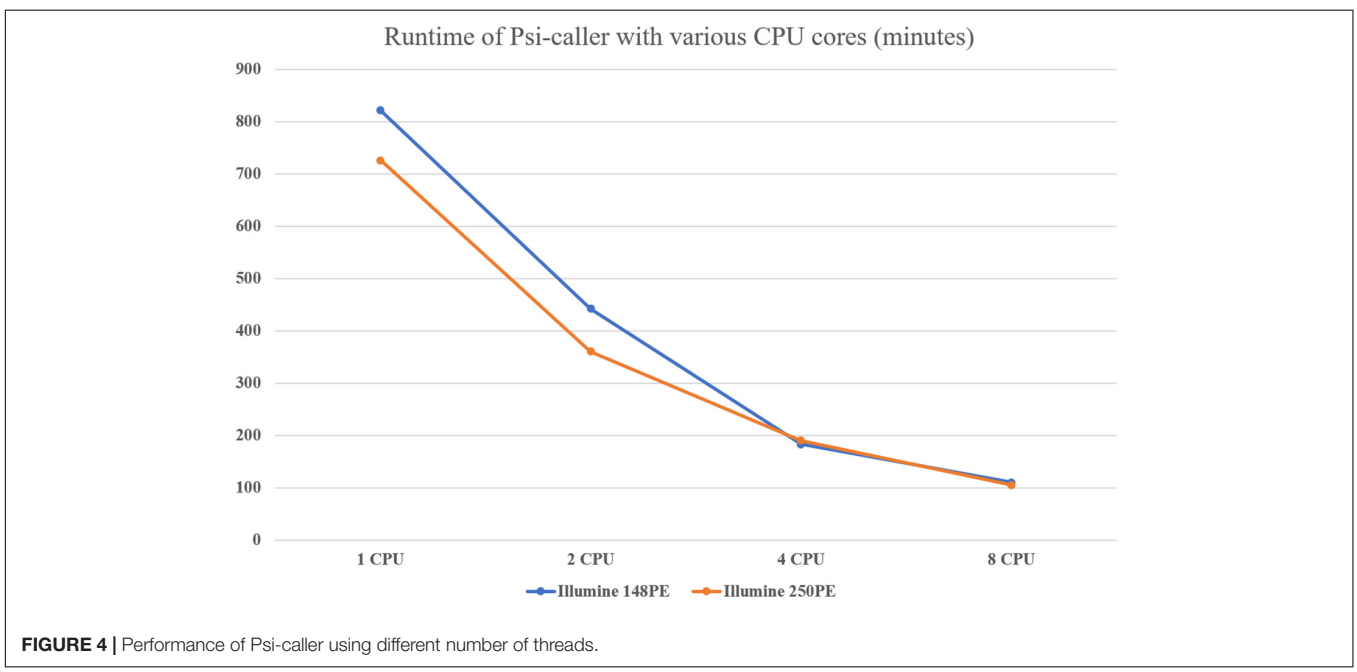


FIGURE 4 | Performance of Psi-caller using different number of threads.

The major advantage of Psi-caller is its outstanding speed, which is tens of times faster than GATK. Meanwhile, benchmarks on both of simulated and real sequencing data sets demonstrate that Psi-caller has equal or higher good sensitivity and accuracy to state-of-the-art variant callers. For SNV calling, Psi-caller achieved comparable or higher F1 scores, and it also had the least number of unique calls compared to other benchmarked state-of-the-art callers. This indicates that Psi-caller is able to produce accurate and reliable callsets. For Indel calling, Psi-caller and GATK are the best two callers with relatively higher F1 scores in the benchmark. In PE150, Psi-caller is the best runner-up, but in PE250 real sequencing data set, Psi-caller outperform GATK by 0.17%.

However, Psi-caller still has a few shortcomings like the relatively lower recalls on Indels. We investigated the detailed intermediate results of Psi-caller and observed the following four issues, which could be important to further improve the sensitivity of this approach.

(1) For some Indels in the ground truth callset, Psi-caller fails to detect them due to that there is no candidate variant site recognized from the input short-read alignments. Most of them are located in short tandem repeat regions. It is still a bottleneck for short read aligners to produce accurate alignments in such regions, so that the detection of variant is affected. An example is in **Supplementary Figure 6**. In this case, the variant is in a typical short tandem repeat region; however, no evident signature is implied by the alignment of the reads around the variant site.
(2) For some Indels, Psi-caller can recognize the corresponding variant site as candidates, although they are also in repeat-rich regions. However, the assembly method produces false-positive or false-negative variant calls due to the failure of local assembly, mainly caused by the high repetitiveness of local genomic sequences. The example displayed in **Supplementary Figure 7** illustrates that only some of candidates can be detected as variants due the power of assembly or fewer supporting alignments.
(3) Some of Indels are recognized by several nearby false positives due to the scoring system of KSW2. An example in **Supplementary Figure 8**, which obtains incorrect variants using a small mismatch penalty score, which causes SKW2 to regard alignment containing one deletion and two mismatches (i.e., SNVs) as the best result. However, it is still a difficult open problem to design a highly effective scoring system to fit all kinds of variants; hence, it could be a potential solution to use specifically designed scoring system for some difficult regions.
(4) Some Indels in low confidence sites cannot be detected by abPOA due to the ratio of supporting reads ratio and the number of reads to recognize the contigs carrying alternative alleles.

In order to improve the sensitivity of Indels, more advanced recognition and filtration strategies for candidate variant sites and more accurate partial order alignment and local assembly algorithm for detecting variants should be employed in the future works.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

YL, TJ, and BL designed the method. YL implemented the method. All authors performed the analysis, wrote the manuscript, read, and approved the final manuscript.

FUNDING

This work has been supported by the National Key Research and Development Program of China (Nos: 2017YFC0907503, 2018YFC0910504, and 2016YFC0901605) and the National Natural Science Foundation of China (32000467 and 62076082).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.731424/full#supplementary-material

REFERENCES

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., and Korbel, J. O. (2015). A global reference for human genetic variation. *Nature* 526:68. doi: 10.1038/nature15393

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms open. *ISME J.* 6, 1621–1624. doi: 10.1038/ismej.2012.8

Gao, Y., Liu, Y., Ma, Y., Liu, B., Wang, Y., and Xing, Y. (2020). abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. *Bioinformatics* doi: 10.1093/bioinformatics/btaa963

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv* [Preprint] arXiv:1207.3907,

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics* 28, 593–594. doi: 10.1093/bioinformatics/btr708

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17:239.

Kosugi, S., Momozawa, Y., Liu, X. X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 20:117.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Luo, R., Wong, C.-L., Wong, Y.-S., Tang, C. L., Liu, C. M., Leung, C. M., et al. (2020). Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat. Mach. Intell.* 2, 220–227. doi: 10.1038/s42256-020-0167-4

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Rigo Aea (2018). *Pypy*. Available online at: https://pypy.org/ (accessed May 21, 2021).

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biol.* 14:405.

Shastry, B. S. (2002). SNP alleles in human disease and evolution. *J. Hum. Genet.* 47, 561–566.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2019). DNA sequencing at 40: past, present and future. *Nature* 568, E11–E11.

Smit, A. F. A., Hubley, R., and Green, P. (2015). *RepeatMasker Open-4.0 2013-2015*. Available online at: http://www.repeatmasker.org

Suzuki, H., and Kasahara, M. (2018). Introducing difference recurrence relations for faster semi-global alignment of long sequences. *BMC Bioinformatics* 19(Suppl 1):45.

Tange, O. (2020). *GNU Parallel 20201122 ('Biden')*. Genève: Zenodo.

Trost, B., Walker, S., Wang, Z. Z., Thiruvahindrapuram, B., MacDonald, J. R., Sung, W. W. L., et al. (2018). A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am. J. Hum. Genet.* 102, 142–155. doi: 10.1016/j.ajhg.2017.12.007

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* 14, 125–138. doi: 10.1038/nrg3373

Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C. C., et al. (2019). Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* 179, 736–749.e15.

Yue, J. X., and Liti, G. N. (2019). simuG: a general-purpose genome simulator. *Bioinformatics* 35, 4442–4444. doi: 10.1093/bioinformatics/btz424

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Liu, Jiang, Gao, Liu, Zang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 23 August 2021
doi: 10.3389/fcell.2021.740267

Prioritizing Susceptible Genes for Thyroid Cancer Based on Gene Interaction Network

Lin-kun Zhong[1†], Chang-lian Xie[2†], Shan Jiang[3†], Xing-yan Deng[4], Xiao-xiong Gan[5], Jian-hua Feng[5], Wen-song Cai[5], Chi-zhuai Liu[1], Fei Shen[5], Jian-hang Miao[1] and Bo Xu[5*]*

[1] Department of General Surgery, Zhongshan City People's Hospital, Zhongshan, China, [2] Intensive Care Unit, Zhongshan Hospital of Traditional Chinese Medicine Affiliated to Guangzhou University of Chinese Medicine, Zhongshan, China, [3] Reproductive Medicine Center, Boai Hospital of Zhongshan, Zhongshan, China, [4] Department of Thyrovascular Surgery, Maoming People's Hospital, Maoming, China, [5] Department of Thyroid Surgery, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Shihua Zhang,
Wuhan University of Science and Technology, China
Junwei Han,
Harbin Medical University, China

***Correspondence:**
Jian-hang Miao
lymjh@yeah.net
Bo Xu
eyboxu@scut.edu.cn

[†]These authors share first authorship

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 12 July 2021
Accepted: 02 August 2021
Published: 23 August 2021

Citation:
Zhong L-k, Xie C-l, Jiang S, Deng X-y, Gan X-x, Feng J-h, Cai W-s, Liu C-z, Shen F, Miao J-h and Xu B (2021) Prioritizing Susceptible Genes for Thyroid Cancer Based on Gene Interaction Network. Front. Cell Dev. Biol. 9:740267. doi: 10.3389/fcell.2021.740267

Thyroid cancer ranks second in the incidence rate of endocrine malignant cancer. Thyroid cancer is usually asymptomatic at the initial stage, which makes patients easily miss the early treatment time. Combining genetic testing with imaging can greatly improve the diagnostic efficiency of thyroid cancer. Researchers have discovered many genes related to thyroid cancer. However, the effects of these genes on thyroid cancer are different. We hypothesize that there is a stronger interaction between the core genes that cause thyroid cancer. Based on this hypothesis, we constructed an interaction network of thyroid cancer-related genes. We traversed the network through random walks, and sorted thyroid cancer-related genes through ADNN which is fusion of Adaboost and deep neural network (DNN). In addition, we discovered more thyroid cancer-related genes by ADNN. In order to verify the accuracy of ADNN, we conducted a fivefold cross-validation. ADNN achieved AUC of 0.85 and AUPR of 0.81, which are more accurate than other methods.

Keywords: thyroid cancer, gene interaction, genetic testing, Adaboost, deep neural network

INTRODUCTION

Thyroid carcinoma (TC) is the most common malignant tumor of endocrine system, accounting for 2.5% of all human cancers, accounting for 90% (Maniakas et al., 2020; Sahu and Pattanayak, 2020; Xia et al., 2020) are cell-derived thyroid malignancies are derived from the follicular cells, including follicular thyroid carcinoma (FTC), papillary thyroid carcinoma (PTC), poorly differentiated thyroid carcinoma (PDTC), and anaplastic thyroid carcinoma (ATC) (Dralle et al., 2015). PTC and FTC with low malignancy are classified as differentiated thyroid carcinoma (DTC), accounting for about 90% of all thyroid cancers (Zanella et al., 2021). And the majority of deaths from thyroid carcinoma was caused by ATC (Molinaro et al., 2017). Medullary thyroid carcinoma (MTC) originates from parafollicular (c) cells, accounts for 2–4% of all thyroid carcinoma (Ceolin et al., 2019; Chen et al., 2020). About 25% of MTC cases are caused by germline genetic mutations, that is, familial medullary thyroid carcinoma (FMTC), while 75% are sporadic cases. Hereditary cases can occur alone, it can also be used as a part of multiple endocrine neoplasia type 2 (men2) syndrome (Vijayan et al., 2021).

According to statistics, the incidence rate of thyroid malignancies in recent decades is almost entirely due to the improvement of diagnostic accuracy and over diagnosis of PTC tumors, while the incidence rate of FTC, ATC, and MTC remains relatively stable (Xing, 2019; Zhang et al., 2021). The degree of differentiation of PTC is relatively high, and the corresponding degree of malignancy is relatively low, but it is not equivalent to the low risk of PTC. There are generally no obvious symptoms in the early stage of PTC (Tsukatani et al., 2020), but once clinical symptoms appear, such as hoarseness, Tracheal compression, etc., usually have entered the local advanced stage, and the best time for treatment has been missed at this time, and the metastasis of cervical lymph nodes and the invasion of local muscles, nerves and other tissues can often be seen during surgical treatment, resulting in postoperative complications. The treatment effect is not satisfactory.

Molecular markers are an effective tool for diagnosis, especially for thyroid nodules whose Fine needle aspiration cytology (FNAC) is uncertain (Sanguedolce et al., 2015). Gene mutation and chromosome rearrangement are important genetic changes in the occurrence and development of thyroid cancer. The molecular pathogenesis of most thyroid cancer involves mitogen activated protein kinase (MAPK) and phosphatidylinositide 3-kinases/protein kinase B (Sui et al., 2014), PI3K/Akt signaling pathway (Petrulea et al., 2015) is out of balance. BRAF and RAS point mutations, RET/PCT and Pax8/PPAR γ Rearrangement can activate MAPK pathway, and mainly occurs in DTC. BRAF mutation and RET/PCT rearrangement are common in PTC, while Ras mutation and Pax8/PPAR γ Rearrangement is a common molecular change in FTC. Pik3/Akt pathway is mainly activated by Ras, TP53 and TERTP mutations. TP53 and TERTP mutations are rare in well-differentiated thyroid cancer (Penna et al., 2016), and the mutation frequency is high in ATC and PDTC, which may be related to tumor invasion. Therefore, FNAC is difficult to determine the benign and malignant thyroid nodules, which can be combined with relevant molecular detection to help diagnosis, so as to improve the diagnostic accuracy of thyroid cancer.

According to the data released by the Cancer Genome Atlas (TCGA) in 2014 (Tomczak et al., 2015), 402 patients with thyroid cancer were analyzed. Compared with other cancers, the frequency of gene somatic mutations in thyroid cancer is relatively low. The frequency of BRAF V600 was 58.5%, which was the highest mutation site in thyroid cancer. In addition, the high-frequency mutation gene also includes three RAS gene family members, such as NRAS and KRAS, which are known tumor related genes, with a mutation frequency of 12.9% in European and American populations. In addition, some new thyroid cancer driving genes eif1ax, ppm1d, and CHEK2 were identified, and some of them also had gene fusion (Agrawal et al., 2014). TCGA research, with large sample size and various analysis methods, not only found a large number of somatic mutations, but also copy number variation and gene fusion information, which has great reference value and clinical significance. BRAF gene is used to assist in the diagnosis of benign and malignant thyroid nodules (Salvatore et al., 2006), which greatly reduces the misdiagnosis rate in clinical diagnosis and improves the accuracy of preoperative diagnosis of patients with papillary thyroid cancer. BRAF gene is used for clinical diagnosis of papillary thyroid cancer (Trovisco et al., 2004). It can also be used as an important factor for clinical prediction of post-operative recurrence and guidance of medication, so as to facilitate the formulation of individualized and precise diagnosis, treatment and follow-up plans. In recent years, deep learning methods have been widely used in the diagnosis thyroid cancer. Lee JH et al. developed a deep learning-based computer-aided diagnosis (CAD) system could accurately classify cervical lymph node metastasis (LNM) on CT images in patients with thyroid cancer (Lee et al., 2019). Li X et al. indicated that compared with a group of skilled radiologists, deep convolutional neural network (DCNN) models that showed similar sensitivity and improved the diagnostic accuracy of thyroid cancer on sonographic images (Li et al., 2019).

Although multiple TC-related genes have been found by collecting samples and implementing gene differential expression analysis (Zhao et al., 2021b), people are still unclear about the pathogenesis and early diagnosis of thyroid cancer. With the increasing computational power and omics data, machine learning methods can identify disease-related molecules on a large scale to reveal the pathogenesis (Zhao et al., 2020c), disease occurrence process (Zhao et al., 2020a) and clinical medication guidance (Tianyi et al., 2020). Most of the calculation methods are based on similarity and interaction (Zhao et al., 2020b, 2021a). In this article, we propose hypotheses: there is a stronger interaction between the core genes that cause thyroid cancer. There is a close relationship between the pathogenic genes of thyroid cancer, but the interaction between the genes only related to thyroid cancer and these genes is not that close. Based on this hypotheses, we constructed a gene interaction network and used Random Walk (RW) to traverse this network. Then, Adaboost and deep neural network (DNN) was fused to identify TC-related genes.

MATERIALS AND METHODS

There are three steps to implement ADNN. First, TC-related genes are obtained from DisGeNET (Piñero et al., 2015). Then, we collected genes which can interaction with TC-related genes to construct gene interaction network. The red points represent TC-related genes and blue points represent other genes. The second step is to use RW to traverse this network. The features of genes can be encoded by this step. The last step is to fuse Adaboost with DNN to prioritize TC-related genes. The whole process of ADNN is shown as **Figure 1**.

Construction of Gene Interaction Network

First, we obtained TC-related genes from DisGeNET. According to DisGeNET, there are 147 genes related to TC. Using String database (Szklarczyk et al., 2016), we draw these genes interaction network as **Figure 2**.

As we can see in **Figure 2**, the interaction between some core genes in the center of the network is very close. Although there are still some marginal genes that cannot interact with other

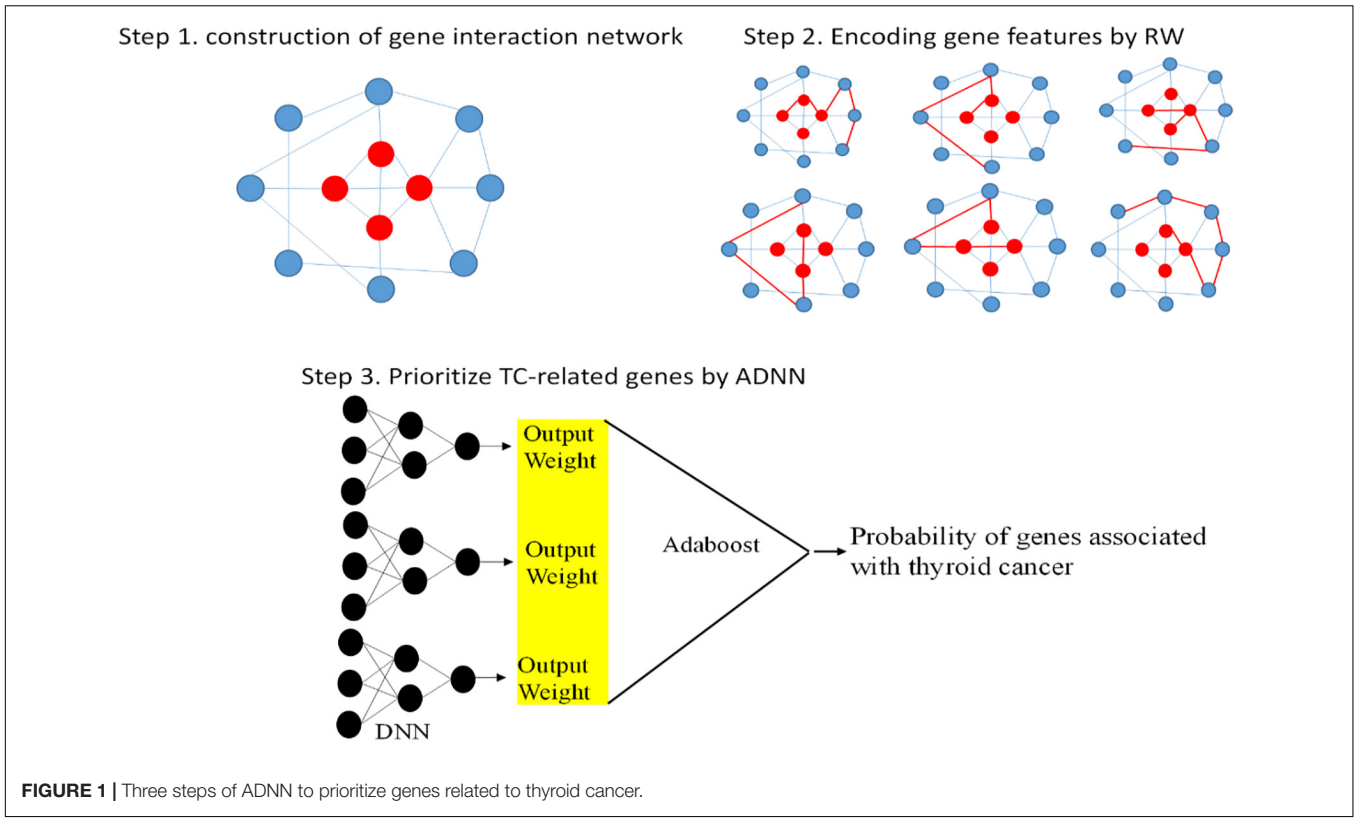


FIGURE 1 | Three steps of ADNN to prioritize genes related to thyroid cancer.

genes, most of the genes have close relationship with other genes. We collected genes which can interaction with 147 TC-related genes to construct a whole gene interaction network.

Encoding Gene Features by RW

Since the gene interaction network we construct is a two-dimensional graph, when we walk through the network in a probabilistic manner based on gene interaction, when the node information of the current gene is known, the historical gene node traversal information and the future gene node The traversal path is irrelevant. Therefore, we can regard the genetic coding method based on random walk as a Markov chain. In each step of the Markov chain, the probability distribution of gene interaction can change from one state to another, or maintain the current state. The change of state is called transition, and the probability associated with different states is called transition probability.

If A is the adjacency matrix of the gene interaction network, we can normalize A as:

$$P = D^{-1}A \tag{1}$$

D is a diagonal matrix and the degree matrix of the gene interaction network:

$$D(i,i) = \sum A(i,j) \tag{2}$$

P is a random walk matrix, the transition probability of each node is 1, and P is the probability matrix associated with TC and all genes.

A random walk matrix corresponds to a Markov chain, and the probability distribution of TC-related genes changes as the state in the Markov chain changes. Starting from any state, the probability of going to the next state is as follows:

$$P_{t+1} = A_t P \tag{3}$$

This process continues, and the relationship between TC and genes is constantly changing. After a period of time, it reaches a state of equilibrium. The equilibrium state is also called steady state, which means that the probability distribution of the association between TC and genes no longer changes. The calculation method of steady state is as follows:

$$\pi = D(i,j) / \sum_i \sum_j A(i,j) \tag{4}$$

When $\pi P = \pi$, the entire system reaches a steady state. This steady state is the final calculated association between TC and gene.

Prioritize TC-Related Genes by ADNN

DNN neural network layers can also be simply divided into three categories: input layer, hidden layer and output layer. Its layers are fully connected, that is, all neurons in the upper layer are connected to any neuron in the next layer. Its partial model is:

$$O = \sigma\left(\sum wx + b\right) \tag{5}$$

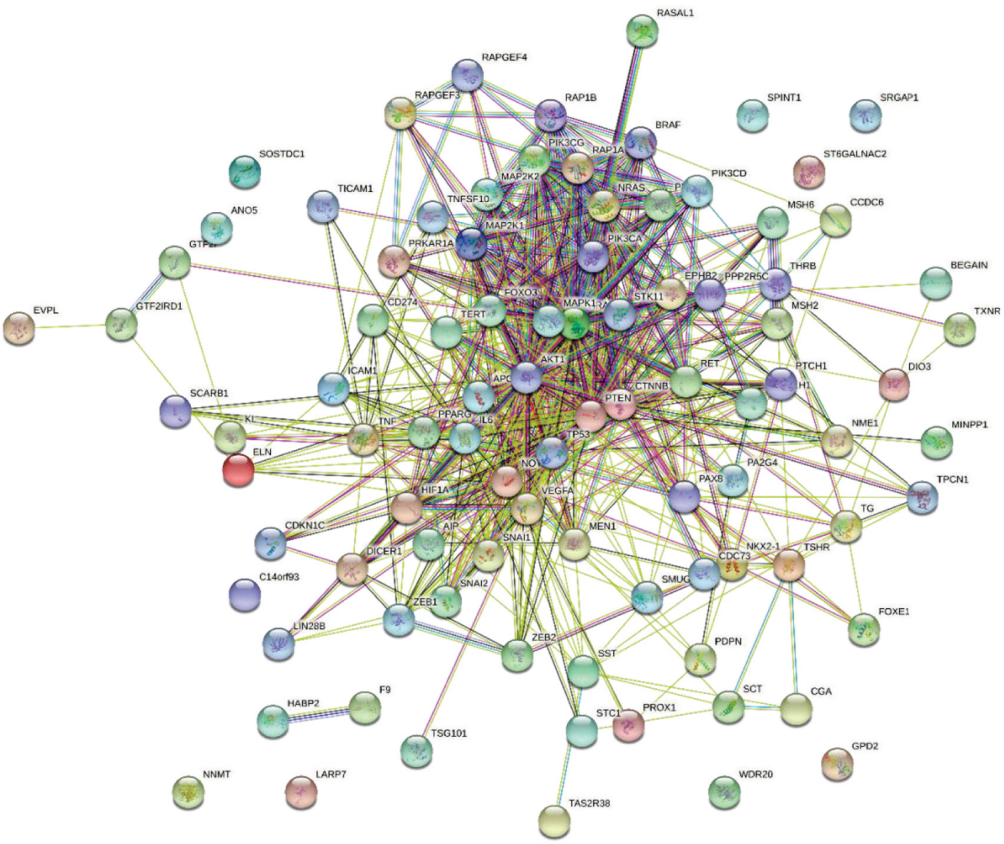


FIGURE 2 | Gene interaction network of TC-related genes.

O is output. σ() is activation function. w is the coefficient of linear relationship, b is bias model parameters.

Using DNN network architecture to identify the interaction pattern between TC and gene, we need to define the objective function to measure the loss of model fitting.

$$J(w, b, x, y) = \frac{1}{2} \parallel a^L - y \parallel_2^2 \tag{6}$$

The process of training DNN is to minimize the loss function. The parameters of DNN model is shown in **Table 1**.

Due to the small sample set, DNN is used as a weak classifier. In order to make the model more accurate, we introduced AdaBoost.

First, set the initial weight of each sample to 1/N. Then, training samples to get the first DNN model, test this DNN model, increase the weight of the unclassified correct samples and reduce the weight of the classified correct samples. At the same time, the weight of the DNN model is obtained. Repeating the above process, we can get multiple DNN models and corresponding weights, thereby obtaining the final strong classifier.

The error rate of each model can be calculated as following:

$$err_m = \sum_{i=1}^{N} w_{mi} || (G_m(x_i) \neq y_i) \tag{7}$$

The weight of the model is:

$$a_m = \frac{1}{2} \log \frac{1 - err_m}{err_m} \tag{8}$$

The final model is the summary of all DNN models:

$$G(x) = \arg\max \sum_{m:\ G_m(x)=y} a_m \tag{9}$$

TABLE 1 | The parameters of DNN model.

Structure	Parameters
Layer 1	Units: 256 Activation function: Tanh Dropout rate: 0.3
Layer 2	Units: 128 Activation function: Tanh Dropout rate: 0.2
Layer 3	Units: 2 Activation function: Tanh
Loss function	Binary cross entropy
Optimizer	RMSprop

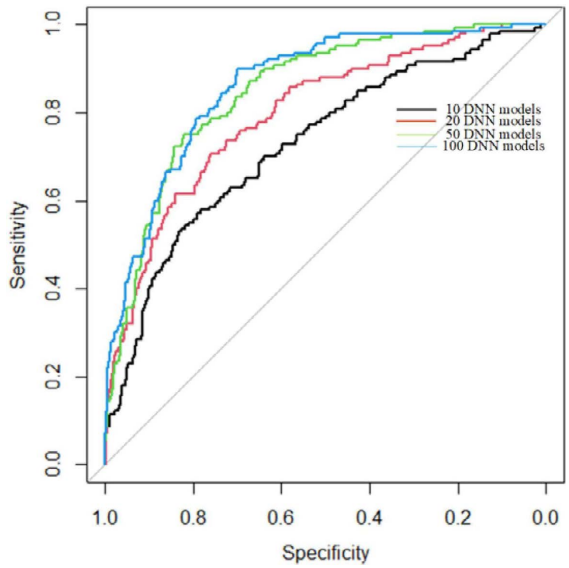


FIGURE 3 | ROC curves of different numbers of DNN models.

RESULTS

Since we used DNN as a weak classifier and the number of DNN models is set by experience, we used 5-cross validation to find the best number of DNN models of ADNN. The process of 5-cross validation is to divide whole sample set into five groups. We used one group for testing and four groups for training each time. After repeating five times, each group has been tested once and trained four times. We use 10, 20, 50, and 100 DNN models to build ADNN, respectively. The experiment results are shown as **Figure 3**.

In **Figure 3**, the AUC are 0.73, 0.79, 0.85, 0.86 for 10, 20, 50, 100 DNN models, respectively. As shown in **Figure 3**, the AUC of 50 and 100 DNN models are very close. However, constructing 100 DNN models is twice time consuming than 50 DNN models. Therefore, we used 50 DNN models to build ADNN model.

In order to show the superiority of ADNN, we compared it with several other methods such as DNN, ASVM, ANB, Random Forest (RF). ASVM is the fusion of Adaboost and Support Vector Machine (SVM). ANB is the fusion of Adaboost and Naïve Bayes (NB).

The comparison results is listed in **Table 2**.

Compared ADNN with ASVM, we can find that DNN is more suitable than SVM in prioritizing susceptible genes for thyroid cancer. Compared ADNN with DNN, we can find that Adaboost can significantly increase the accuracy of prioritizing susceptible genes for thyroid cancer.

TABLE 2 | Comparison of ADNN and other methods.

Method	AUC	AUPR
ADNN	0.85	0.81
DNN	0.69	0.65
ASVM	0.82	0.79
ANB	0.76	0.71
RF	0.78	0.76

CONCLUSION

Genetic factors are an important cause of thyroid cancer. Exploring the susceptibility genes of thyroid cancer is the key to understanding the pathogenesis and developing new treatment options. Collecting samples from patients and healthy individuals and analyzing differential gene expression is very costly and time-consuming. After years of research, researchers have found only 147 genes related to thyroid cancer. The role of these genes in thyroid cancer is unknown. In addition, there are more genes associated with thyroid cancer. To prioritize susceptible genes of thyroid cancer in large-scale, we proposed a novel method, named ADNN, to identify TC-related genes by gene interaction network. We constructed gene interaction network based on known TC-related genes and used RW to encode the features of genes. Then, we fused Adaboost with DNN to classify whether a gene is related to TC and obtain the probability of genes associated with TC. We get the best number of DNN models needed to construct ADNN through experiments. Finally, we compared ADNN with several other methods. Overall, we propose a precise and efficient method for prioritizing susceptible genes for thyroid cancer.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

L-KZ, C-LX, SJ, J-HM, and BX participated in the study design. L-KZ, C-LX, SJ, X-XG, J-HF, W-SC, C-ZL, and FS analyzed the data. L-KZ, C-LX, SJ, and X-YD wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

Financial support came from the Zhongshan Medical Scientific Research Project (2021A020606) and the Guangzhou Science and Technology Plan Project (202102080170).

REFERENCES

Agrawal, N., Akbani, R., Aksoy, B. A., Ally, A., Arachchi, H., Asa, S. L., et al. (2014). Integrated genomic characterization of papillary thyroid carcinoma. *Cell* 159, 676–690. doi: 10.1016/j.cell.2014.09.050

Ceolin, L., Duval, M., Benini, A. F., Ferreira, C. V., and Maia, A. L. (2019). Medullary thyroid carcinoma beyond surgery: advances, challenges, and perspectives. *Endocr. Relat. Cancer* 26, R499–R518. doi: 10.1530/erc-18-0574

Chen, L., Zhao, K., Li, F., and He, X. (2020). Medullary thyroid carcinoma with elevated serum CEA and normal serum calcitonin after surgery: a case report and literature review. *Front. Oncol.* 10:526716. doi: 10.3389/fonc.2020.526716

Dralle, H., Machens, A., Basa, J., Fatourechi, V., Franceschi, S., Hay, I. D., et al. (2015). Follicular cell-derived thyroid cancer. *Nat. Rev. Dis. Primers* 1:15077. doi: 10.1038/nrdp.2015.77

Lee, J. H., Ha, E. J., and Kim, J. H. (2019). Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT. *Eur. Radiol.* 29, 5452–5457. doi: 10.1007/s00330-019-06098-8

Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., et al. (2019). Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet. Oncol.* 20, 193–201. doi: 10.1016/s1470-2045(18)30762-9

Maniakas, A., Dadu, R., Busaidy, N. L., Wang, J. R., Ferrarotto, R., Lu, C., et al. (2020). Evaluation of overall survival in patients with anaplastic thyroid carcinoma, 2000-2019. *JAMA Oncol.* 6, 1397–1404. doi: 10.1001/jamaoncol.2020.3362

Molinaro, E., Romei, C., Biagini, A., Sabini, E., Agate, L., Mazzeo, S., et al. (2017). Anaplastic thyroid carcinoma: from clinicopathology to genetics and advanced therapies. *Nat. Rev. Endocrinol.* 13, 644–660. doi: 10.1038/nrendo.2017.76

Penna, G. C., Vaisman, F., Vaisman, M., Sobrinho-Simões, M., and Soares, P. (2016). Molecular markers involved in tumorigenesis of thyroid carcinoma: focus on aggressive histotypes. *Cytogenet. Genome. Res.* 150, 194–207. doi: 10.1159/000456576

Petrulea, M. S., Plantinga, T. S., Smit, J. W., Georgescu, C. E., and Netea-Maier, R. T. (2015). PI3K/Akt/mTOR: a promising therapeutic target for non-medullary thyroid carcinoma. *Cancer Treat. Rev.* 41, 707–713. doi: 10.1016/j.ctrv.2015.06.005

Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015:bav028. doi: 10.1093/database/bav028

Sahu, R., and Pattanayak, S. P. (2020). Strategic developments & future perspective on Gene therapy for breast cancer: role of mTOR and Brk/PTK6 as molecular targets. *Curr. Gene. Ther.* 20, 237–258. doi: 10.2174/1566523220999200731002408

Salvatore, G., De Falco, V., Salerno, P., Nappi, T. C., Pepe, S., Troncone, G., et al. (2006). BRAF is a therapeutic target in aggressive thyroid carcinoma. *Clin. Cancer Res.* 12, 1623–1629. doi: 10.1158/1078-0432.ccr-05-2378

Sanguedolce, F., Cormio, A., Bufo, P., Carrieri, G., and Cormio, L. (2015). Molecular markers in bladder cancer: Novel research frontiers. *Crit. Rev. Clin. Lab. Sci.* 52, 242–255. doi: 10.3109/10408363.2015.1033610

Sui, T., Ma, L., Bai, X., Li, Q., and Xu, X. (2014). Resveratrol inhibits the phosphatidylinositide 3-kinase/protein kinase B/mammalian target of rapamycin signaling pathway in the human chronic myeloid leukemia K562 cell line. *Oncol. Lett.* 7, 2093–2098. doi: 10.3892/ol.2014.2014

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 2020:bbaa044. doi: 10.1093/bib/bbaa044

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Trovisco, V., Vieira de Castro, I., Soares, P., Máximo, V., Silva, P., Magalhaes, J., et al. (2004). BRAF mutations are associated with some histological types of papillary thyroid carcinoma. *J. Pathol.* 202, 247–251. doi: 10.1002/path.1511

Tsukatani, T., Niwa, H., Komori, T., Yoneyama, T., Tsuji, H., Michigishi, T., et al. (2020). Superior mediastinal lymphadenopathy by silicosis mimicking metastasis of papillary thyroid carcinoma–Case report and literature review. *Auris Nasus Larynx* 47, 1054–1057. doi: 10.1016/j.anl.2019.12.003

Vijayan, R., Nair, V., Menon, U., and Kumar, H. (2021). A rare RET mutation in an Indian pedigree with familial medullary thyroid carcinoma. *Indian J. Cancer* 58, 98–100. doi: 10.4103/ijc.IJC_639_19

Xia, Y., Li, X. Q., and Sun, W. (2020). Applications of recombinant adenovirus-p53 gene therapy for cancers in the clinic in china. *Curr. Gene. Ther.* 20, 127–141. doi: 10.2174/1566523220999200731003206

Xing, M. (2019). Genetic-guided risk assessment and management of thyroid cancer. *Endocrinol. Metab. Clin.* 48, 109–124. doi: 10.1016/j.ecl.2018.11.007

Zanella, A. B., Scheffel, R. S., Weinert, L., Dora, J. M., and Maia, A. L. (2021). New insights into the management of differentiated thyroid carcinoma in children and adolescents. *Int. J. Oncol.* 58:13. doi: 10.3892/ijo.2021.5193

Zhang, S., He, X., Zhang, R., and Deng, W. (2021). LncR2metasta: a manually curated database for experimentally supported lncRNAs during various cancer metastatic events. *Brief. Bioinform.* 22:bbaa178. doi: 10.1093/bib/bbaa178

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 2020:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020c). MRTFB regulates the expression of NOMO1 in colon. *Proc. Nat. Acad. Sci.* 117, 7568–7569. doi: 10.1073/pnas.2000499117

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021a). Prediction and collection of protein–metabolite interactions. *Brief. Bioinform.* 2021:bbab014. doi: 10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Zhong, Xie, Jiang, Deng, Gan, Feng, Cai, Liu, Shen, Miao and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 26 August 2021
doi: 10.3389/fcell.2021.712931

Recognizing Pattern and Rule of Mutation Signatures Corresponding to Cancer Types

Lei Chen[1,2†], Xianchao Zhou[3,4†], Tao Zeng[5], Xiaoyong Pan[6], Yu-Hang Zhang[7], Tao Huang[5,8], Zhaoyuan Fang[9]* and Yu-Dong Cai[1]**

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [3] School of Life Sciences and Technology, ShanghaiTech University, Shanghai, China, [4] Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [5] CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [6] Key Laboratory of System Control and Information Processing, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Ministry of Education of China, Shanghai, China, [7] Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [8] Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China, [9] Zhejiang University-University of Edinburgh Institute, Zhejiang University School of Medicine, Haining, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Yun Li,
University of Pennsylvania, United States
Jie Yin,
Zhejiang University, China

***Correspondence:**
Tao Huang
tohuangtao@126.com
Zhaoyuan Fang
fangzhaoyuan@sibs.ac.cn
Yu-Dong Cai
cai_yud@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *21 May 2021*
Accepted: *02 July 2021*
Published: *26 August 2021*

Citation:
Chen L, Zhou X, Zeng T, Pan X, Zhang Y-H, Huang T, Fang Z and Cai Y-D (2021) Recognizing Pattern and Rule of Mutation Signatures Corresponding to Cancer Types. Front. Cell Dev. Biol. 9:712931. doi: 10.3389/fcell.2021.712931

Cancer has been generally defined as a cluster of systematic malignant pathogenesis involving abnormal cell growth. Genetic mutations derived from environmental factors and inherited genetics trigger the initiation and progression of cancers. Although several well-known factors affect cancer, mutation features and rules that affect cancers are relatively unknown due to limited related studies. In this study, a computational investigation on mutation profiles of cancer samples in 27 types was given. These profiles were first analyzed by the Monte Carlo Feature Selection (MCFS) method. A feature list was thus obtained. Then, the incremental feature selection (IFS) method adopted such list to extract essential mutation features related to 27 cancer types, find out 207 mutation rules and construct efficient classifiers. The top 37 mutation features corresponding to different cancer types were discussed. All the qualitatively analyzed gene mutation features contribute to the distinction of different types of cancers, and most of such mutation rules are supported by recent literature. Therefore, our computational investigation could identify potential biomarkers and prediction rules for cancers in the mutation signature level.

Keywords: cancer, subtype, mutation signature, pattern, rule, classification

INTRODUCTION

Cancer is generally defined as a systematic disease with abnormal cell proliferation and invasion potentials. The general symptoms of cancer include cough, weight loss, lump, and abnormal bleeding, depending on the pathological regions of cancer. Such symptoms are common but not specific in cancers, which may also be shared in other diseases. Genetic factors from either noxious environmental factors or inherited variations trigger the initiation and progression of cancers (Anand et al., 2008; Tang et al., 2018). The regulation and functioning of genetic alterations vary at different omics levels with alternative pathological potentials. However, one of the shared and most

common pathological alterations at the phenotypic level is abnormal cell proliferation, which is consistently regulated at the transcriptomics and proteomics level, apart from genomic regulations (Feitelson et al., 2015; Li Z. et al., 2020). Generated from and regulated by tumor microenvironment, cancer is consistently shaped by surrounding cells and tissues, making tumor microenvironment a crucial factor for tumorigenesis in most kinds of cancers, including lung, colorectal, cervical, and breast cancers.

Various biological or pathological behavior affect cancer during initiation, progression, and metastasis. For instance, a bidirectional functioning has been identified on autophagy, one of the most common biological behavior during tumorigenesis. Autophagy suppresses malignant transformation during cancer initiation (Kondo et al., 2005); however, with the invasion and progression of cancer, autophagy may promote the malignant proliferation of cancer cells (de Visser et al., 2006; Nadia and Ramana, 2020). Apart from biological processes, such as autophagy, cancer immune responses are another key factor affecting cancer initiation, progression, and metastasis. Complicated relationships between immune cells and general tumor biological processes, such as initiation, progression, and metastasis, have been recognized in multiple cancer subtypes. Similarly, oxygen-mediated cell damage (Srinivas et al., 2019), growth factor abnormality (Normanno et al., 2017; Umino et al., 2018), and various similar malignant alterations have been recognized during tumorigenesis, describing a complicated profiling of tumor biology at different omics levels.

Clustering cancer related SNPs have been one of the most important research fields in cancer biology for a long time. Multiple previous clustering methods have been presented to summarize the mutational patterns of different cancers. According to one of the most famous cancer mutation databases, COSMIC (Catalogue of Somatic Mutations In Cancer) (Tate et al., 2019), three groups of mutational signatures, including Single Base Substitution (SBS), Doublet Base Substitution (DBS), and Small insertions and deletions (ID), have been summarized. However, such clustering only involves in the molecular biological features of cancer mutations without functional interpretation. Furthermore, in 2019, investigators from University Medical Centre Utrecht summarized current contribution of cancer mutations on clinical diagnosis and identified functional interpretations of existed identified cancer mutation biomarkers (Van Hoeck et al., 2019). However, such results only summarized the effects of mutations with validated functional interpretation. As for computational contribution of cancer mutation clustering, few methods have been presented and these algorithms can only recognize single nucleotide variants (SNVs) and copy number variants (CNVs) (Maura et al., 2019), but not mutation patterns involving multiple base pairs. Therefore, as a summary, the method we presented here have two major innovations: (1) we are targeting essential malignant signatures at the level of mutation patterns not just SNVs or just CNVs; (2) we established quantitative rules to evaluate the contribution of mutation patterns to tumorigenesis.

Among such biological characteristics, cancer-associated genetic mutations are one of the key and essential malignant signatures related to the initiation, invasion, and metastasis of cancers. Specific mutation patterns (ACA to AAA, ACC to AAC, ACG to AAG, etc.) on target regions/genes of the genome have been widely reported to participate in tumorigenesis. For instance, mutant *KAI1*, regulating a batch of downstream tumor-associated genes, has been observed during metastasis in human prostate cancer (Dong et al., 1996; Yang L. et al., 2020). hMLH1 functioning in DNA mismatch repair is one of the most important molecular biomarkers for hereditary non-polyposis colon cancer at the genomics level (Bronner et al., 1994; Ran et al., 2020). The frequency of *Smad4* gene mutation in human colorectal cancer is higher than in any type of cancer (Miyaki et al., 1999; Zhang et al., 2020). Other genes, such as P53 (Fujimoto et al., 1992) and VHL (Kim and Kaelin, 2004), influence cancer development to some extent. However, studies on the mutation features and rules that affect cancers are limited. In the present study, we gave a computational investigation on mutation profiles of cancer samples in 27 types. The powerful feature selection method, Monte Carlo Feature Selection (MCFS) (Dramiński et al., 2007), was adopted to analyze such profiles. A feature list was generated. Then, the incremental feature selection (IFS) (Liu and Setiono, 1998) was applied to this list to extract essential mutation features, find out interesting rules and construct efficient classifiers. As a result, 207 rules and many mutation features related to the 27 types of cancers were obtained. We discussed the top 37 mutation features corresponding to different types of cancers. Our study may serve as a reference in establishing a novel qualitative and quantitative standard in identifying tumor type-specific mutation patterns for tumor classification, and thus provide a new tool for the study of tumorigenesis mechanism based on mutation signatures.

MATERIALS AND METHODS

Datasets

We downloaded the relative mutation frequency of 96 mutation types in 2,892 patients from 27 cancer types (Lawrence et al., 2013). The sample sizes of each cancer type are listed in **Table 1**. The relative mutation frequency of each mutation type in each cancer patient was calculated by Lawrence et al. (2013) and defined as $R_{cs} = (\frac{n_{cs}}{N_{cs}})/(\sum_c n_{cs}/\sum_c N_{cs})$, where s is the sample, c is the mutation type, n_{cs} is the number of observed mutations, and N_{cs} is the number of bases with enough coverage ($\geq$14 reads in tumor cases and $\geq$8 reads in normal cases) to observe mutation. The mutation types were summarized according to their base pair changes identified in a data set of 3,083 tumor–normal pairs across 27 tumor types (Lawrence et al., 2013). Mutations were specified in the middle of three base pair patterns with all possible mutational directions. The detailed mutation types are provided in **Supplementary Table 1**. Several studies have suggested that the mutation signatures of different cancers vary and involve combinations of the above mutation types (Alexandrov et al., 2013; Lawrence et al., 2013; Huang et al., 2018; Wojtowicz et al., 2019). We investigated the cancer mutation signatures quantitatively through advanced machine learning

methods and identified the mutation rules for explaining and understanding each cancer type.

Feature Selection

We applied feature selection to discriminate influential mutation types from the unrelated ones in the dataset. First, MCFS (Dramiński et al., 2007) was used to evaluate the importance of each mutation type. A feature list was generated. Then, a set of optimal mutation types with strong distinctions between different cancer types was obtained by applying IFS (Liu and Setiono, 1998; Ma et al., 2020) with a supervised classifier on such list.

Monte Carlo Feature Selection

In this study, we used the MCFS method to assess the importance of mutation types. MCFS is a feature selection method based on the random features of the original features (Dramiński et al., 2007). Given a dataset with d features (in this study, 96 mutation types were deemed as features), MCFS first randomly constructed p feature subsets, each of which contains m features, where m is much smaller than d. Second, for each feature subset, t decision trees are built. Each tree is constructed based on 66% samples that are randomly selected from the original dataset, and the rest samples are used to test such tree. Thus, $p \times t$ trees can be constructed in total. Based on these trees, the importance of each feature, called relative importance (RI) score in MCFS, is evaluated by the following equation

$$RI_f = \sum_{\tau=1}^{pt} (wAcc)^u IG(n_f(\tau))\left(\frac{no.in\ n_f(\tau)}{no.in\ \tau}\right)^v, \quad (1)$$

where wAcc is the weighted accuracy of the decision tree τ and $n_f(\tau)$ is a node of feature f in decision tree τ. The information gain of $n_f(\tau)$ is expressed as $IG(n_f(\tau))$, and $no.in\ n_f(\tau)$ is the number of training samples in $n_f(\tau)$. u and v are two different weighting factors.

The MCFS program used in this study was retrieved from http://www.ipipan.eu/staff/m.draminski/mcfs.html. For convenience, default parameters were adopted. In detail, $u = v = 1$, $p = 3{,}000$, $t = 5$, $m = 5$. The 96 features (mutation types) were analyzed by the MCFS program. Each feature was assigned a RI score. Evidently, a feature with a high RI score was more important than that with a low RI score. Thus, we sorted all 96 features with the decreasing order of their RI scores. For formulation, this list was denoted by F.

Incremental Feature Selection

Incremental Feature Selection (IFS) (Liu and Setiono, 1998) is a feature selection method that filters out a set of optimal features to accurately distinguish different sample classes. As mentioned in section "Monte Carlo Feature Selection," a feature list F was generated by MCFS method. Clearly, the high-ranked features should have positive contributions to classification and can help the classification algorithm to produce good performance. To perform IFS, we first created a series of feature subsets with a step 1 from the feature list F. In detail, the first feature subset included the top feature in the list F, the second feature subset contained the top two features, and so forth. For each constructed feature subset, a random forest (RF) classifier was built based on samples represented by features in the subset and it was further evaluated by 10-fold cross-validation (Kohavi, 1995; Li J. et al., 2020; Zhou et al., 2020; Liu et al., 2021; Pan et al., 2021; Zhang et al., 2021a,b,c; Zhu et al., 2021). After testing all feature subsets, we obtained the optimal feature subset with the optimal performance. This feature subset was termed as the optimal feature subset and the classifier with such subset was called the optimal classifier.

TABLE 1 | Sample sizes of 27 cancer types.

Index	Cancer type	Sample size
1	Acute myeloid leukemia	119
2	Bladder	35
3	Breast	120
4	Carcinoid	21
5	Cervical	20
6	Chronic lymphocytic leukemia	87
7	Colorectal	230
8	Diffuse large B-cell lymphoma	49
9	Esophageal adenocarcinoma	76
10	Ewing sarcoma	9
11	Glioblastoma multiforme	213
12	Head and neck	165
13	Kidney clear cell	212
14	Kidney papillary cell	11
15	Low-grade glioma	55
16	Lung adenocarcinoma	327
17	Lung squamous cell carcinoma	177
18	Medulloblastoma	16
19	Melanoma	121
20	Multiple myeloma	62
21	Neuroblastoma	61
22	Ovarian	382
23	Pancreas	9
24	Prostate	196
25	Rhabdoid tumor	3
26	Stomach	87
27	Thyroid	29

Synthetic Minority Over-Sampling Technique

Synthetic Minority Over-Sampling Technique (SMOTE) (Chawla et al., 2002; Chao et al., 2019; Yang X.F. et al., 2020) is a classic technology used to address the potential sample imbalance issue during classification learning. SMOTE can add new samples into the minority class as the same number of samples in the majority class, in an oversampling manner. SMOTE includes several computational steps: (1) it randomly selects a sample x in the minority class; (2) it finds k neighboring samples in such class with x; (3) it randomly select again a sample y from these neighboring samples to generate a new sample z by a linear combination of x and y; (4) it places each new sample z into the minority class; and (5) it repeats the above steps with predefined times. We directly adopted SMOTE in WEKA (Witten and Frank, 2005).

When evaluating the performance of classifiers in the IFS method, SMOTE was used to decrease the influence of imbalanced problem. In detail, in each round of 10-fold cross-validation, the training dataset was processed by SMOTE so that all classes had same number of samples. The classifier was built on such dataset and further used to predict testing samples.

Random Forest

A RF is a classifier that is a predictive model for establishing classification and regression problems. It determines the output class for one sample by aggregating votes from different decision trees (Breiman, 2001). As one of the common methods of machine learning, we built a RF by constructing a large number of decision trees. Averaging the predictions of all decision trees to reduce the variance will slightly increase the bias of the predictions, but at the same time, the performance of the model will be considerably improved while avoiding over-learning. As one of the common methods in the field of machine learning, it has wide applications in tackling different biological problems (Pugalenthi et al., 2011; Pan et al., 2014; Marques et al., 2016; Zhao X. et al., 2018; Ru et al., 2019; Zhang et al., 2019; Zhao R. et al., 2019; Zhao X. et al., 2019; Jia et al., 2020; Liang et al., 2020). In this study, we used the tool "RandomForest" in Weka (Witten and Frank, 2005), which implements the RF. Default parameters were used, where the number of decision tree was 10.

Rule Learning

In this study, we also used the interpretable machine learning method repeated incremental pruning to produce error reduction (RIPPER) (Cohen, 1995) to learn the classification rules. In RIPPER, each rule is an IF-ELSE statement; for instance, if gene1 $>$ 1.3 and gene2 $<$ 5, then breast cancer occurs. The learned rules can be used to make predictions for new samples. We used the RIPPER implemented tool "JRip" in WEKA.

Performance Measurement

Matthew's Correlation Coefficient (MCC) (Matthews, 1975) is a commonly used method for estimating performance measurements of classification models (Chen et al., 2017a,b, 2018; Cui and Chen, 2019). However, the original MCC was designed for binary classification problem. In this study, 27 cancer types were involved. Thus, we used the MCC in multi-class version, which was proposed by Gorodkin (2004). To calculate such MCC, two matrices X and Y must be constructed first, where X stands for the true labels of all samples and Y represents the predicted labels of all samples. Then, the MCC in multi-class can be computed by

$$MCC = \frac{cov(X, Y)}{\sqrt{cov(X, X)cov(Y, Y)}}, \quad (2)$$

where $cov(\cdot)$ indicates the correlation coefficient of two matrices. Similar to the original MCC, MCC in multi-class ranges between –1 and 1. A high MCC implies the good performance. For convenience, we still called MCC in multi-class as MCC in the following text.

Besides, we also reported the accuracy of each cancer type and overall accuracy to give a complete picture on the performance of each classifier.

RESULTS

In this study, several machine learning methods were adopted to investigate the mutation profiles of cancer samples in 27 types. The entire procedures are illustrated in **Figure 1**.

Results of MCFS Method

The mutation profiles were first analyzed by the MCFS method to assess the importance of each mutation type. Each mutation type was assigned a RI score, which is listed in **Supplementary Table 2**. Then, all mutation types were sorted by the decreasing order of their RI scores, resulting in a feature list F. Such list is also provided in **Supplementary Table 2**.

Results of IFS Method

A feature list F was generated according to the results of MCFS. Such list was fed into the IFS method. Two classification algorithms: RF and RIPPER, were integrated in the IFS method. For RF, its performance on each feature subset is provided in **Supplementary Table 3**. An IFS curve was plotted, as shown in **Figure 2**, for an easy observation. The number of features was set as X-axis and MCC was set as the Y-axis. It can be observed that the highest MCC was 0.772 when top 96 features were adopted. Surprisingly, all mutation features were used in this case, indicating that each mutation type gave less or more contributions to the distinction of different cancer types. Accordingly, all mutation features comprised the optimal feature subset for RF and the RF classifier with these features was called the optimal RF classifier. The overall accuracy of such classifier was 0.784, as listed in **Table 2**. Its detailed performance on 27 cancer types is illustrated in **Figure 3**. 14 cancer types received the accuracies higher than 0.900. All these suggested the good performance of such RF classifier. Furthermore, we also employed the rule learning algorithm, RIPPER, to do the same procedures. The performance of RIPPER on all feature subsets is also available in **Supplementary Table 3**. Likewise, an IFS curve was plotted, as illustrated in **Figure 2**. Evidently, the highest MCC was 0.408 when top 61 features were adopted. Subsequently, the optimal RIPPER classifier was constructed using these 61 features and these features constituted the optimal feature subset for RIPPER. The MCC (0.408) was much lower than that of the optimal RF classifier (0.772). The overall accuracy was 0.443, as listed in **Table 2**, also much lower than that of the optimal RF classifier (0.784). The detailed performance of the optimal RIPPER classifier on all cancer types is illustrated in **Figure 3**. It can be observed that almost all cancer types received lower accuracies than those of the optimal RF classifier. Thus, the optimal RF classifier was much superior to the optimal RIPPER classifier. However, the optimal RF classifier was an absolute black-box classifier, few insights can be extracted from such classifier. The RIPPER classifier was much better in this regard because it can learn some classification rules

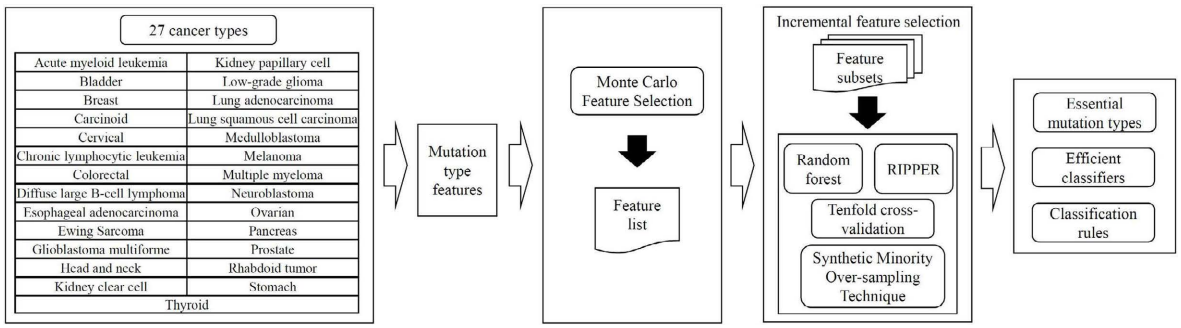


FIGURE 1 | Entire procedures for the investigation on mutation profiles of cancer samples in 27 types. The profiles are analyzed by the MCFS method, resulting in a feature list. Such list is fed into the incremental feature selection, incorporating random forest or RIPPER as the classification algorithm, to extract essential mutation types, build efficient classifiers and construct classification rules.

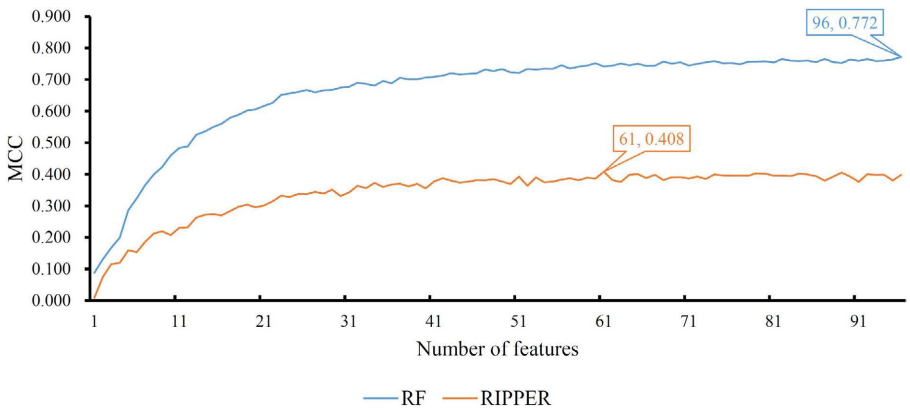


FIGURE 2 | Performance of RF and RIPPER on different feature subsets. The RF yields the highest MCC of 0.772 when all features are used, whereas RIPPER generates the highest MCC of 0.408 when top 61 features are used.

for explaining and understanding particular cancer differences. In detail, generally, several mutation features were involved in one classification rule that can be used to predict one cancer type. These mutation features together with their corresponding thresholds can comprise a mutation pattern on such cancer type. Further investigation on such pattern was helpful to understand the mechanism of this cancer type in the mutation signature level.

From **Table 1**, we can see that some cancer types contained much more samples than other types, that is, the type sizes were of great differences. Here, we investigated the performance of two optimal classifiers on cancer types with different sizes. To this end, we classified 27 cancer types into three categories. The first category contained types with less than 10 samples, the second category included the types with 10–100 samples, and the third category contained types with more than 100 samples. For convenience, these three categories were called small, middle and large cancer types. The performance of the optimal RF and RIPPER classifiers on three categories is illustrated in **Figure 4**. It can be observed that the middle cancer type received relative higher accuracies, whereas the small cancer type received slightly higher accuracies than the large cancer type. The reason may be the application of SMOTE. It can increase the performance on minor classes (small cancer type) and decrease the performance on major classes (large cancer type).

TABLE 2 | Performance of IFS with RF and RIPPER for classifying samples from different cancers.

Classifier	Number of features	Overall accuracy	MCC
RF	96	0.784	0.772
RIPPER	61	0.443	0.408

Classification Rules

The optimal RIPPER classifier used the top 61 mutation features. Accordingly, the RIPPER was applied on all cancer samples that were represented by these 61 features. As a result, 207 rules were obtained, which are provided in **Supplementary Table 4**. Each cancer type had at least one rules. The number of rules on each cancer type is shown in **Figure 5**. The cancer types "Esophageal adenocarcinoma" and "Neuroblastoma" had most rules, whereas "Kidney clear cell" and "Rhabdoid tumor" had only one rule. In section "Mutation Pattern Rules Associated With Cancer Subgrouping," some representative rules would be discussed.

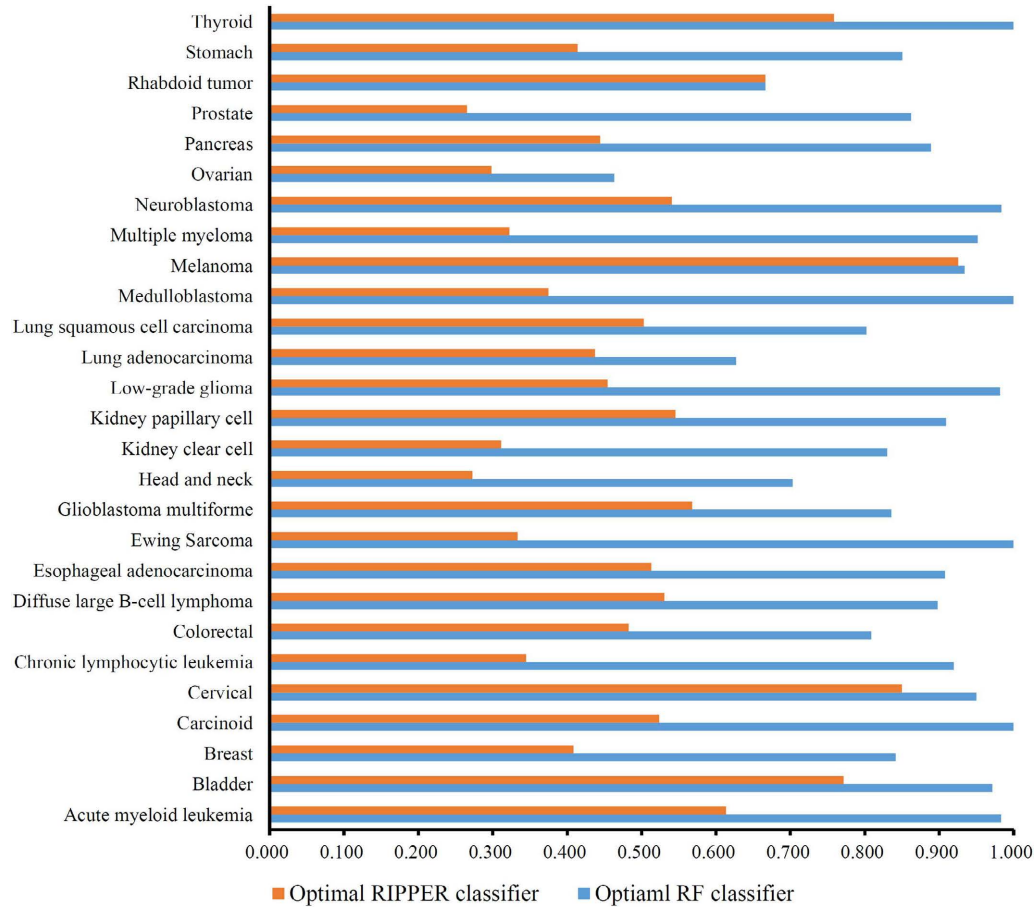


FIGURE 3 | Accuracy on each cancer type yielded by the optimal RF and RIPPER classifiers. The optimal RF classifier gives high performance (accuracy > 0.900) on 14 cancer types, indicating its good performance. The optimal RIPPER classifier provides much lower performance.

DISCUSSION

We summarized optimal features describing the distinct mutation types to generate an applicable classifier for all the 27 cancer subtypes. The contribution of typical mutation patterns was functionally connected to certain cancer biological behavior, such as the initiation and progression of tumorigenesis. The top 37 features were supported by previous studies as biomarker candidates. In addition to these potential cancer biomarkers, we identified 27 rules of mutation types associated with these 27 cancers. Such group of quantitative rules may contribute to the detailed classification of cancers and serve as guidance in future research.

Mutation Patterns Associated With Cancer Subgrouping

First, we screened a group of functional mutation patterns with different distributions in various cancer types. Recent publications confirmed such mutation patterns consistent with our prediction. Here, we selected the top five mutation patterns for following detailed discussions.

The top mutation pattern ACG to ATG refers to amino acid substitution from Thr to Met. Recent publications indicated that such mutation pattern exists in multiple cancer-associated genes, including TP53, PRSS1, and XRCC3. For TP53, such mutation pattern has been recognized in multiple tumor types, especially for head and neck cancer (Boyle et al., 1993) and lung cancer (Veldore et al., 2015), rather than other malignant proliferative diseases, such as esophageal adenocarcinoma and melanoma; this finding reflects the distinctive potentials of such mutation pattern. For PRSS1, such mutation pattern has only been found in pancreatic cancer, validating our prediction (Yi et al., 2016). Similarly, XRCC3 has such a mutation pattern in multiple cancer types, including thyroid cancer (Wang et al., 2015) and breast cancer (Lee et al., 2007). Therefore, as we have mentioned above, this mutation pattern has been discovered in limited tumor types but not in all tumor categories, validating our prediction of mutation pattern distinctive with certain tumor types.

CCG to CTG, indicating the transition from Pro to Leu, is another mutation pattern we identified. According to previous studies, such mutation pattern in protein IL-10 contributes to the anti-tumor immune responses in prostate cancer but not in other tumor types, corresponding with our prediction

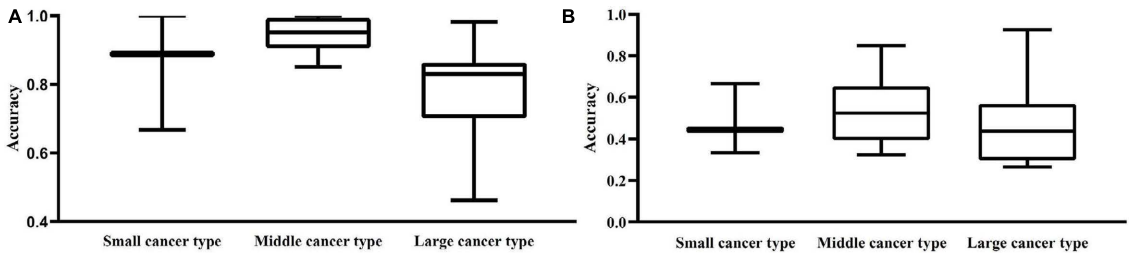


FIGURE 4 | Box plot to show the performance of the optimal RF and RIPPER classifier on three different cancer types. Small cancer type includes the types with less than 10 samples; middle cancer type contains types containing 10–100 samples; larger cancer type contains types with more than 100 samples. **(A)** Box plot for the optimal RF classifier; **(B)** Box plot for the optimal RIPPER classifier. The performance on middle cancer type of two optimal classifiers is best.

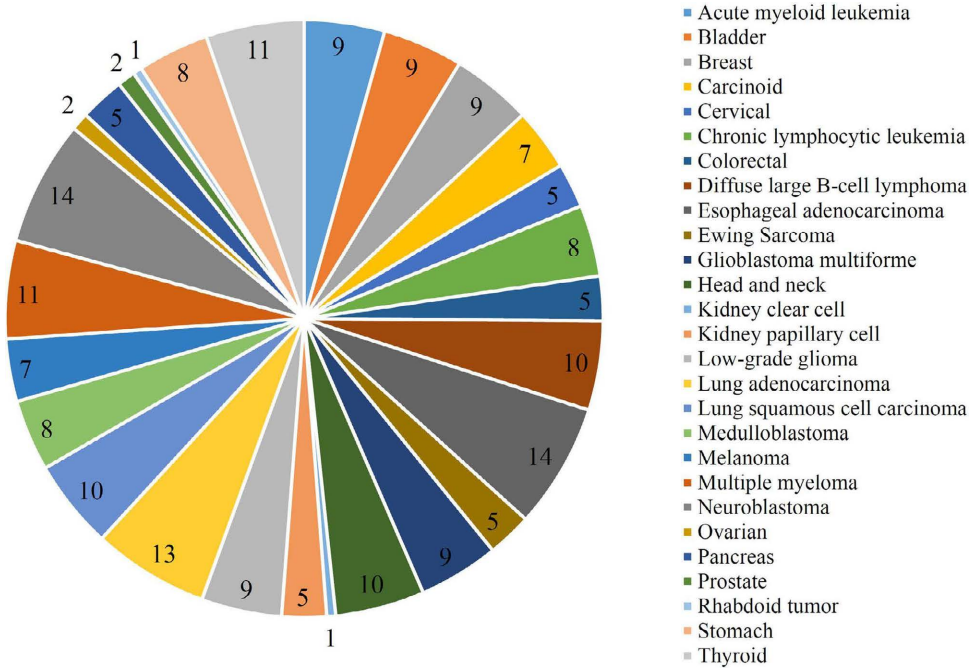


FIGURE 5 | Pie chart to show the number of rules on each cancer type.

(Bandil et al., 2017). Such mutation pattern has also been identified in the gene p16 in multiple tumor types but not in all of our candidate tumor types (Zhang and Peng, 2002), further validating our analysis.

The mutation pattern TCG to TTG describes the amino acid transition from serine to leucine. According to recent reports, such mutation pattern may also distinguish certain tumor types from the other ones, which has been validated to be pathogenic in a specific oncogene named RET for thyroid carcinoma (Colombo-Benkmann et al., 2008), confirming our prediction. Further, SMAD2 and SMAD4 as the two essential components of the TGF beta-Smad signaling pathway also have such variant pattern contributing to the tumorigenesis of head and neck squamous cell carcinoma (Qiu et al., 2007) in contrast with other tumor types.

Apart from the optimal mutation patterns analyzed above, the mutation patterns from GCG to GTG (i.e., from alanine to valine) and from TCA to TTA (i.e., from serine to leucine) (Zhang and Peng, 2002; Bandil et al., 2017) have been distinctively identified in different tumor types, which is consistent with to our prediction.

Mutation Pattern Rules Associated With Cancer Subgrouping

A total of 207 mutation rules are related to 27 cancers. This work focused on a few representative cancer types and several top mutation rules for each cancer.

Melanoma

The first rule is the mutation of codon CCC to CTC (Leu to Pro). In 1998, it was reported that the case of melanoma is particularly peculiar, showing Leu-to-Pro mutations in codons 31 and 35, both of which are located in the highly conserved regions (Kumar et al., 1998). P53 mutations in melanoma cell lines, metastases,

and primary tumors include the Leu-to-Pro mutations (Zerp et al., 1999). The second rule is the mutations of codons from CCC to CTC, TCC to TTC, TAA to TTA, and the third rule includes the mutations from CCC to CTC, and CCG to CTG, which are some refinements of the first rule.

Head and Neck

With typical symptoms as a lump or sore, head and neck cancer is a general description of all cancers related to mouth, nose, and other accessory organs around the head. The first rule for head and neck cancer involves mutations of codons TCA to TTA (Ser to Leu), CAG to CGG (Glu to Arg), and TAT to TGT (Cys to Tyr). In 2007, researchers identified an effective variation on SMAD2. Such variation located in exon 8 at codon 276 (Ser to Leu) contributes to the progression of human head and neck squamous cell carcinoma. The mutations of SMAD2 disrupts a famous tumor associated pathway named transforming growth factor β-Smad signaling pathway (Qiu et al., 2007). The second rule in head and neck cancer involves the mutation of codons TCA to TGA (Ser to a stop codon), ACG to ATG (Thr to Met), AAG to AGG (Lys to Arg), GAG to GGG (Glu to Gly), and TCC to TTC (Ser to Phe). TP53 mutation rate increases following the development and progression of head and neck cancer, and Thr to Met is a well-known p53 mutation (Boyle et al., 1993). The third rule in head and neck cancer includes mutations from codons CCG to CAG, GAC to GGC, and GCT to GTT; these mutations are especially related to CYP1 gene, and the genetic polymorphisms of CYP are associated with head and neck cancer (Gattás et al., 2010).

Esophageal Adenocarcinoma

Arising from the esophagus, esophageal adenocarcinoma is one of the most common subtypes of malignancies affecting the digestive tract. The clinical symptoms of such cancer include difficulty in swallowing and weight loss. The first rule in esophageal cancer involves six mutations, such as AAG to ACG, CAT to CGT, GCC to GGC, CCG to CAG, GCA to GAA, and ACA to ATA. A highly significant association exists between P53 mutations in the molecular pathogenesis of esophageal adenocarcinoma and esophageal malignancy, and AAG to ACG is a remarkable type of TP53 mutation (Vaninetti et al., 2008). The second rule in esophageal adenocarcinoma comprises mutations, such as TAA to TTA, TCT to TGT, GCT to GTT, and CCT to CAT. The third rule involves mutations, such as AAG to AGG, CAT to CGT, CAG to CTG, TCG to TAG, and ACC to AAC. The missense polymorphism of human AGT gene (at codon 276, Ser to Leu) is pathogenic for such disease, revealing the specific characteristics of such diseases at the mutational pattern level. In addition, we confirmed the existence of a codon 84 genetic polymorphism previously, which converts leucine to phenylalanine (Deng et al., 1999).

Neuroblastoma

With cancer-associated bone pain, neck, and chest lump as typical clinical symptoms, neuroblastoma derives from nerve tissues with high cellular diversity. The first rule in neuroblastoma is the involvement of five mutations, including ACG to ATG, GCA to GAA, CAC to CCC, CCA to CTA, and TCC to TTC. The second rule comprises the mutations, such as CAC to CCC, TCC to TAC, ACA to ATA, and CCA to CAA. The third comprises the mutations ACA to AAA, TCA to TTA, TCT to TAT, GCC to GTC, and ACC to ATC.

Comparison With COSMIC Database

Here, we further compared our results with previously reported effective mutation patterns in COSMIC database with solid publication supports. Although in COSMIC, the mutations are summarized based on single base not a combination of three constitutive bases, we did find the consistency of COSMIC validated mutation patterns and our results.

For instance, in melanoma, we identified CCC to CTC, TCC to TTC, and CCG to CTG as three typical mutation patterns, which all involved specific cosmic mutation type as C > T. In COSMIC database, more than 23% of patients have such mutation pattern in melanoma associated genomic regions (Tate et al., 2019). Apart from that, such mutation patterns have been identified in melanoma associated genes, like OR4F5 and SAMD11, validating the specific role of such mutational patterns in melanoma.

Furthermore, as for esophageal adenocarcinoma, mutation patterns like GCA to GAA, CCT to CAT, and TCG to TAG are significant mutations identified in this study with the same single base pair alterations as C > A. According to COSMIC database, C > A pattern has been shown to be identified in more than 54% of all patients associated with esophageal adenocarcinoma (Tate et al., 2019), indicating the specific biological functions of such mutation patterns on such cancer subtypes.

Apart from mutation patterns associated with specific cancer subtypes, we also identified a group of effective mutation patterns associated with APOBEC (Chen et al., 2019). As an apolipoprotein B mRNA editing enzyme, APOBEC family has been shown to be associated with multiple cancer mutations and contribute to the variety of cancer mutation burdens. In our study, we also identified some specific APOBEC associated mutation patterns like TCG > TTG and TCA > TTA, validating that identified mutation patterns are associated with the initiation and progression of different cancer subtypes.

Therefore, the mutation patterns identified in this study to be associated with different cancer subtypes have been validated by COSMIC database, implying reliability of our results.

CONCLUSION

All of the qualitatively analyzed mutation signatures contribute to the distinction of different types of cancers. Most of the quantitative analyzed mutation rules are supported by recent literature. Our computational approach could efficiently identify mutation signatures and rules for cancers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.nature.com/articles/nature12213.

AUTHOR CONTRIBUTIONS

TH, ZF, and Y-DC designed the study. LC, XZ, and XP performed the experiments. TZ and Y-HZ analyzed the results. LC and XZ wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), the National Key R&D Program of China (2017YFC1201200), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the National Key R&D Program of China (2018YFC0910403), the National Natural Science Foundation of China (31701151), the Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.712931/full#supplementary-material

Supplementary Table 1 | Information of 96 mutation types.

Supplementary Table 2 | RI scores of mutation types.

Supplementary Table 3 | IFS corresponding to different numbers of features used.

Supplementary Table 4 | Classification rules generated by RIPPER.

REFERENCES

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421.

Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., et al. (2008). Cancer is a preventable disease that requires major lifestyle changes. *Pharm. Res.* 25, 2097–2116. doi: 10.1007/s11095-008-9661-9

Bandil, K., Singhal, P., Dogra, A., Rawal, S. K., Doval, D. C., Varshney, A. K., et al. (2017). Association of SNPs/haplotypes in promoter of TNF A and IL-10 gene together with life style factors in prostate cancer progression in Indian population. *Inflamm. Res.* 66, 1085–1097. doi: 10.1007/s00011-017-1088-5

Boyle, J. O., Hakim, J., Koch, W., Van Der Riet, P., Hruban, R. H., Roa, R. A., et al. (1993). The incidence of p53 mutations increases with progression of head and neck cancer. *Cancer Res.* 53:4477.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Bronner, C. E., Baker, S. M., Morrison, P. T., Warren, G., Smith, L. G., Lescoe, M. K., et al. (1994). Mutation in the DNA mismatch repair gene homologue hMLH 1 is associated with hereditary non-polyposis colon cancer. *Nature* 368:258. doi: 10.1038/368258a0

Chao, L., Wei, L., and Zou, Q. (2019). SecProMTB: a SVM-based Classifier for Secretory Proteins of Mycobacterium tuberculosis with Imbalanced Data Set. *Proteomics* 19:e1900007.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, L., Chu, C., Zhang, Y.-H., Zheng, M.-Y., Zhu, L., Kong, X., et al. (2017a). Identification of drug-drug interactions using chemical interactions. *Curr. Bioinform.* 12, 526–534.

Chen, L., Pan, X., Hu, X., Zhang, Y. H., Wang, S., Huang, T., et al. (2018). Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* 143, 1731–1740. doi: 10.1002/ijc.31554

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017b). Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* 5, 26582–26590. doi: 10.1109/access.2017.2775703

Chen, Z., Wen, W., Bao, J., Kuhs, K. L., Cai, Q., Long, J., et al. (2019). Integrative genomic analyses of APOBEC-mutational signature, expression and germline deletion of APOBEC3 genes, and immunogenicity in multiple cancer types. *BMC Med. Genomics* 12:131. doi: 10.1186/s12920-019-0579-3

Cohen, W. W. (1995). "Fast effective rule induction," in *Proceedings the Twelfth International Conference on Machine Learning*, Tahoe City, CA, 115–123. doi: 10.1016/b978-1-55860-377-6.50023-2

Colombo-Benkmann, M., Li, Z., Riemann, B., Hengst, K., Herbst, H., Keuser, R., et al. (2008). Characterization of the RET protooncogene transmembrane domain mutation S649L associated with nonaggressive medullary thyroid carcinoma. *Eur. J. Endocrinol.* 158, 811–816. doi: 10.1530/eje-07-0817

Cui, H., and Chen, L. (2019). A binary classifier for the prediction of EC numbers of enzymes. *Curr. Proteom.* 16, 381–389.

de Visser, K. E., Eichten, A., and Coussens, L. M. (2006). Paradoxical roles of the immune system during cancer development. *Nat. Rev. Cancer* 6, 24–37. doi: 10.1038/nrc1782

Deng, C., Xie, D., Capasso, H., Zhao, Y., Wang, L. D., and Hong, J. Y. (1999). Genetic polymorphism of human O6-alkylguanine-DNA alkyltransferase: identification of a missense variation in the active site region. *Pharmacogenetics* 9, 81–87. doi: 10.1097/00008571-199902000-00011

Dong, J.-T., Suzuki, H., Pin, S. S., Bova, G. S., Schalken, J. A., Isaacs, W. B., et al. (1996). Down-Regulation of the KAI1 metastasis suppressor gene during the progression of human prostatic cancer infrequently involves gene mutation or allelic loss. *Cancer Res.* 56, 4387–4390

Dramiński, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2007). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Feitelson, M. A., Arzumanyan, A., Kulathinal, R. J., Blain, S. W., Holcombe, R. F., Mahajna, J., et al. (2015). Sustained proliferation in cancer: mechanisms and novel therapeutic targets. *Semin. Cancer Biol.* 35(Suppl.), S25–S54.

Fujimoto, K., Yamada, Y., Okajima, E., Kakizoe, T., Sasaki, H., Sugimura, T., et al. (1992). Frequent association of p53 gene mutation in invasive bladder cancer. *Cancer Res.* 52, 1393–1398.

Gattás, G. J. F., Marcos Brasilino, D. C., Maria Salete, S., Curioni, O. A., Priscila, K., Jose, E. N., et al. (2010). Genetic polymorphisms of CYP1A1, CYP2E1, GSTM1, and GSTT1 associated with head and neck cancer. *Head Neck* 28, 819–826. doi: 10.1002/hed.20410

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Huang, P. J., Chiu, L. Y., Lee, C. C., Yeh, Y. M., Huang, K. Y., Chiu, C. H., et al. (2018). mSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Res.* 46, D964–D970.

Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-Based machine learning model for predicting the metabolic pathways of compounds. *IEEE Access* 8, 130687–130696. doi: 10.1109/access.2020.3009439

Kim, W. Y., and Kaelin, W. G. (2004). Role of VHL gene mutation in human cancer. *J. Clin. Oncol.* 22, 4991–5004. doi: 10.1200/jco.2004.05.061

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, (Mahwah, NJ: Lawrence Erlbaum Associates Ltd), 1137–1145.

Kondo, Y., Kanzawa, T., Sawaya, R., and Kondo, S. (2005). The role of autophagy in cancer development and response to therapy. *Nat. Rev. Cancer* 5, 726–734. doi: 10.1038/nrc1692

Kumar, R., Rozell, B. L., Louhelainen, J., and Hemminki, K. (1998). Mutations in the CDKN2A (p16INK4a) gene in microdissected sporadic primary melanomas. *Int. J. Cancer* 75, 193–198. doi: 10.1002/(sici)1097-0215(19980119)75:2<193::aid-ijc5>3.0.co;2-p

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218.

Lee, S. A., Lee, K. M., Park, S. K., Choi, J. Y., Kim, B., Nam, J., et al. (2007). Genetic polymorphism of XRCC3 Thr241Met and breast cancer risk: case-control study in Korean women and meta-analysis of 12 studies. *Breast Cancer Res. Treat.* 103, 71–76. doi: 10.1007/s10549-006-9348-z

Li, J., Chang, M., Gao, Q., Song, X., and Gao, Z. (2020). Lung cancer classification and gene selection by combining affinity propagation clustering and sparse group lasso. *Curr. Bioinform.* 15, 703–712. doi: 10.2174/1574893614666190171035557

Li, Z., Zhang, T., Lei, H., Wei, L., Liu, Y., Shi, Y., et al. (2020). Research on gastric Cancer's drug-resistant gene regulatory network model. *Curr. Bioinform.* 15, 225–234. doi: 10.2174/1574893614666190722102557

Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of drug side effects with a refined negative sample selection strategy. *Comput. Math. Methods Med.* 2020:1573543.

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230.

Liu, H., Hu, B., Chen, L., and Lu, L. (2021). Identifying protein subcellular location with embedding features learned from networks. *Curr. Proteom.* 17.

Ma, X., Xi, B., Zhang, Y., Zhu, L., Sui, X., Tian, G., et al. (2020). A machine learning-based diagnosis of thyroid cancer using thyroid nodules ultrasound images. *Curr. Bioinform.* 15, 349–358. doi: 10.2174/1574893614666190171091959

Marques, Y. B., De Paiva Oliveira, A., Ribeiro Vasconcelos, A. T., and Cerqueira, F. R. (2016). Mirnacle: machine learning with SMOTE and random forest for improving selectivity in pre-miRNA ab initio prediction. *BMC Bioinformatics* 17:474. doi: 10.1186/s12859-016-1343-8

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta BBA Protein Struct.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., et al. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* 10:2969.

Miyaki, M., Iijima, T., Konishi, M., Sakai, K., Ishii, A., Yasuno, M., et al. (1999). Higher frequency of Smad4 gene mutation in human colorectal cancer with distant metastasis. *Oncogene* 18, 3098–3103. doi: 10.1038/sj.onc.1202642

Nadia, and Ramana, J. (2020). The human OncoBiome database: a database of cancer microbiome datasets. *Curr. Bioinform.* 15, 472–477. doi: 10.2174/1574893614666190902152727

Normanno, N., Denis, M. G., Thress, K. S., Ratcliffe, M., and Reck, M. (2017). Guide to detecting epidermal growth factor receptor (EGFR) mutations in ctDNA of patients with advanced non-small-cell lung cancer. *Oncotarget* 8, 12501–12516. doi: 10.18632/oncotarget.13915

Pan, X. Y., Zhu, L., Fan, Y. X., and Yan, J. C. (2014). Predicting protein-RNA interaction amino acids using random forest based on submodularity subset selection. *Comput. Biol. Chem.* 53, 324–330. doi: 10.1016/j.compbiolchem.2014.11.002

Pan, X., Li, H., Zeng, T., Li, Z., Chen, L., Huang, T., et al. (2021). Identification of protein subcellular localization with network and functional embeddings. *Front. Genet.* 11:626500. doi: 10.3389/fgene.2020.626500

Pugalenthi, G., Kandaswamy, K., Chou, K.-C., Vivekanandan, S., and Kolatkar, P. (2011). RSARF: prediction of residue solvent accessibility from protein sequence using random forest method. *Protein Peptide Lett.* 19, 50–56. doi: 10.2174/092986612798472875

Qiu, W., Schonleben, F., Li, X., and Su, G. H. (2007). Disruption of transforming growth factor beta-Smad signaling pathway in head and neck squamous cell carcinoma as evidenced by mutations of SMAD2 and SMAD4. *Cancer Lett.* 245, 163–170. doi: 10.1016/j.canlet.2006.01.003

Ran, W., Chen, X., Wang, B., Yang, P., Li, Y., Xiao, Y., et al. (2020). Whole-exome sequencing of tumor-only samples reveals the association between somatic alterations and clinical features in pancreatic cancer. *Curr. Bioinform.* 15, 1160–1167. doi: 10.2174/1574893615999200626190346

Ru, X. Q., Li, L. H., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteome Res.* 18, 2931–2939. doi: 10.1021/acs.jproteome.9b00250

Srinivas, U. S., Tan, B. W., Vellayappan, B. A., and Jeyasekharan, A. D. (2019). ROS and the DNA damage response in cancer. *Redox Biol.* 25:101084. doi: 10.1016/j.redox.2018.101084

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947.

Umino, K., Fujiwara, S.-I., Ikeda, T., Toda, Y., Ito, S., Mashima, K., et al. (2018). Clinical outcomes of myeloid/lymphoid neoplasms with fibroblast growth factor receptor-1 (FGFR1) rearrangement. *Hematology* 23, 470–477. doi: 10.1080/10245332.2018.1446279

Van Hoeck, A., Tjoonk, N. H., Van Boxtel, R., and Cuppen, E. (2019). Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* 19:457. doi: 10.1186/s12885-019-5677-2

Vaninetti, N. M., Geldenhuys, L., Porter, G. A., Risch, H., Hainaut, P., Guernsey, D. L., et al. (2008). Inducible nitric oxide synthase, nitrotyrosine and p53 mutations in the molecular pathogenesis of Barrett's esophagus and esophageal adenocarcinoma. *Mol. Carcinog.* 47, 275–285. doi: 10.1002/mc.20382

Veldore, V. H., Patil, S., Satheesh, C. T., Shashidhara, H. P., Tejaswi, R., Prabhudesai, S. A., et al. (2015). Genomic profiling in a homogeneous molecular subtype of non-small cell lung cancer: an effort to explore new drug targets. *Indian J. Cancer* 52, 243–248. doi: 10.4103/0019-509x.175843

Wang, X., Zhang, K., Liu, X., Liu, B., and Wang, Z. (2015). Association between XRCC1 and XRCC3 gene polymorphisms and risk of thyroid cancer. *Int. J. Clin. Exp. Pathol.* 8, 3160–3167.

Witten, I. H., and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann Pub.

Wojtowicz, D., Sason, I., Huang, X., Kim, Y.-A., Leiserson, M. D. M., Przytycka, T. M., et al. (2019). Hidden Markov models lead to higher resolution maps of mutation signature activity in cancer. *Genome Med.* 11:49.

Yang, L., Gao, H., Wu, K., Zhang, H., Li, C., and Tang, L. (2020). Identification of cancerlectins by using cascade linear discriminant analysis and optimal g-gap tripeptide composition. *Curr. Bioinform.* 15, 528–537. doi: 10.2174/1574893614666190730103156

Yang, X.-F., Zhou, Y.-K., Zhang, L., Gao, Y., and Du, P.-F. (2020). Predicting LncRNA subcellular localization using unbalanced pseudo-k nucleotide compositions. *Curr. Bioinform.* 15, 554–562. doi: 10.2174/1574893614666190902151038

Yi, Q., Dong, F., Lin, L., Liu, Q., Chen, S., Gao, F., et al. (2016). PRSS1 mutations and the proteinase/antiproteinase imbalance in the pathogenesis of pancreatic cancer. *Tumour Biol.* 37, 5805–5810. doi: 10.1007/s13277-015-3982-1

Zerp, S. F., Elsas, A. V., Peltenburg, L. T. C., and Schrier, P. I. (1999). p53 mutations in human cutaneous melanoma correlate with sun exposure but are not always involved in melanomagenesis. *Br. J. Cancer* 79, 921–926. doi: 10.1038/sj.bjc.6690147

Zhang, B., and Peng, Z. Y. (2002). Structural consequences of tumor-derived mutations in p16INK4a probed by limited proteolysis. *Biochemistry* 41, 6293–6302. doi: 10.1021/bi0117100

Zhang, L., He, Y., Song, H., Wang, X., Lu, N., Sun, L., et al. (2020). Elastic net regularized softmax regression methods for multi-subtype classification in cancer. *Curr. Bioinform.* 15, 212–224. doi: 10.2174/1574893613666181112141724

Zhang, X., Chen, L., Guo, Z.-H., and Liang, H. (2019). Identification of human membrane protein types by incorporating network embedding methods. *IEEE Access* 7, 140794–140805. doi: 10.1109/access.2019.2944177

Zhang, Y.-H., Li, H., Zeng, T., Chen, L., Li, Z., Huang, T., et al. (2021a). Identifying transcriptomic signatures and rules for SARS-CoV-2 infection. *Front. Cell Dev. Biol.* 8:627302. doi: 10.3389/fcell.2020.627302

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021b). Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles. *Front. Genet.* 11:599970. doi: 10.3389/fgene.2020.599970

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021c). Determining protein–protein functional associations by functional rules based on gene ontology and KEGG pathway. *Biochim. Biophys. Acta BBA Proteins Proteom.* 1869:140621. doi: 10.1016/j.bbapap.2021.140621

Zhao, R., Chen, L., Zhou, B., Guo, Z.-H., Wang, S., and Aorigele. (2019). Recognizing novel tumor suppressor genes using a network machine learning strategy. *IEEE Access* 7, 155002–155013. doi: 10.1109/access.2019.2949415

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144. doi: 10.1016/j.mbs.2018.09.010

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinform.* 14, 709–720. doi: 10.2174/1574893614666190220114644

Zhou, J.-P., Chen, L., and Guo, Z.-H. (2020). iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. *Bioinformatics* 36, 1391–1396.

Zhu, Y., Hu, B., Chen, L., and Dai, Q. (2021). iMPTCE-Hnetwork: a multi-label classifier for identifying metabolic pathway types of chemicals and enzymes with a heterogeneous network. *Comput. Math. Methods Med.* 2021:6683051.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chen, Zhou, Zeng, Pan, Zhang, Huang, Fang and Cai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

METHODS
published: 08 September 2021
doi: 10.3389/fcell.2021.735687

Heterogeneous Information Network-Based Patient Similarity Search

Hao-zhe Huang[1], *Xu-dong Lu*[1]*, *Wei Guo*[1,2], *Xin-bo Jiang*[1,3], *Zhong-min Yan*[1] and *Shi-peng Wang*[1]

[1] School of Software, Shandong University, Jinan, China, [2] Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan, China, [3] Shandong Provincial Key Laboratory of Software Engineering, Shandong University, Jinan, China

Patient similarity search is a fundamental and important task in artificial intelligence-assisted medicine service, which is beneficial to medical diagnosis, such as making accurate predictions for similar diseases and recommending personalized treatment plans. Existing patient similarity search methods retrieve medical events associated with patients from Electronic Health Record (EHR) data and map them to vectors. The similarity between patients is expressed by calculating the similarity or dissimilarity between the corresponding vectors of medical events, thereby completing the patient similarity measurement. However, the obtained vectors tend to be high dimensional and sparse, which makes it hard to calculate patient similarity accurately. In addition, most of existing methods cannot capture the time information in the EHR, which is not conducive to analyzing the influence of time factors on patient similarity search. To solve these problems, we propose a patient similarity search method based on a heterogeneous information network. On the one hand, the proposed method uses a heterogeneous information network to connect patients, diseases, and drugs, which solves the problem of vector representation of mixed information related to patients, diseases, and drugs. Meanwhile, our method measures the similarity between patients by calculating the similarity between nodes in the heterogeneous information network. In this way, the challenges caused by high-dimensional and sparse vectors can be addressed. On the other hand, the proposed method solves the problem of inaccurate patient similarity search caused by the lack of use of time information in the patient similarity measurement process by encoding time information into an annotated heterogeneous information network. Experiments show that our method is better than the compared baseline methods.

Keywords: heterogeneous information network, clinical similarity, electronic health records, patient similarity search, weighted meta path

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Hongbo Sun,
Yantai University, China
Di Wang,
Nanyang Technological University, Singapore

***Correspondence:**
Xu-dong Lu
dongxul@sdu.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *03 July 2021*
Accepted: *30 July 2021*
Published: *08 September 2021*

Citation:
Huang H-z, Lu X-d, Guo W, Jiang X-b, Yan Z-m and Wang S-p (2021) Heterogeneous Information Network-Based Patient Similarity Search. Front. Cell Dev. Biol. 9:735687. doi: 10.3389/fcell.2021.735687

1. INTRODUCTION

Patient similarity search has been identified as one of the key techniques in artificial intelligence (AI) medicine service, which is beneficial to medical diagnosis, such as making accurate predictions for similar diseases and recommending personalized treatment plans (Sharafoddini et al., 2017). Generally speaking, patient similarity analysis involves selecting certain clinical records as features of patients in a

specific medical environment, then quantitatively analyzing the distance between them. A proper similarity measure should support various downstream applications, such as personalized medicine recommendation (Zhang et al., 2014; Lee et al., 2015), target patient retrieval (Sun et al., 2012), medical diagnoses (Gottlieb et al., 2013), and cohort study (Che et al., 2017).

The wide availability of Electronic Health Records (EHRs) makes it possible to quickly and accurately calculate the similarity between patients. Many similarity learning methods have been proposed (Tsevas and Iakovidis, 2011; Wang et al., 2012b; Barkhordari and Niamanesh, 2015; Wang and Sun, 2015; Sha et al., 2016; Zhan et al., 2016; Sharafoddini et al., 2017; Huai et al., 2018; Suo et al., 2018) on healthcare datasets. Existing methods have successfully derived the similarity measure from EHRs data through mapping the medical events into vector spaces. However, EHRs contain a variety of data (diagnostics, drugs, etc.) and a large number of medical events, which usually results in high-dimensional embedding vectors.

Heterogeneous information network (HIN) contains rich structure and semantic information, and it can effectively solve the problem caused by the high-dimensional and sparse embedding vectors. For calculating the similarity of patients, the diseases and drugs used by patients provide essential information. The patient's disease is critical to the doctor's clinical decision. At the same time, the patient's disease is basically determined by the patient's clinical symptoms and clinical indicators. It can be said that the disease is a comprehensive reflection of clinical indicators. The medicine is the solution made by the doctor to the patient's disease and symptoms, and is the final manifestation of the doctor's clinical decision. Therefore, it is easy to think that patients, diseases, and drugs can be connected to form HIN.

However, there are many duplicate diseases and drugs in the EHRs, meaning that if we were to use classic HIN modeling techniques with the above schema, we would lose the correlation information between patients and drugs. Considering this problem, we propose a kind of HIN with annotation: that is, in links connecting diseases and drugs, we add an annotation of patient information to enrich the original network with the information between patients and drugs. We call it annotated HIN. On the annotated HIN, we propose a novel node similarity measure S-PathSim to calculate patient similarity. As a node similarity measure, S-PathSim enjoys some good properties, like symmetric and self-maximum.

On the other hand, temporal information is crucial to understand the dynamics of medical expressions. To leverage the essential temporal information for patient similarity evaluation, we propose to use N-disease to encode temporal information into annotated HINs. N-disease is inspired by the N-grams model in natural language processing. Its basic idea is to arrange the patients' diseases into time series according to the time they are developed, sequentially collect the N-grams from the disease sequences, and then replace the disease object with the disease N-grams in the annotated HIN. The collected N-grams from the disease time series are called N-diseases.

Finally, two patient similarity search methods, MBH (method based on annotated HIN) and MBHT (method based on annotated HIN and temporal information), were defined according to S-PathSim and N-disease.

The remainder of this paper is structured as follows. The second section reviews the related research work on the topic of patient similarity analysis and heterogeneous information network, while the third section provides some preliminaries on HIN and shows the limitation of HIN to the calculation of patient similarity. In the fourth section, we introduced our method in detail. The experimental results and comparative analysis are shown in section five. Finally, the last section summarizes this paper and discusses some possible avenues for future research.

2. RELATED WORK

In this section, we review some related works on evaluating patient similarity and heterogeneous information network.

Studying patient similarity has practical significance in many applications (Lee et al., 2015; Li et al., 2015). Ng et al. provided personalized predictive healthcare model by matching clinical similar patients with a locally supervised metric learning measure (Ng et al., 2015). An integrated method for personalized modeling (IMPM) was proposed to provide personalized treatment and personalized drug design (Kasabov and Hu, 2010). The data-driven clinical decision support system was combined with patient similarity (Xia et al., 2019).

At present, there are many studies to calculate the similarity of patients. Zhang et al. combined patient similarity and drug similarity analysis and proposed a heterogeneous label propagation method to identify which drug is likely to be effective for a given patient (Zhang et al., 2014). Chan et al. proposed a patient similarity algorithm named SimSvm that uses support vector machine to weight the similarity measures (Chan et al., 2010). Wang et al. proposed a patient similarity based disease prognosis strategy named SimProX (Wang et al., 2012a). This model used a local spline regression based method to embed these patient events into an intrinsic space, and then measure the patient similarity by the Euclidean distance in an embedded space. However, these methods do not leverage temporal information to evaluate patient similarities, which prevents them from delivering.

Cheng et al. (2016) took temporal information into consideration and proposed an adjustable temporal fusion scheme using CNN-extracted features. This method is a supervised model, but the label data are not easy to obtain, which limits its use, and the method lacks interpretability. Zhu et al. proposed the method to solve the problem of high-dimensional vectors and time series (Zhu et al., 2016). They embed medical events from HER into fixed-length vectors, but fixed-length vectors are difficult to obtain complete medical event information.

As mentioned above, the current method of measuring patient similarity is limited, and a better method is needed to calculate patient similarity.

Since, Sun et al. proposed the concept of HIN (Sun and Han, 2010), and the meta path concept subsequently (Sun and Han, 2011), HIN analysis becomes a hot topic rapidly in the

fields of data mining, database, and information retrieval. He et al. incorporated temporal information for similarity search in HINs by assigning different weights to the paths built at different time (He et al., 2014). But this method is not suitable for the annotated HIN proposed in this paper. In order to evaluate the relevance of different-typed objects, Shi et al. (2014) proposed HeteSim to measure the relevance of any object pairs under arbitrary meta paths. As an adaption of HeteSim, LSH-HeteSim (Li et al., 2014) is proposed to mine the drug–target interaction in heterogeneous biological networks where drugs and targets are connected with complicated semantic paths. In order to overcome the shortcoming of HeteSim in high computation and memory demand, Meng et al. (2014) proposed the AvgSim measure that evaluates similarity score through two random walk processes along the given meta path and the reversed meta path, respectively. In order to overcome the problem that the meta path can only express simple information, Cheng et al. (2017) proposed meta structure to measure the similarity between the objects. Until today, HINs have been widely used in other fields (Wang, 2019; Wang et al., 2020; Zhang et al., 2020).

HIN rarely results in high-dimensional vectors, and most similarity calculation methods based on HIN have good interpretability. But it cannot be perfectly applied to patient similarity calculation, so in this paper,we propose an improved method, annotated HIN, which can be well-applied to calculate the similarity of patients.

3. PRELIMINARIES

In this section, as preliminaries, we will detail the HIN and its limitation in measures patient similarity.

3.1. HIN

An information network is defined as a directed graph $G = (V, E)$ with an object type mapping function $\psi : V \rightarrow A$ and a link type mapping function $\varphi : E \rightarrow R$, in which each object $v \in V$ belongs to a particular object type $\psi(v) \in A$ while each link $e \in E$ belongs to a particular relation $\varphi(e) \in R$.

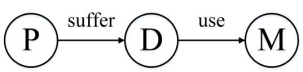


FIGURE 1 | Network schema of the patient heterogeneous information network.

Different from the traditional network definition, we explicitly distinguish the object types and relationship types in these networks. When the types of objects $|A| > 1$ or the types of relations $|R| > 1$, the network is referred to as a heterogeneous information network; otherwise, it is a homogeneous information network.

3.2. Limitation of HIN

HIN can link patients, diseases, and drugs. As shown in **Figure 1**, we can get the network schema of the patient HIN. P, D, and M represent patient, disease, and medicine, respectively.

There may be many kinds of drugs to treat one disease, and one drug can also cure many diseases, which leads to some incorrect information in the traditional heterogeneous information network when connecting patients, diseases, and drugs. We use a specific example below to illustrate this problem.

Table 1 presents three inpatient records for two patients, all of which were diagnosed with the same disease; patient 231 was hospitalized twice. From the data in **Table 1**, the HIN in **Figure 2** is obtained. However, the HIN shown in **Figure 2** has two problems. First of all, we need to know that patient 231 has been hospitalized twice, but this information cannot be obtained through **Figure 2**. Second, patient 231 does not use perindopril in treatment, but the information we get from the heterogeneous information network is that there is a relation between patient 231 and perindopril, which leads to the incorporation of misleading information. Therefore, traditional HIN-based measurement methods are not suitable for our problem.

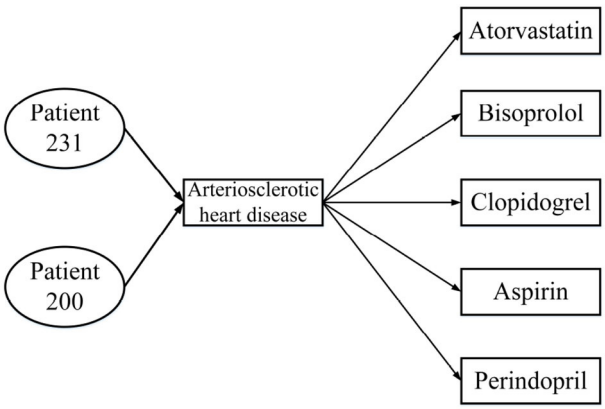


FIGURE 2 | Heterogeneous information network from **Table 1**.

TABLE 1 | Example of case information.

Hospital ID	Patient ID	Disease	Medicine
564435	231	Arteriosclerotic heart disease	Atorvastatin, Bisoprolol, Clopidogrel
561657	200	Arteriosclerotic heart disease	Aspirin, Atorvastatin, Perindopril
564677	231	Arteriosclerotic heart disease	Atorvastatin, Clopidogrel

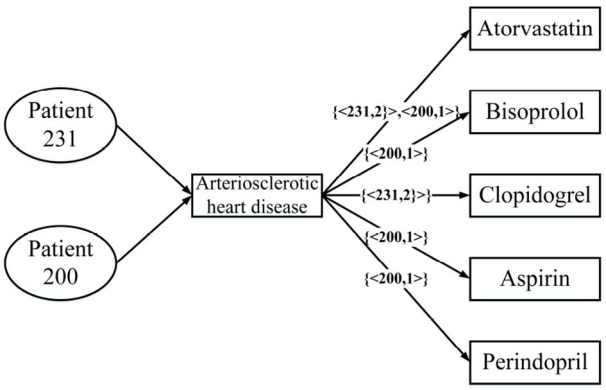


FIGURE 3 | An annotated HIN from **Table 1**.

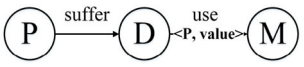


FIGURE 4 | Schema of patient-annotated HIN.

4. THE PROPOSED METHOD

4.1. Annotated HIN

As mentioned in section 3, HIN is not suitable for our problem. In order to measure patient similarity, we propose a new graph model-annotated HIN.

Definition 1. ***Annotated Heterogeneous Information Network.*** Annotated HIN is a special heterogeneous information network $G = (V, E, C)$. In the annotated HIN, there is a set of one or more link types annotated by $< key, value >$ pairs. For each $< key, value >$ pair, key corresponds to a specific type of object $\psi(key) \in V$, while value is used to record the number of links.

As above mentioned, we regard the set of $< key, value >$ pairs as the annotations of a heterogeneous information network, represented by C. The number of key-value pairs in the set is referred to the length of the annotation, which is represented by L. Annotations can be added to one or more link types of the classic heterogeneous information network. These annotations can be used to record the source and number of connections and can thus represent more information.

Figure 3 is a real example diagram of an annotated HIN, and we named it patient-annotated HIN. It can be seen that the connection with "Clopidogrel" has annotation $C_{Clopidogrel} = \{< 231, 2 >\}$, and that the annotation length is $L = 1$. Combined with the annotated heterogeneous information network, we can interpret it as follows: Patient 231 was diagnosed with atherosclerotic heart disease in both hospitalizations, and clopidogrel was used in both treatments. Moreover, there is no corresponding record of patient 200 in the note, so it can be concluded that clopidogrel was not used in the treatment of patient 200. In this way, the two problems described in the previous section are solved.

For a given annotated HIN, in order to help readers better understand the object type, link type, and annotation type in the network, we provide its meta-description.

Definition 2. ***AHIN Network Schema.*** The network schema of AHIN is recorded as $SG = (A, R, I)$. This is a meta template of AHIN $G = (V, E, C)$. It has object type mapping $\psi(v) \in A$, relation type mapping $\varphi(e) \in R$, and annotation type mapping $\theta : C \to I$. It is defined on object type set A, relation type set R, and annotation type set I.

4.2. Weighted Meta Path and S-PathSim

The weighted meta path, designed to capture complex relationship between two annotated HIN objects, is based on network expansion structure. And the network expansion structure is defined as follows.

Definition 3. ***Network Expansion Structure.*** Network expansion structure S is a set of directed weighted graphs, which is defined on an annotated HIN schema $SG = (A, R, I)$. It expands the annotated heterogeneous information network into an easy-to-process format. Formally, $S = (D_1, D_2, \ldots, D_n)$, where $D_n = (V_n, E_n)$ is a directed weighted graph with D_n, V_n being the set of nodes and edges, respectively. For any edge $e \in E_n$, a weight $w(e)$ is associated, with the default value 1.

Below we use an example to introduce the expansion of the network structure. **Figure 5A** demonstrates the expansion from a given annotated heterogeneous information network into the network expansion structure. There are annotations $\{< P_1, 2 >, < P_2, 3 >\}$, and $\{< P_1, 3 >\}$ in graph G. The key-value pairs $< P_1, 2 >$ and $< P_1, 3 >$ correspond to the entity P_1, so we can get the graph D_1, and the corresponding edge weights are 2, 3, respectively. And the key-value pair $< P_2, 3 >$ corresponds to the entity P_2, so we get the graph D_2, and the corresponding edge weight is 3. For the other edges, our default weight is 1.

After introducing the network expansion structure, we propose the concept of weighted meta path.

Definition 4. ***Weighted Meta Path.*** Weighted meta path P is a path defined on the network schema $SG = (A, R, I)$, and based on network expansion structure $S = (D_1, D_2, \ldots, D_n)$. Weighted meta path is denoted in the form of $A_1 \xrightarrow{R_1, w(e_1)} A_2 \xrightarrow{R_2, w(e_2)} \ldots \xrightarrow{R_l, w(e_l)} A_{l+1}$, which defines a composite relation between object A_1 and A_{l+1}, where R_l represents the relationship between A_1 and A_{l+1}, and $w(e_l)$ represents the weight of the relationship.

Just like the meta path, if the relationship of the weighted meta path P is symmetric, then we say that it is symmetric. For a specified weighted meta path, it has a specified template. If there is no multiple relationship between the same object types, we can use the type name to represent the template of the weighted meta path: $P = (A_1 A_2 \ldots A_{l+1})$. As shown in **Figure 5B**, P_1 and P_2 have the same template $PDMDP$. P_1 and P_2 are symmetric weighted meta paths.

When $A_{l+1} = A'_1$, the weighted meta paths $P = (A_1 A_2 \ldots A_{l+1})$ and $P' = (A'_1 A'_2 \ldots A'_{l+1})$ are concatenable,

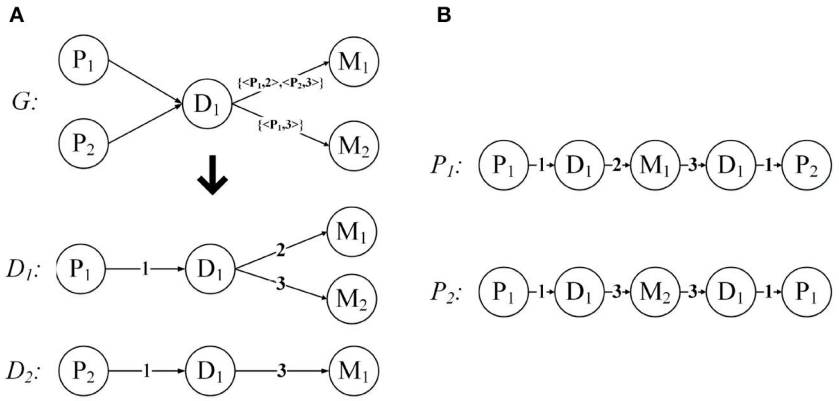


FIGURE 5 | Network expansion structure and weighted meta path. **(A)** Network expansion structure. **(B)** Weighted meta path.

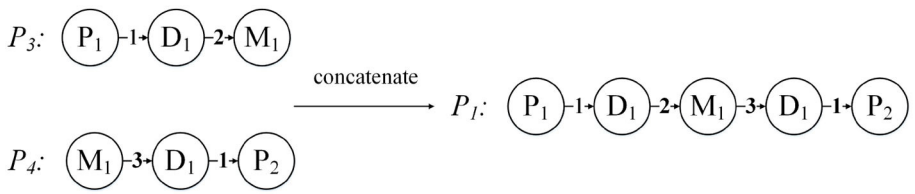


FIGURE 6 | Concatenation of weighted meta path.

so that a new weighted meta path $(A_1A_2 \ldots A_{l+1}A'_2 \ldots A'_{l+1})$ is obtained.

For each weighted meta path P, there is a score $S(P)$, and $S(P)$ is the product of the weights of the relationships in P. For example, the weighted meta path P_1, $S(P_1) = 1 * 2 * 3 * 1 = 6$. In fact, $S(P)$ represents the weight of the relationship between the first and last objects in the weighted meta path P, and can also be understood as the number of connection paths between the two objects. As shown in **Figure 6**, the weighted meta path P_3, $W(\{D_1, M_1\}_{P3}) = 2$, represents that patient P_1 has used the drug M_1 twice because of disease D_1. Therefore, $S(P_3) = W(\{P_1, D_1\}_{P_3}) * W(\{D_1, M_1\}_{P_3}) = 2$ can also be obtained, then the number of connection paths between P_1 and M_1 is 2. In the same way, $S(P_4) = W(\{M_1, D_1\}_{P_4}) * W(\{D_1, P_2\}_{P_4}) = 3$, then the number of connection paths between patient P_2 and drug M_1 is 3. P_1 can be obtained by concatenating P_3 and P_4, then we can get that the number of connection paths between patient P_1 and patient P_2 is $S(P_1) = S(P_3) * S(P_4) = 6$.

Based on the annotated HIN and weighted meta path, we propose a new measure, named S-PathSim.

Definition 5. ***S-PathSim.*** Given a symmetric weighted meta path, S-PathSim between two objects of the same type x and y is:

$$s(x, y) = \frac{2 \times S_{sum}(P_{x \to y})}{S_{sum}(P_{x \to x}) + S_{sum}(P_{y \to y})}, \quad (1)$$

where $S_{sum}(P_{x \to y})$ is the sum of score of the weighted meta path between x and y, $S_{sum}(P_{x \to x})$ is that between x and x, and $S_{sum}(P_{y \to y})$ is that between y and y. If there are two weighted meta-paths P_a and P_b between x and y, and $S(P_a) = 4$, $S(P_b) = 3$, then $S_{sum}(P_{x \to y}) = S(P_a) + S(P_b) = 7$.

Take the patients in **Table 1** as an example, and patient 231 has two admissions. During his first hospitalization, he developed arteriosclerotic heart disease and had some medicine including atorvastatin, bisoprolol, and clopidogrel. Patient 200 also developed arteriosclerotic heart disease and he had the medicine aspirin, atorvastatin, and perindopril. According to these information, we can get an heterogeneous information network G as shown in **Figure 5A**. According to Definition 5, we can get $S_sum(patient231 \to patient200) = 6$, $S_{sum}(patient231 \to patient231) = 22$, $S_{sum}(patient200 \to patient200) = 9$, therefore $s(patient231, patient200) = 6/11$.

As mentioned before, $S(P)$ can be understood as the number of connecting paths of the first and last two objects in the weighted meta path P. If there are more connection paths between two objects, then we can consider them to have a higher similarity. However, the result obtained by using the number of paths as the judgment condition will be biased toward high-visibility objects. Therefore, we use the number of connection paths from two objects to their own as a balance factor. This idea has been applied to PathSim, and we extend it to the annotated HIN here, and propose S-PathSim.

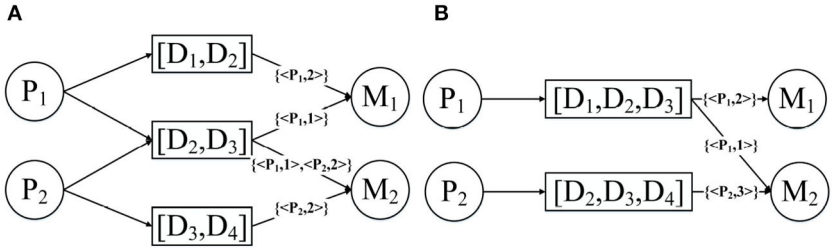


FIGURE 7 | Annotated HIN after **(A)** 2-disease and **(B)** 3-disease operations.

Properties of S-PathSim:

- (1) Symmetric: $s(x, y) = s(y, x)$. Considering the semantics of $S_{sum}(P_{x\to y})$, it is easy to understand $S_{sum}(P_{x\to y}) = S_{sum}(P_{y\to x})$, so $s(x, y) = s(y, x)$.
- (2) Self-maximum: $s(x, y) \in [0, 1]$, and $s(x, x) = 1$. The weighted meta path template mn and nm can be concatenated into a new weighted meta path $mnm.mnm_i$ is the ith path of the weighted meta-path template mnm, as mentioned before, $S(mnm_i) = S(mn_i) * S(nm_i)$. Assuming that mn is the weighted meta path template, the kth weighted element path is expressed as a_k, and nm is the weighted meta path template, and the kth weighted element path is expressed as b_k, then $S_{sum}(P_{x\to y}) = \sum_{k=1}^{p} S(a_k) * S(b_k)$; the same can be obtained as $S_{sum}(P_{x\to x}) = \sum_{k=1}^{q} S(a_k)^2$, $S_{sum}(P_{y\to y}) = \sum_{k=1}^{o} S(b_k)^2$. There must be $p \le q, p \le o$. Then $2\sum_{k=1}^{p} S(a_k) * S(b_k) \le \sum_{k=1}^{p} S(a_k)^2 + \sum_{k=1}^{p} S(b_k)^2 \le \sum_{k=1}^{q} S(a_k)^2 + S(b_k)^2$, so $S(x, y) \le 1$. And it is easy to understand that $s(x, y) \ge 0$, so $s(x, y) \in [0, 1]$, $s(x, x) = 1$. In the above formula, p represents the number of weighted meta path between x and y, q represents the number of weighted meta path between x and x, and o represents the number of weighted meta path between y and y.

4.3. Temporal Information Encoding

Temporal information is critical to understanding the patients' dynamics. However, the AHIN described previously cannot capture the temporal information, so for the problem to be solved in this article, we propose an N-disease method to embed temporal information into the AHIN.

N-disease is inspired by the natural language processing model N-grams. Its basic idea is to arrange the patients' diseases set into time series according to the time when they were developed, sequentially collect the N-grams from the disease sequences, and then replace the disease object with the disease N-grams in the annotated HIN. Assuming that P_1 has the diseases $[D_1, D_2, D_3]$ and P_2 has the disease $[D_2, D_3, D_4]$, then the results obtained after the 2-disease operation and the 3-disease operation are shown in **Figure 7**. In fact, the patient annotation HIN given in **Figure 4** is essentially the patient annotation HIN after 1-disease operation.

It should be noted that as N becomes larger and larger, the accuracy of the patient's annotation of diseases and drug connections in the HIN will gradually decrease. As shown in **Figure 7A**, the node $[D_1, D_2]$ is connected to the drug; then you do not know whether this drug is used to treat disease D_1 or disease D_2. Fortunately, we can trade off the accuracy and temporal information by changing N.

4.4. MBH and MBHT

Retrieving top-k similar patients of specified patients has practical significance. It allows doctors to analyze similar patients to provide better treatment options. Previously, we have introduced the annotated HIN–based measurement method S-PathSim and temporal information embedding method N-disease. In this section, we define two patient similarity search methods, MBH and MBHT, according to the definition introduced earlier.

MBH is a method based on annotated HIN. In detail, first, annotated HIN is constructed using the patient's medical record information. After specifying a patient, S-PathSim is used to calculate the patient similarity and return the top-k similar patient.

MBHT is a method based on annotating HIN and temporal information. The difference between MBHT and MBH is that MBHT needs to construct the annotated HIN processed by the N-disease based on patient's medical record information, and embed the temporal information into the annotated HIN, then use S-PathSim to calculate the patient similarity and return the top-k similar patient.

It is easy to understand that MBHT is the combination of N-disease and MBH. When $N = 1$, MBHT is MBH. MBHT uses the temporal information in the patient's medical records, but it also loses some accuracy, and we need to make a trade-off between timing and accuracy.

5. SIMULATION EXPERIMENTS AND RESULTS ANALYSIS

5.1. Data Description

We perform experiments on a real dataset, which primarily includes information about the medical treatments and drug details of each person. Each person has multiple records ($n > 2$). Moreover, each record contains a diagnosis (i.e., ICD10) and information about multiple drugs. To improve the experiment quality, we randomly divided the data into four sub-datasets. **Table 2** shows the description of the divided datasets. In addition,

TABLE 2 | Description of datasets.

Sub-dataset	Dataset A	Dataset B	Dataset C	Dataset D	Total
Number of patient	13,461	13,461	13,461	13,460	53,853
Disease types	946	953	946	943	1,928
Drug types	1,400	1,412	1,403	1,390	2,217
Number of diagnoses per capita	6.653	6.540	6.647	6.456	6.620

we did not perform any other desensitization treatment (such as removing diseases with less than five patients), so our experiment is performed on a real-world dataset without any unjustifiable data manipulations.

5.2. Experimental Settings

In application, comparative analysis is often performed by retrieving top-k similar patients of designated patients to support clinical decision making. In the experiment, we also evaluate the model by retrieving the top-k similar patients of the specified patients. We set $k = 10$. We used two metrics for quantitative evaluation.

nDCG (normalized Discounted Cumulative Gain, with the value between 0 and 1, the higher the better) Zhang et al. (2020) is an indicator used to measure the quality of the ranking. The main idea is that the products that the user likes are supposed to be ranked in front of the recommendation list rather than in the back so as to significantly increase the user experience. It is obtained by *DCG* (Discounted Cumulative Gain) normalization, where *rel* is a sorted list, *i* is the position number of the current result, and *IDCG* is the largest *DCG* in the ideal state.

$$DCG = \sum_{i}^{p} \frac{rel_i}{log_2 i + 1} \tag{2}$$

$$nDCG = \frac{DCG}{IDCG} \tag{3}$$

The *HL* (half-life utility) (Sarwar et al., 2001) index is proposed under the assumption that the probability that the user browses the product and the specific ranking value of the product in the recommendation list decrease exponentially. It measures the practicality of the recommendation system for a user. It is the difference between the user's actual rating and the model rating. So *HL* can also be used to evaluate top-k search results.

$$HL_u = \sum_{\alpha} \frac{max(r_{ua} - d, 0)}{2^{(l_{ua}-1)/(h-1)}} \tag{4}$$

Among them, r_{ua} represents the true similarity of patient u and patient a, d is the default score, in the experiment we set d to the average similarity, and l_{ua} is the ranking of patient a in the recommended list of patient u. h is the half-life of the system, that is, there is a 50% probability that the user will browse the recommended list position, we set $h = 3$.

In order to verify the effectiveness of the proposed MBH based on S-PathSim, we set up a comparison experiment between MBH and the similarity search method based on PathSim. In addition, in order to explore the effect of N-disease on the results, N was set to 1, 2, 3, 4, respectively, and count the results of MBHT for comparative analysis. Finally, we explored the effect of N-disease on algorithm efficiency. The experimental environment is as follows: INTELCorei5 CPU, 2.80 GHz; 4G memory.

5.3. Comparison of Patient Similarity Search Method

This article proposes annotated HIN and S-PathSim, and defines MBH, a patient similarity search method based on the annotated HIN and S-PathSim. PathSim is an excellent object similarity measurement method based on HIN. PathSim can be used to retrieve the similarity of patients. Here, we compare MBH with PathSim-based methods to verify the effectiveness of MBH:

(1) MBH: Map the patient information to the annotated HIN, the schema is shown in **Figure 4**, through the weighted meta path as shown in **Figure 5B**; the S-PathSim is used to measure the similarity of patients, and get the top-k similar patients of the specified patients.
(2) Baseline: Map patient information into HIN. The schema is shown in **Figure 1**. The meta path used is (*PDMDP*). PathSim is used to calculate the patient similarity, and the top-k search result of the specified patient is obtained.

It is worth mentioning that the above steps are run simultaneously in 4 sets of datasets, effectively avoiding accidental.

Figure 8 shows the experimental results of the two models on 4 sets of datasets. **Figure 8A** uses *nDCG* as the evaluation criterion, and it can be observed that MBH is superior to baseline on four datasets. **Figure 8B** uses *HL* as the evaluation criterion, which proves that MBH has better practicability than baseline.

5.4. The Impact of N-Disease

We propose N-disease to embed temporal information into annotated HIN, and the difference between MBH and MBHT is whether N-disease is used or not. In this section, we explore the comparison results of MBH and MBHT, and the effect of N-disease on MBHT. We set N to 1, 2, 3, and 4, respectively. When $N = 1$, the annotated HIN does not contain temporal information, and MBHT is MBH. When $N = 4$, annotated HIN contains the largest amount of temporal information. However, after a threshold, with the increase of N, the annotated HIN captures increasingly more temporal information while its patient similarity search performance decreases steadily. We

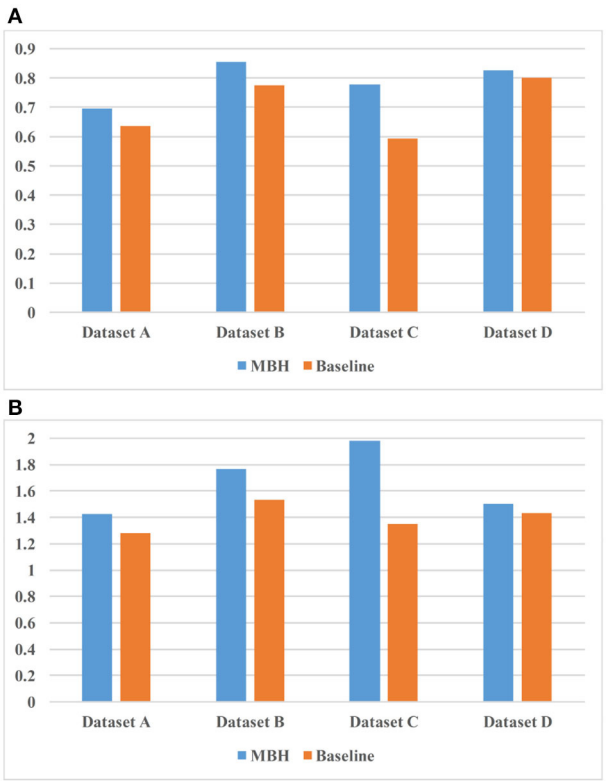


FIGURE 8 | Comparison result of MBH (method based on annotated HIN) and baseline. **(A)** uses nDCG as the evaluation criterion and **(B)** uses HL as the evaluation criterion.

TABLE 3 | The effect of N-disease measured by *nDCG* (normalized Discounted Cumulative Gain).

Dataset	$N = 1$	$N = 2$	$N = 3$	$N = 4$
Dataset A	0.696	**0.704**	0.695	0.694
Dataset B	0.855	**0.859**	0.807	0.761
Dataset C	0.778	0.767	**0.800**	0.791
Dataset D	0.825	**0.837**	0.799	0.787
Mean	0.788	**0.791**	0.775	0.758

Note: The bold values are the best results.

should carefully choose the threshold for N to obtain the best results.

The experimental results are shown in **Table 3**. In datasets A, C, and D, MBHT has the best results when $N = 2$; in dataset B, MBHT achieved the best results when $N = 3$. Among the average values of the 4 datasets, $N = 2$ makes MBHT achieve the best results. In general, $N = 2$ can achieve the best results of MBHT, and $N = 2$ can balance the time-consuming and accuracy of annotated HIN. At the same time, the experimental results also show that MBHT is better than MBH.

In the following, we explore the effect of N-disease on MBHT efficiency. We assume that when N-disease method is not used (i.e., $N = 1$), the running time of the program is unit 1. The experimental results are as follows.

TABLE 4 | The effect of N-disease on efficiency.

Dataset	$N = 1$	$N = 2$	$N = 3$	$N = 4$
Dataset A	1	0.918	0.978	0.972
Dataset B	1	0.828	0.902	0.893
Dataset C	1	0.767	0.957	0.893
Dataset D	1	0.893	0.929	0.954
Mean	1	0.883	0.941	0.928

It can be seen from **Table 4** that the efficiency of the algorithm is improved by using N-disease; especially when $N = 2$, the algorithm has the highest efficiency. The use of N-disease changes the number of annotated HIN nodes and the relationship between the nodes, which in turn changes the efficiency of the algorithm. Since N-disease will affect the efficiency of MBHT, this paper gives an explanation from a practical point of view.

When the program is implemented, we divide MBHT into two steps. The first step is data statistics, and the second step is S-PathSim calculation. The use of MBHT has more data statistics steps than the use of MBH alone, but we know from practice that the time consumed by the data statistics step is quite small and can even be ignored. When we calculate S-PathSim, we use a lot of multiplication, which takes most of the total running time. We found that when $N = 2$, the number of multiplication operations is significantly smaller than when MBH is used alone. This explains why the running time of the program when $N = 2$ is shorter than that when using MBH alone.

In short, we conclude that when N =2, annotated HIN achieves a balance between time consuming and accuracy, and can effectively improve the efficiency of the algorithm.

6. CONCLUSION

In this paper, a new method of patient similarity calculation is proposed that uses the disease and drug data of patients, and further uses the annotated HIN proposed in this paper to create a model. The heterogeneous network adds the annotation of patient information to the connecting links between diseases and drugs, which solves the problem of the classic HIN in losing the information regarding these associations. At the same time, based on the annotated HIN, we propose S-PathSim to measure patient similarity. Furthermore, N-disease is proposed to encode temporal information into the annotated HIN. Our measurement does not rely on high-dimensional and sparse vectors, and effectively captures the patient's medical events and the temporal information in EHRs. Finally, based on S-PathSim and N-disease, two patient similarity search methods, MBH and MBHT, are proposed. The experimental results show that the method proposed in this paper is superior to competitive baseline method.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

H-zH designed and conducted the experiments and finished this paper writing. X-dL determined the technical route and methods. WG devised the assessment method for the methods. X-bJ did some research on the related work. Z-mY discussed the methods and ideas. S-pW discussed the methods and ideas and submitted the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key Research and Development Project of China (No. 2019YFB1705904), National Nature Science Foundation of China (Nos.91846205, 61772316, 61907026), Innovation Methods Work Special Project (No. 2020IM020100), Science and Technology Development Plan Project of Shandong Province (No. 2019JZZY020505), Key Research & development Project of Shandong Province (No. 2019GGX101009), Shandong-Chongqing Technological Collaboration Plan (No. cstc2020jscx-lyjsAX0010), and Project of Shandong Province Higher Educational Science and Technology Program (No. J18KA392).

REFERENCES

Barkhordari, M., and Niamanesh, M. (2015). ScaDiPaSi: an effective scalable and distributable mapreduce-based method to find patient similarity on huge healthcare networks. *Big Data Res.* 2, 19–27. doi: 10.1016/j.bdr.2015.02.004

Chan, L. W.-C., Chan, T.-K., Cheng, L., and Mak, W. (2010). "Machine learning of patient similarity: a case study on predicting survival in cancer patient after locoregional chemotherapy," in *IEEE International Conference on Bioinformatics and Biomedicine Workshops* (HongKong), 467–470. doi: 10.1109/BIBMW.2010.5703846

Che, C., Xiao, C., Liang, J., Jin, B., Zho, J., and Wang, F. (2017). "An RNN architecture with dynamic temporal matching for personalized predictions of Parkinson's disease," in *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)* (Houston, TX), 198–206. doi: 10.1137/1.9781611974973.23

Cheng, R., Huang, Z., Zheng, Y., Yan, J., Wong, K. Y., and Ng, E. (2017). "Meta structure: computing relevance in large heterogeneous information networks," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data* (Beijing), 3–7.

Cheng, Y., Wang, F., Zhang, P., and Hu, J. (2016). "Risk prediction with electronic healthrecords: a deep learning approach," in *Proceedings of the 2016 SIAM International Conference on Data Mining* (Miami, FL), 432–440. doi: 10.1137/1.9781611974348.49

Gottlieb, A., Stein, G. Y., Ruppin, E., Altman, R. B., and Sharan, R. (2013). A method for inferring medical diagnoses from patient similarities. *BMC Med.* 11:194. doi: 10.1186/1741-7015-11-194

He, J., Bailey, J., and Zhang, R. (2014). "Exploiting transitive similarity andtemporal dynamics for similarity search in heterogeneous information net-works," in *International Conference on Database Systems for Advanced Applications* (Bali), 141–155. doi: 10.1007/978-3-319-05813-9_10

Hu, L., Gong, Y., Xing, H., and Wang, F. (2019). Semantic representation with heterogeneous information network using matrix factorization for clustering in the internet of things. *IEEE Access* 7, 31233–31242. doi: 10.1109/ACCESS.2019.2903310

Huai, M., Miao, C., Suo, Q., and Li, Y. (2018). "Uncorrelated patient similarity learning," in *Proceedings of the 2018 SIAM International Conference on Data Mining* (San Diego, CA), 270–278. doi: 10.1137/1.9781611975321.31

Kasabov, N., and Hu, Y. (2010). Integrated optimization method for personalised modelling and case studies for medical decision support. *Int. J. Funct. Inform. Pers. Med.* 3, 236–236. doi: 10.1504/IJFIPM.2010.039123

Lee, J., Maslove, D. M., and Dubin, J. A. (2015). Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLoS ONE* 10:e0127428. doi: 10.1371/journal.pone.0127428

Li, C., Sun, J., Xiong, Y., and Zheng, G. (2014). "An efficient drug-target interaction mining algorithm in heterogeneous biological networks," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Tainan; Taiwan), 65–76. doi: 10.1007/978-3-319-13186-3_7

Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., et al. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* 7:311ra174. doi: 10.1126/scitranslmed.aaa9364

Meng, X., Shi, C., Li, Y., Zhang, L., and Wu, B. (2014). "Relevance measure in large scale heteroge-neous networks," in *Asia-Pacific Web Conference* (Changsha), 636–643. doi: 10.1007/978-3-319-11116-2_61

Ng, K., Sun, J., Hu, J., and Wang, F. (2015). Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits Transl. Sci. Proc.* 2015, 132–136.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web* (Hong Kong), 285–295. doi: 10.1145/371920.372071

Sha, Y., Venugopalan, J., and Wang, M. D. (2016). "A novel temporal similarity measure for patients based on irregularly measured data in electronic health records," in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Seattle, WA), 337–344. doi: 10.1145/2975167.2975202

Sharafoddini, A., Dubin, J. A., and Lee, J. (2017). Patient similarity in prediction models based on health data: a scoping review. *JMIR Med. Inform.* 5:e7. doi: 10.2196/medinform.6730

Shi, C., Kong, X., Huang, Y., Yu, P. S., and Wu, B. (2014). HeteSim: a general framework for relevancemeasure in heterogeneous networks. *IEEE Trans. Knowledge Data Eng.* 26, 2479–2492. doi: 10.1109/TKDE.2013.2297920

Sun, J., Wang, F., Hu, J., and Edabollahi, S. (2012). Supervised patient similarity measure of heterogeneous patient records. *ACM SIGKDD Explorat. Newslett.* 14, 16–24. doi: 10.1145/2408736.2408740

Sun, Y., and Han, J. (2010). "Rankclus: integrating clustering with ranking for heterogeneous information network analysis," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (Saint Petersburg), 565–576. doi: 10.1145/1516360.1516426

Sun, Y., and Han, J. (2011). "Pathsim: meta path-based top-k similarity search in heterogeneous information networks," in *Proceedings of the VLDB Endowment* (Seattle, WA), 992–1003. doi: 10.14778/3402707.3402736

Suo, Q., Zhong, W., Ma, F., Ye, Y., Huai, M., and Zhang, A. (2018). "Multi-task sparse metric learning for monitoring patient similarity progression," in *2018 IEEE International Conference on Data Mining (ICDM)* (Singapore), 477–486. doi: 10.1109/ICDM.2018.00063

Tsevas, S., and Iakovidis, D. K. (2011). "Fusion of multimodal temporal clinical data for the retrieval of similar patient cases," in *10th International Workshop on Biomedical Engineering* (Kos), 1–4. doi: 10.1109/IWBE.2011.6079049

Wang, F., Hu, J., and Sun, J. (2012a). "Medical prognosis based on patient similarity and expert feedback," in *Proceedings of the 21st International Conference on Pattern Recognition* (Tsukuba), 1799–1802.

Wang, F., Lee, N., Hu, J., Sun, J., and Ebadollahi, S. (2012b). "Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing), 453–461. doi: 10.1145/2339530.2339605

Wang, F., and Sun, J. (2015). Psf: A unedified patient similarity evaluation framework through metric learning with weak supervision. *IEEE J. Biomed. Health Inform*. 19, 1053–1060. doi: 10.1109/JBHI.2015.2425365

Wang, R., Ma, X., Jiang, C., Ye, Y., and Zhang, Y. (2020). Heterogeneous information network-based music recommendation system in mobile networks. *Comput. Commun*. 150, 429–437. doi: 10.1016/j.comcom.2019.12.002

Xia, E., Wang, K., Zhang, Y., Yu, Y., Mei, J., and Li, S. (2019). "A data-driven clinical decision support system for acute coronary syndrome patient similarity," in *2019 IEEE International Conference on Healthcare Informatics* (Xi'an), 1–6. doi: 10.1109/ICHI.2019.8904614

Zhan, M., Cao, S., Qian, B., Chang, S., and Wei, J. (2016). "Low-rank sparse feature selection for patient similarity learning," in *2016 IEEE 16th International Conference on Data Mining (ICDM)* (Barcelona), 1335–1340. doi: 10.1109/ICDM.2016.0182

Zhang, C., Wang, G., Yu, B., Xie, Y., and Pan, K. (2020). Proximity-aware heterogeneous information network embedding. *Knowledge Based Syst*. 193:105468. doi: 10.1016/j.knosys.2019.105468

Zhang, P., Wang, F., Hu, J., and Sorrentino, R. (2014). Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA. Jt Summits. Transl. Sci. Proc*. 2014, 132–136.

Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., and Wang, F. (2016). "Measuring patient similarities via a deep architecture with medical concept embedding," in *2016 IEEE 16th Intermnational Conference on Data Mining* (Barcelona), 749–758. doi: 10.1109/ICDM.2016.0086

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Huang, Lu, Guo, Jiang, Yan and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 23 September 2021
doi: 10.3389/fcell.2021.686664

Systematic Characterization of Novel Immune Gene Signatures Predicts Prognostic Factors in Hepatocellular Carcinoma

*Dafeng Xu[1], Yu Wang[2], Jincai Wu[1], Yuliang Zhang[3], Zhehao Liu[1], Yonghai Chen[1] and Jinfang Zheng[1]**

[1] Department of Hepatobiliary and Pancreatic Surgery, Hainan General Hospital, Hainan Affiliated Hospital of Hainan Medical University, Haikou, China, [2] Geriatric Medicine Center, Hainan General Hospital, Hainan Affiliated Hospital of Hainan Medical University, Haikou, China, [3] Department of Otolaryngology-Head and Neck Surgery, Hainan General Hospital, Hainan Affiliated Hospital of Hainan Medical University, Haikou, China

Background: The prognosis of patients with hepatocellular carcinoma (HCC) is negatively affected by the lack of effective prognostic indicators. The change of tumor immune microenvironment promotes the development of HCC. This study explored new markers and predicted the prognosis of HCC patients by systematically analyzing immune characteristic genes.

Methods: Immune-related genes were obtained, and the differentially expressed immune genes (DEIGs) between tumor and para-cancer samples were identified and analyzed using gene expression profiles from TCGA, HCCDB, and GEO databases. An immune prognosis model was also constructed to evaluate the predictive performance in different cohorts. The high and low groups were divided based on the risk score of the model, and different algorithms were used to evaluate the tumor immune infiltration cell (TIIC). The expression and prognosis of core genes in pan-cancer cohorts were analyzed, and gene enrichment analysis was performed using clusterProfiler. Finally, the expression of the hub genes of the model was validated by clinical samples.

Results: Based on the analysis of 730 immune-related genes, we identified 64 common DEIGs. These genes were enriched in the tumor immunologic related signaling pathways. The first 15 genes were selected using RankAggreg analysis, and all the genes showed a consistent expression trend across multi-cohorts. Based on lasso cox regression analysis, a 5-gene signature risk model (ATG10, IL18RAP, PRKCD, SLC11A1, and SPP1) was constructed. The signature has strong robustness and can stabilize different cohorts (TCGA-LIHC, HCCDB18, and GSE14520). Compared with other existing models, our model has better performance. CIBERSORT was used to assess the landscape maps of 22 types of immune cells in TCGA, GSE14520, and HCCDB18 cohorts, and found a consistent trend in the distribution of TIIC. In the high-risk score group, scores of Macrophages M1, Mast cell resting, and T cells CD8 were significantly lower than those of the low-risk score group. Different immune expression characteristics, lead to the different prognosis. Western blot demonstrated

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Qingping Jiang,
Third Affiliated Hospital of Guangzhou Medical University, China
Valerie Heong,
Tan Tock Seng Hospital, Singapore

***Correspondence:**
Jinfang Zheng
zhengjf71@sina.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *27 March 2021*
Accepted: *20 August 2021*
Published: *23 September 2021*

Citation:
Xu D, Wang Y, Wu J, Zhang Y, Liu Z, Chen Y and Zheng J (2021) Systematic Characterization of Novel Immune Gene Signatures Predicts Prognostic Factors in Hepatocellular Carcinoma. Front. Cell Dev. Biol. 9:686664. doi: 10.3389/fcell.2021.686664

that ATG10, PRKCD, and SPP1 were highly expressed in cancer tissues, while IL18RAP and SLC11A1 expression in cancer tissues was lower. In addition, IL18RAP has a highly positive correlation with B cell, macrophage, Neutrophil, Dendritic cell, CD8 cell, and CD4 cell. The SPP1, PRKCD, and SLC11A1 genes have the strongest correlation with macrophages. The expression of ATG10, IL18RAP, PRKCD, SLC11A1, and SPP1 genes varies among different immune subtypes and between different T stages.

Conclusion: The 5-immu-gene signature constructed in this study could be utilized as a new prognostic marker for patients with HCC.

Keywords: immune gene signatures, HCC, prognosis, tumor immune infiltration cell, pan-cancer

BACKGROUND

Hepatocellular Carcinoma (HCC) is the most common type of liver cancer and one of the leading causes of cancer death worldwide. In 2020, there were an estimated 42,810 new liver cancer cases and 30,160 related deaths in the United States (Siegel et al., 2020). The major risk factors for liver cancer include chronic Hepatitis B virus infection, hepatitis C infection, obesity, and alcoholism (El-Serag and Rudolph, 2007). At present, the treatment of HCC mainly includes surgical resection, transvascular chemoembolization, radiofrequency ablation, liver transplantation, molecular targeted therapy, and systemic chemotherapy. However, due to the concealed onset of HCC, more than 80% of patients are diagnosed at an advanced stage. The above treatment methods are usually insufficient, and the 5-year overall survival rate is less than 10% (Forner et al., 2018). Therefore, new therapies with different mechanisms are urgently needed to improve the prognosis of HCC patients. Immunotherapy has gradually become another milestone after traditional radiotherapy and chemotherapy.

The liver is a special immune tolerance organ which can effectively escape the immune response. Immunotherapy can enhance the immune response, stimulate tumor-specific immunity, break immune tolerance, and reactivate immune cells, for the purpose of recognizing and destroying tumor cells. In patients with Liver Hepatocellular Carcinoma (LIHC), CTLA-4 is often overexpressed in the liver's dendritic cells, which inhibits T cell activation and proliferation and reduces its ability to recognize tumor antigens (Sangro et al., 2013). PD-1 can inhibit the activation of antigen-specific T cells by binding with its ligand PD-L1, which can decrease the immune response of tumor microenvironment to T cells and lead to immune escape (Yarchoan et al., 2019). Anti-PD-1/PD-L1 nivolumab and pembrolizumab are currently used in the treatment of HCC. Although immunotherapy has made progress in patients with liver cancer, the response rate to immunotherapy varies from patient to patient, even regarding drug resistance. Therefore, sensitive and specific predictive biomarkers must be identified to maximize the efficacy of immunotherapy. It is necessary to determine the role of tumor immune microenvironment and immune-related genes in the diagnosis and prognosis of HCC.

In recent years, the rapid development of high-throughput sequencing technology has caused the rapid emergence of retrospective studies using public databases to provide new guidance for prognosis prediction and clinical treatment of tumors. An immune-related 7-microRNA signature was constructed to predict the prognosis of patients with HCC (Li et al., 2021). Zhang et al. (2020) used an immune-associated lncRNA signature to predict the efficacy of immune checkpoint therapy in patients with HCC. Zhao et al. (2020) constructed a 4-gene signature using immune checkpoint-related to guide the diagnosis and prognosis of HCC patients. A 5-gene signature constructed by Tian et al. (2019) using tissue-associated immune biomarkers can be used to predict the survival outcome of patients with early/middle stage HCC. In this study, we systematically described the immune characteristics and immunophenotypes of HCC tumor microenvironment using 730 immune-related genes based on four data sets. 64 immune-related genes were identified. These genes were closely related to the immune pathways including NF–kappa B signaling pathway, TNF signaling pathway, and Cytokine receptor interaction. A 5-gene signature was constructed to predict the prognosis of patients with HCC. The results of this study highlight how tumor microenvironment promotes the prognosis of patients with HCC, and can explain the complexity of immune microenvironment of HCC and provide new ideas regarding immunotherapy for HCC patients.

MATERIALS AND METHODS

Data Source and Preprocessing

RNA-Seq data and clinical follow-up data of LIHC data set and HCCDB18 data set were downloaded from the TCGA database and the HCCDB database[1], respectively. GSE14520 data set with survival data, in addition to GSE22058, GSE25097, GSE64041, and GSE36376 data sets containing tumor and para-cancerous samples without survival data, were downloaded from the Gene Expression Omnibus (GEO) database.

Abbreviations: AIC, the Akaike information criterion; DEGs, differentially expressed genes; DEIGs, differentially expressed immune genes; FDR, false discovery rate; GEO, gene expression omnibus; GO, gene ontology; GSEA, gene set enrichment analysis; HCC, hepatocellular carcinoma; KEGG, kyoto encyclopedia of genes and genomes; LIHC, liver hepatocellular carcinoma; TIIC, Tumor immune infiltration cell.

[1]http://lifeome.net/database/HCCDB/home.html

The RNA-seq data of TCGA-LIHC were processed in the following steps: (1) removing samples with no clinical follow-up information; (2) removing samples with no survival time; (3) removing samples with no Status; (4) transforming ensemble into gene symbols; and (5) taking the median with multiple gene symbol expressions. GEO data sets were processed in the following steps: (1) removing samples without clinical follow-up information; (2) removing samples without survival time and survival status; (3) converting probes into gene symbols; (4) removing the probe when corresponding to multiple genes; and (5) taking the median value of the expression of multiple gene symbols. The RNA-seq data of HCCDB18 were processed in the following steps: (1) removing samples without clinical follow-up information; (2) removing samples without survival time; (3) removing samples without status; and (4) removing samples without expression spectrum data.

After data-preprocessing, there were 197 samples, including 100 tumor samples and 97 para-cancerous samples, in GSE22058; 511 samples, including 268 tumor samples and 243 para-cancerous samples, in GSE25097; 120 samples, including 60 tumor samples and 60 para-cancerous samples, in GSE64041; and 433 samples, including 240 tumor samples and 193 para-cancerous samples, in GSE36376.The samples clinical statistics were shown in **Table 1**.

There were 365 samples for TCGA-LIHC, 203 samples for HCCDB18, and 221 samples for GSE14520 with survival data after pre-processing. Clinical statistics of the sample are shown in **Table 2**.

TABLE 1 | Clinical statistics of the sample.

Data	Tumor	Adjacent	Total
GSE22058	100	97	197
GSE25097	268	243	511
GSE64041	60	60	120
GSE36376	240	193	433
GSE14520	225	220	445
HCCDB18	212	177	389
TCGA	365	50	415
			2095

TABLE 2 | Samples with survival information.

Clinical features	TCGA-LIHC	HCCDB18	GSE14520
OS			
0	235	168	136
1	130	35	85
T stage			
T1	180		
T2	91		
T3	78		
T4	13		
TX	3		
N stage			
N0	248		
N1	4		
NX	113		
M stage			
M0	263		
M1	3		
MX	99		
Stage			
I	170		
II	84		
III	83		
IV	4		
X	24		
Grade			
G1	55		
G2	175		
G3	118		
G4	12		
GX	5		
Gender			
Male	246		
Female	119		
Age			
≤60	173		
>60	192		
Recurrence			
YES	167		
NO	198		

Identification of Differentially Expressed Genes (DEGs) and Analysis of KEGG Pathway and GO Enrichment

The limma package (Ritchie et al., 2015) was used to analyze the differentially expressed immune genes of 730 immune genes in tumor samples and para-cancerous samples from four data sets, and filter with False discovery rate (FDR) < 0.05 and $|FC| > 1.2$ as the threshold. The differentially expressed immune genes were analyzed by using the R package WebGestaltR (v0.4.2) for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis and Gene ontology (GO) function enrichment analysis. Rankaggreg was used to identify genes with the most significant variation in the differentially expressed immune genes.

Random Grouping of Training Set Samples

The 365 samples in the TCGA data set were divided into a training set and a validation set. In order to prevent random assignment bias from compromising the stability of subsequent modeling, all samples were placed in random grouping 100 times in advance. The grouping sampling was performed according to the ratio of training set: validation set = 1:1. The most suitable training sets and validation sets were selected according to the following conditions: (1) the two groups were similar in age distribution, sex, follow-up time, and patient mortality ratio;

(2) the sample size of the two groups was close after randomly clustering the gene expression profiling data sets. There were 182 samples in the final training set and 183 samples in the validation set. The training set and validation set of the TCGA data are shown in **Table 3**. The training set and validation set were tested by chi-square test. The result showed that our grouping is reasonable.

Univariate Cox Regression and Lasso Regression Analysis

R package survival coxph function was used to conduct univariate Cox proportional hazard regression analysis based on the survival data of each immune gene in the training set data. $P < 0.05$ was selected as the threshold for filtering. Lasso regression (Kukreja et al., 2006) was further compressed using the R package glmnet to reduce the number of genes in the risk model. Lasso regression is a kind of compression estimation. A more refined model is obtained by constructing a penalty function, which compresses some coefficients and sets some to zero. This retains the advantage of subset contraction, which is a kind of biased estimation with complex collinear data. Consequently, one can realize the selection of variables as well as parameter estimation and better solve the problem of multicollinearity in regression analysis. The Akaike information criterion (AIC) was further used to reduce the number of model genes. The STEPAIC method in the MASS package starts from the most complex model and removes one variable, to reduce the AIC. The smaller the value, the better the model is, as it demonstrates that the model can obtain enough fitting degree with fewer parameters.

TABLE 3 | TCGA training set and validation set sample information.

Clinical features	TCGA-LIHC train	TCGA-LIHC test
OS		
0	122	113
1	60	70
T stage		
T1	93	87
T2	47	44
T3	32	46
T4	7	6
TX	3	0
N stage		
N0	114	134
N1	0	4
NX	68	45
M stage		
M0	122	141
M1	2	1
MX	58	41
Stage		
I	87	83
II	42	42
III	33	50
IV	2	2
X	18	6
Grade		
G1	29	26
G2	89	86
G3	54	64
G4	6	6
GX	4	1
Gender		
Male	121	125
Female	61	58
Age		
≤60	91	82
>60	91	101
Recurrence		
YES	76	91
NO	106	92

Expression and Prognosis Value of the Five Genes in Pan-Cancer

The transcriptome data of 33 cancers and the normal tissue data were downloaded from the UCSC Xena database and the GTEX database, respectively. The box maps of ATG10, IL18RAP, PRKCD, SLC11A1, and SPP1 in tumor tissues and normal tissues of 33 cancer types were plotted using the ggplot2 package. The prognosis forest map of each gene in 33 cancer types was also drawn with the ggplot2 package.

Sample Collection and Western Blotting

Liver hepatocellular carcinoma and adjacent normal tissues were collected from 4 patients, immediately placed in liquid nitrogen, and preserved at −80°C. Take the tumor tissue and normal tissue adjacent to the cancer into small pieces and put them into the tube, add lysis buffer RIPA (1% Triton X-100, 50 mM Tris-HCl pH7.4, 150 mM Na Cl, 10 mM EDTA, 100 mM Na F, 1 mM Na 3 VO 4, 1 mM PMSF, 2 μg/ml Aprotinin) (1 ml lysate is added to 250mg tissue). Use a homogenizer to homogenize at low speed for 30 s each time, and ice bath for 1 min between each time until the tissue is completely lysed. Centrifuge at 13,000 rpm for 25 min, take the supernatant, and quantify the protein by Coomassie brilliant blue method. After mixing with 3 × sample buffer, boil for 5 min. The sample (30–50 μg/lane) was electrophoresed in a 12% SDS-polypropylene gel for 3 h, and then transferred to a nitrocellulose membrane (voltage: 2 mV/cm^2; time: 120 min). After sealing with 5% skimmed milk for 1 h, cut the transfer film according to the molecular weight marked by the pre-stained Marker, and add the primary antibodies separately at 4°C overnight. After washing 4 times with TTBS, add secondary antibody (1:2000) for 30 min at room temperature. After washing 4 times with TTBS again, the color will be developed by ECL method. The primary antibodies are as follows: ATG10 (1:1000, ab124711, abcam), PRKCD (1:1000, SAB4300539, sigma), IL18RAP (1:1000, AV42154, sigma), SLC11A1 (1:10000, SAB2108019, sigma), SPP1 (1:1000, ab214050, Sigma).

After rinsing 3 times (10 min each time) with tris-buffered saline, the membrane was incubated with horseradish peroxidase-conjugated secondary antibody against rabbit IgG

(1:5000, Amersham Bioscience, Piscataway, NJ, United States) for 1 h at room temperature. After washout, the membrane was developed using enhanced chemiluminescence reagents (Pierce, Rockford, IL, United States) and visualized using a chemiluminescence system (PTC-200, Bio-Rad Laboratories, Hercules, CA, United States). All Western blots were repeated 3 times.

Immunologic Correlation of the Five Genes in Pan-Cancer

The immune score and matrix score of each patient were assessed using the estimate package, while the levels of ATG10, IL18RAP, PRKCD, SLC11A1, and SPP1 in pan-cancer were analyzed. The corrplot package was used to visualize the correlation between each gene and the immune score and matrix score of each cancer type. A timer database[2] was used to analyze the association of five genes with immune cells.

Clinical Correlation and Pathway Enrichment of the Five Genes in HCC

The levels of the five genes in different clinical stages, tumor grade, T stage, N stage, and M stage in HCC was further analyzed.

[2]https://cistrome.shinyapps.io/timer/

The patients with HCC were divided into high- and low-risk groups according to levels of the five genes, and the single gene Gene Set Enrichment Analysis (GSEA) of the five genes was analyzed using the clusterProfiler package.

RESULTS

64 Differentially Expressed Immune Genes Were Identified

A total of 443, 201, 184, and 205 differentially expressed immune genes were identified in the GSE22058, GSE25097, GSE64041, and GSE36376 data sets, respectively. The differentially expressed immune genes were mapped by volcanoes and heatmaps (**Figures 1A–H**, S2-S5.CSV). A total of 64 differentially expressed immune genes were obtained from the intersection of differentially expressed immune genes in four data sets (**Supplementary Figure 1A**). The results of GO functional enrichment analysis demonstrate that there were 277 items with significant differences in BP, the top ten of which are shown in **Figure 1I**. There were 15 items with significant differences in CC, the top ten of which are shown in **Figure 1J**. There were 16 items with significant differences in MF, the top 16 of which are shown in **Figure 1K**. The results of the first ten KEGG pathways

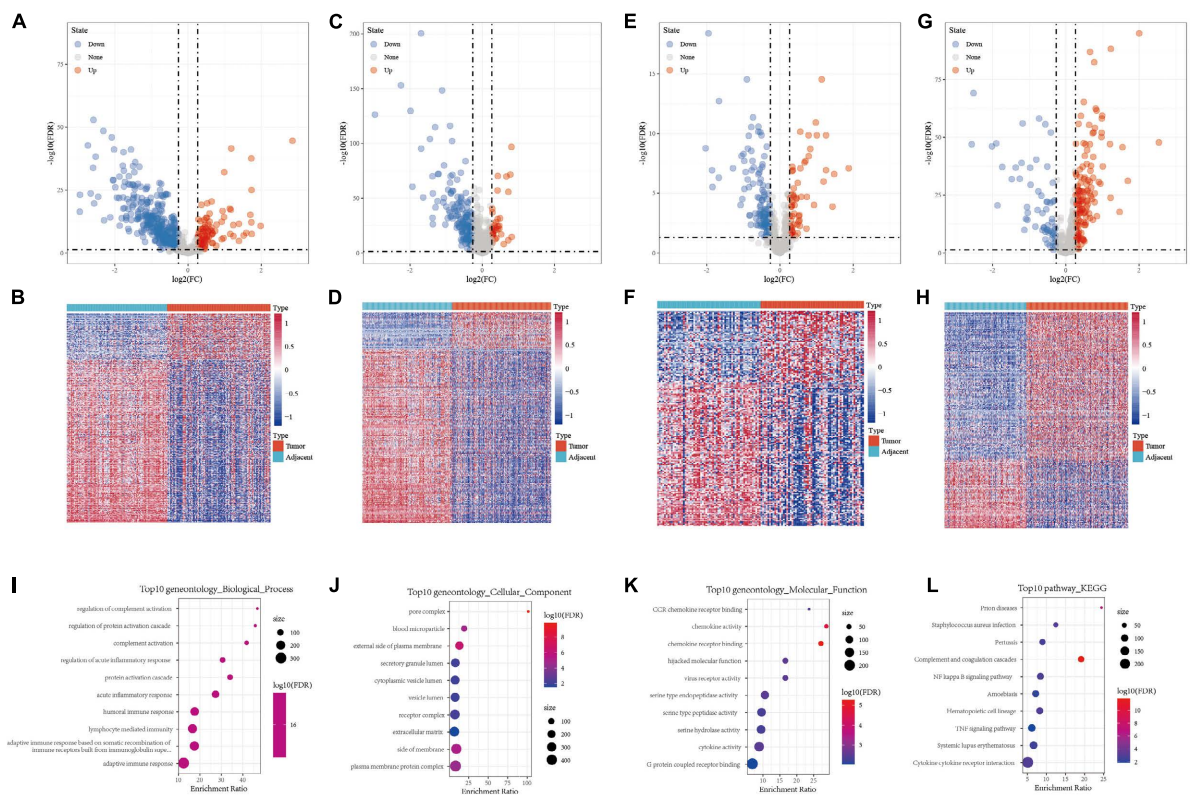


FIGURE 1 | (A,B) Volcano map and heat map of differentially expressed genes in GSE22058. **(C,D)** Volcano map and heat map of differentially expressed genes in GSE25097. **(E,F)** Volcano map and heat map of differentially expressed genes in GSE64041. **(G,H)** Volcano map and heat map of differentially expressed genes in GSE36376. **(I)** Bubble chart of differentially expressed genes in biological process. **(J)** Bubble chart of differentially expressed genes in cellular component. **(K)** Bubble chart of differentially expressed genes in molecular function. **(L)** Bubble chart of differentially expressed genes in KEGG pathway.

showed that the gene was significantly enriched in the NF–kappa B signaling pathway, TNF signaling pathway, and Cytokine receptor interaction (**Figure 1L**).

The Expression of Differentially Expressed Immune Genes in HCC Differed From That in Normal Tissues

The first 15 genes with the most significant differentially expressed genes were identified including HAMP, C9, CXCL14, MARCO, CXCL12, HSD11B1, C7, C8A, MBL2, C6, EGR1, CFP, C8B, CXCL2, and CCL19. GSE22058, GSE25097, GSE64041, GSE36376, TCGA, HCCDB18, and GSE14520 were used to analyze the expression of these genes in tumor samples and para-cancer samples. HAMP, C9, CXCL14, MARCO, CXCL12, HSD11B1, C7, C8A, and MBL2 were differentially expressed, as shown in **Figure 2**. The differentially expressed profile of the other genes are shown in **Supplementary Figure 1B**. The results demonstrated significant differences in the expression of these hub genes between HCC and adjacent tissues, suggesting that immune-related genes play an important role in HCC.

The Construction of Risk Model

A univariate cox regression model was used to analyze 730 immune-related genes, which produced 132 DEGs, as shown in S8.csv. Lasso regression was used to further compress the genes. The change track of each independent variable is shown in **Figure 3A**, in which we can see that with the gradual increase of lambda, the number of independent variable coefficients gradually trends to 0. We used fivefold cross-validation to build the model and analyzed the confidence interval under each lambda, as shown in **Figure 3B**. The results showed that the model is optimal when lambda = 0. 0904, and the 12 genes were obtained when lambda = 0. 0904. The AIC regression further compressed the number of the hub genes into five: ATG10, IL18RAP, PRKCD, SLC11A1, and SPP1. The KM curve of the five genes is shown in **Supplementary Figure 2**, which shows that all the five genes could separate TCGA training set samples significantly ($p < 0.05$). The 5-gene signature formula is as follows: RiskScore = 0.841 * ATG10-0.989 * IL18RAP + 0.507 * PRKCD + 0.437 * SLC11A1 + 0.104 * SPP1.

RiskScore was calculated for each sample based on level of expression. The survival time of the high RiskScore sample is significantly less than that of the low RiskScore sample, suggesting that the former has a worse prognosis (**Figure 3C**). timeROC was used to analyze the ROC of RiskScore. The predictive classification efficiency of 1, 3, and 5 years is shown in **Figure 3D**, which demonstrates that the model has a high AUC. The samples greater than zero after zscore (85 samples) were divided into the high-risk group (85 samples) and samples less than zero after zscore were divided into the low-risk group (97 samples). The difference between the survival curves of the high- and low-risk groups is shown in **Figure 3E** ($p < 0.0001$).

Verifying the Robustness of the Risk Model

The robustness of the risk model was verified using internal data sets (TCGA validation set and all data sets) and external data sets (GSE14520 and HCCDB18), respectively. Using the same model and coefficients as the training set, the risk score of

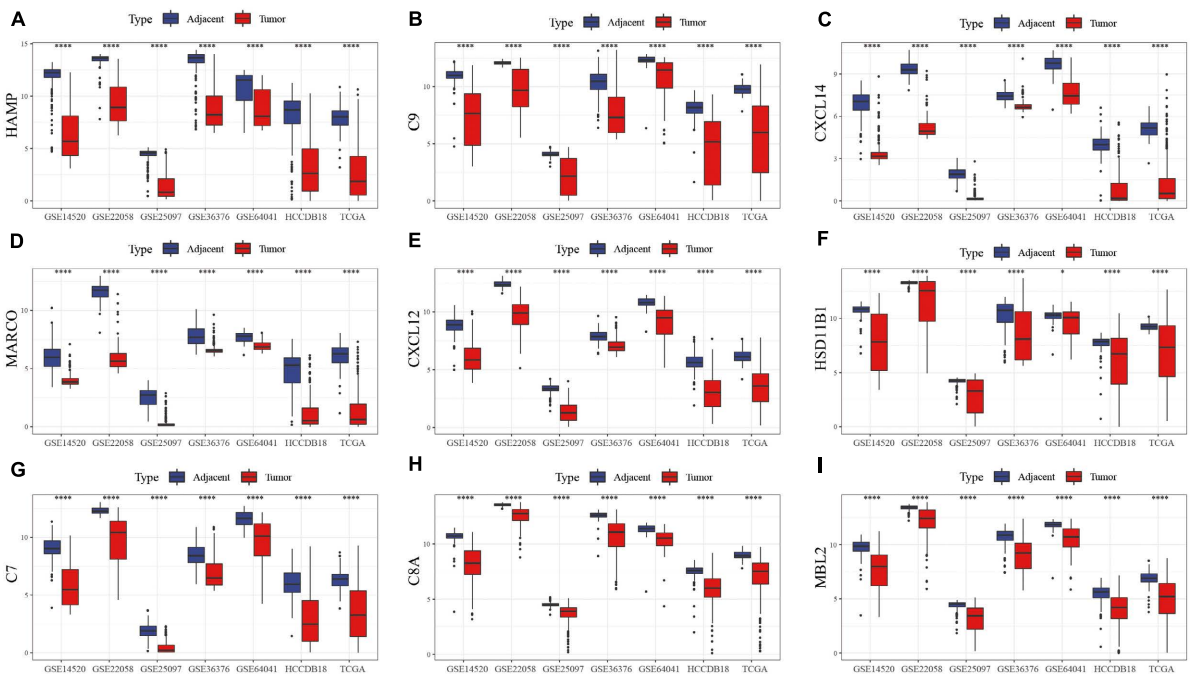


FIGURE 2 | Box map of DEGs expression in tumor and para-cancer samples in different data sets. The differences in the expression of **(A)** HAMP, **(B)** C9, **(C)** CXCL14, **(D)** MARCO, **(E)** CXCL12, **(F)** HSD11B1, **(G)** C7, **(H)** C8A, and **(I)** MBL2 genes, respectively, in tumor and adjacent samples.

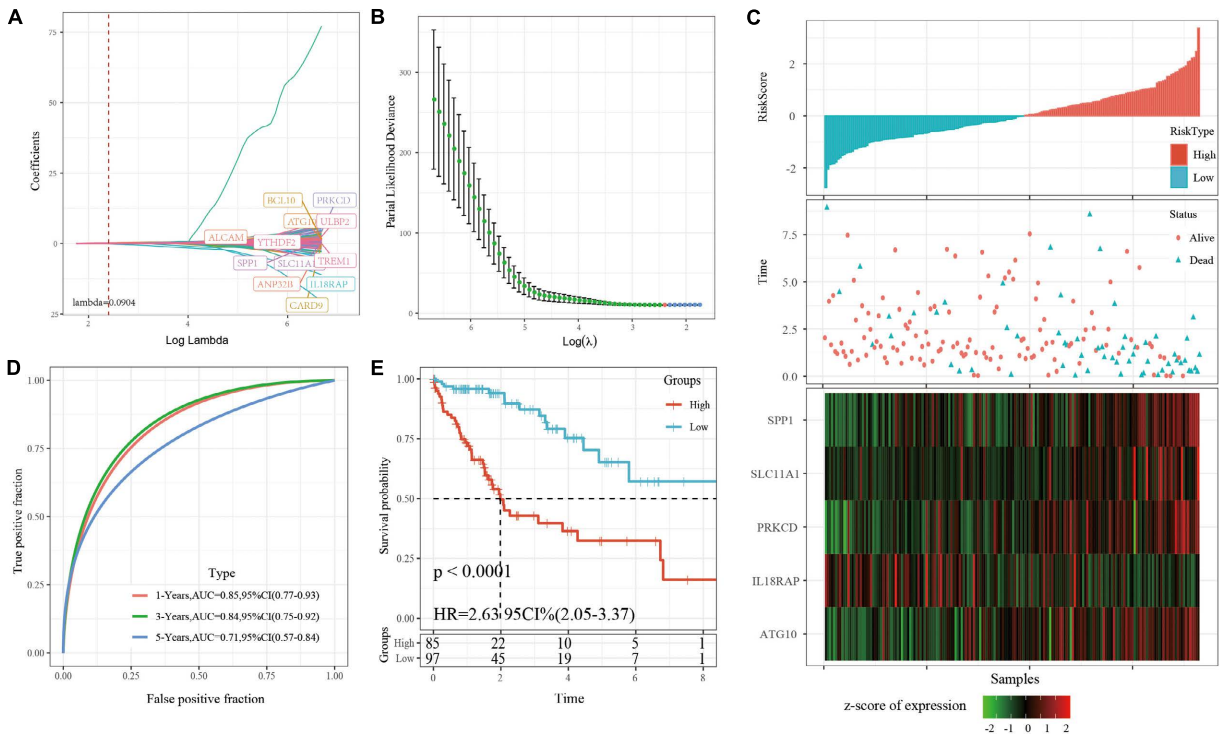


FIGURE 3 | (A) For the changing trajectory of each independent variable, the horizontal axis represents the log value of the independent variable lambda, and the vertical axis represents the coefficient of the independent variable. **(B)** The confidence interval under each lambda. **(C)** RiskScore, survival time, survival status, and 5-gene expression in TCGA training set. **(D)** ROC curve and AUC of 5-gene signature classification. **(E)** The KM survival curve distribution of 5-gene signature in training set.

each sample was calculated according to the level of expression, and the RiskScore distribution of the sample was plotted. The RiskScore distributions of the TCGA validation set, all TCGA data sets, independent validation data sets GSE14520, and independent validation data sets HCCDB18 are shown in **Figures 4A**, **5A** and **Supplementary Figures 3A**, **4A**, respectively. The survival time of the high RiskScore sample is evidently less than that of the low RiskScore sample, indicating that the former has worse prognosis. The ROC curves of 1-year, 3-year, and 5-year predictive classification efficiency are shown in **Figures 4B**, **5B** and **Supplementary Figures 3B**, **4B**, respectively. The samples whose RiskScore was greater than zero after zscore were divided into the high-risk group and those with less than zero after zscore were divided into the low-risk group. In TCGA, 86 samples were divided into the high-risk group and 97 samples were divided into low-risk group. The survival curve indicated a difference between the two groups ($p < 0.01$) (**Figure 4C**). In all TCGA data sets, 171 samples were classified as high-risk and 194 were classified as low-risk. The survival curve indicated a difference between the two groups ($p < 0.0001$) (**Figure 5C**). In GSE14520, 110 samples were divided into the high-risk group and 111 samples were divided into the low-risk group. The survival curve indicated a difference between the two groups ($P = 0.0084$) (**Supplementary Figure 3C**). In the independent validation data set HCCDB18, 98 samples were divided into the high-risk group and 105 samples were divided into the low-risk group. The results indicated a difference between the two groups ($p < 0.01$) (**Supplementary Figure 4C**).

Prognostic Analysis of Clinical Subgroups Based on RiskScore

A stratified analysis of clinical subgroups variables including Age, Gender, TNM stage, Clinical Satge, and Grade were performed based on the expression of RiskScore. The results showed that RiskScore can classify Age ≤ 60, Age > 60, Female, Male, T1 + T2, N0, The M0, Stage I + II, Stage III + IV, Grade 1 + 2, Grade 3 + 4 groups into two groups with significant prognosis differences (**Figures 6A–L**). Therefore, our prognostic score can be used as a prognostic marker for clinical subgroups. Comparison of the distribution of RiskScore among TCGA clinical feature groups indicated that the RiskScore differed between T Stage, Stage, Grade, and relapse groups (**Figures 7A–D**, $P < 0.05$).

The Expression of RiskScore on Different Clinical Features and the Construction of Nomogram

In the TCGA data set, univariate and multivariate regression analysis showed a correlation between RiskScore and survival time ($p < 0.05$), indicating that 5-gene signature was independent in prognostic prediction (**Figures 8A,B**).

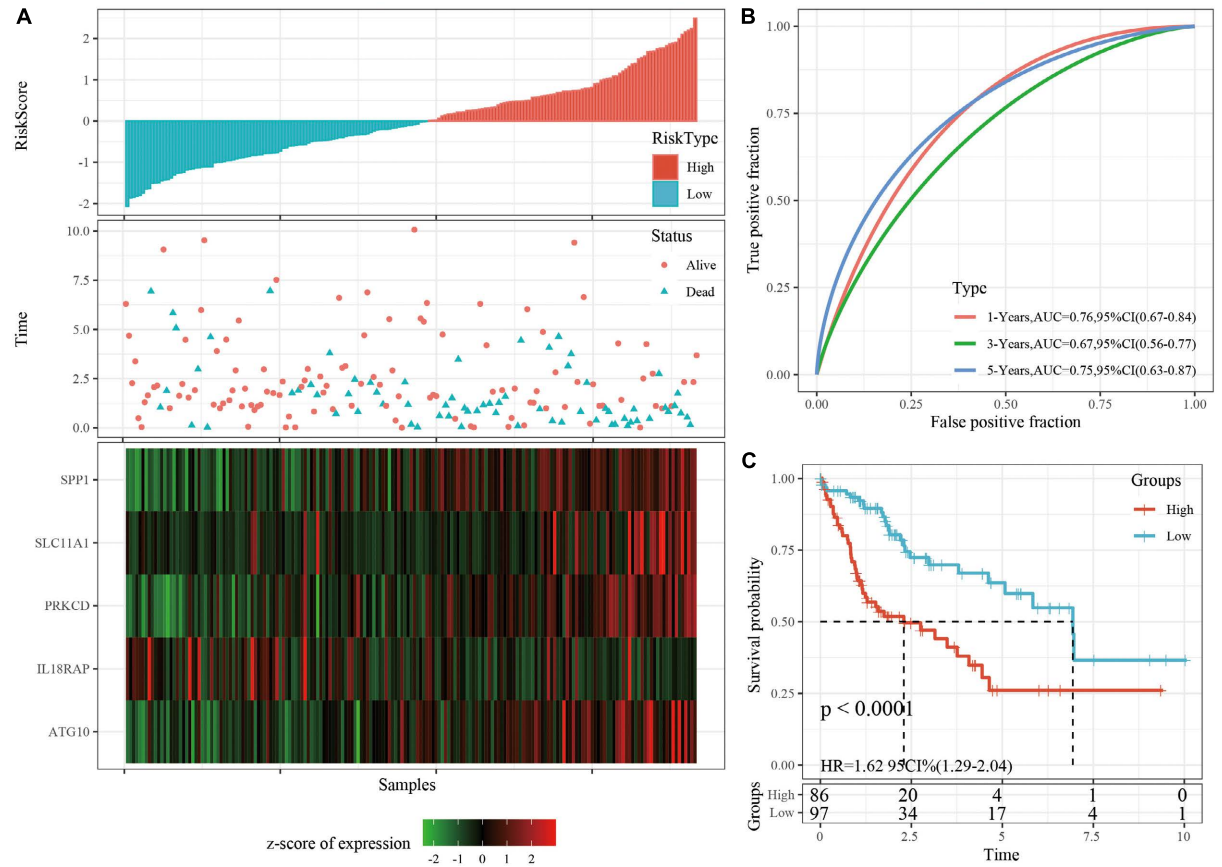


FIGURE 4 | (A) RiskScore, survival time, survival status, and 5-gene expression in TCGA validation set. **(B)** ROC curve and AUC of 5-gene signature classification. **(C)** The KM survival curve distribution of 5-gene signature in validation set.

Nomogram can display risk model results directly and effectively. We constructed a nomogram according to the results of univariate and multivariate regression analysis (**Figure 8C**). Based on the results of RiskScore, as the RiskScore has the greatest influence on the survival prediction, the 5-gene risk model can predict the prognosis better. In addition, the accuracy of the 1-, 3-, and 5-year nomogram was corrected, which proved that the method has good performance.

Comparison of Immune Cell Score Between High-Risk and Low-Risk Groups

The distribution of 22 types of immune cells in TCGA, GSE14520, and HCCDB18 showed a consistent trend (**Figures 9A–C**). To visualize the differences, we analyzed the expression of StromalScore, ImmuneScore, ESTIMATE Score, and 22 types of immune cells in different RiskScore groups in the TCGA data set. Box plots showed that the high RiskScore group had significantly lower scores of Macrophages M1, Mast cell resting, and T cells CD8 than the low score group (**Figure 9D**). ESTIMATE, Immune Score, and Stromal Score had significantly higher immune scores in the low RiskScore group than the high RiskScore group (**Figure 9E**), suggesting that different expression characteristics of immune cells may promote prognosis in different populations. The association of CIBERSORT score and estimated scores of the high-risk and low-risk groups is shown in **Figure 9F**.

Comparison of Risk Models

We selected four prognostic risk models including Ke et al. (2018); Zheng et al. (2018), Liu et al. (2019), and Hu et al. (2020) to compare with our 5-gene model. In order to make the model comparable, we calculated the risk score of each OS sample in TCGA using the same method according to the corresponding genes in these four models. The samples greater than zero after zscore were divided into the high-risk group and those less than zero after zscore were divided into the low-risk group, after which the differences in OS outcomes between the two groups were calculated. The ROC and KM curves of the four models are shown in **Figures 10A–H**. Although OS prognosis differed between the high-risk and low-risk group samples among the four prognostic models, the AUC values at 1-, 3-, and 5-year are lower than those of our models. To compare the prediction performance of these models in LIHC samples, the rms package in R was used to calculate the concordance index (C-index) of

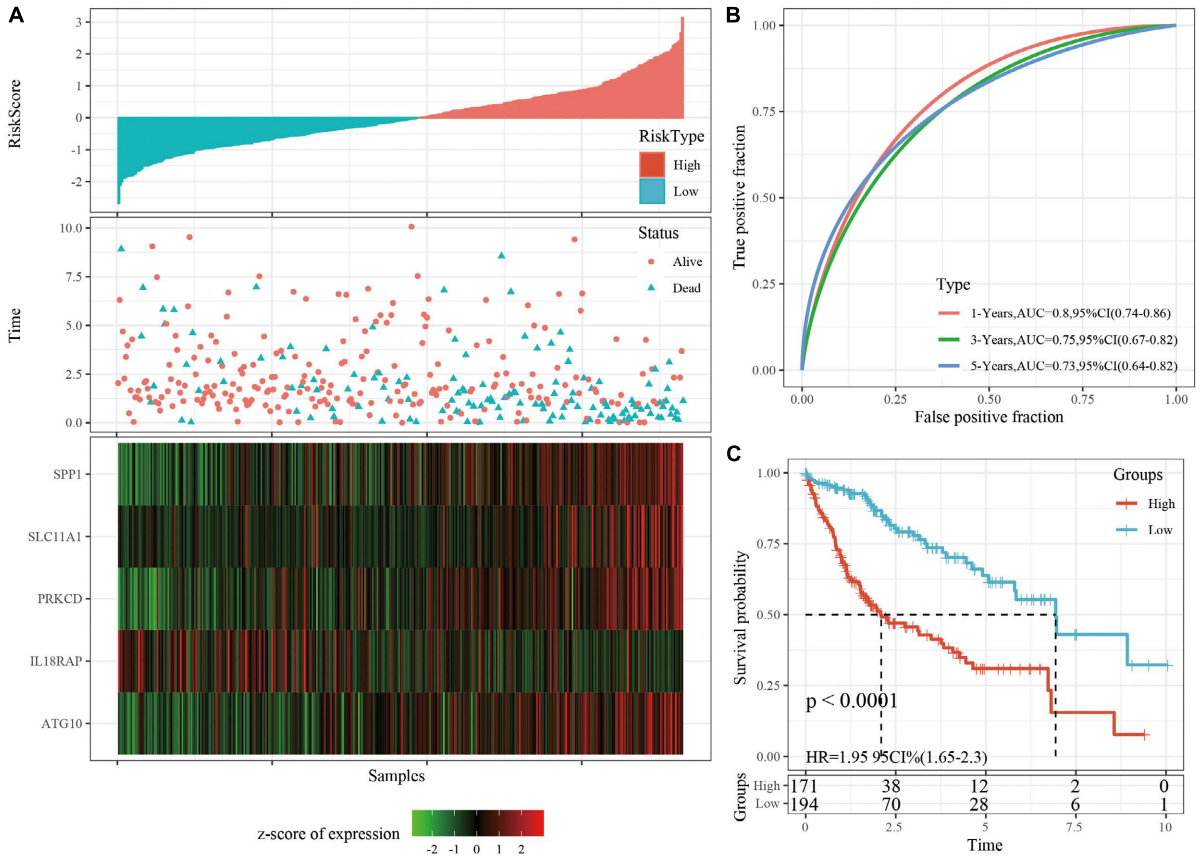


FIGURE 5 | (A) RiskScore, survival time, survival status, and 5-gene expression in TCGA data sets. **(B)** ROC curve and AUC of 5-gene signature classification. **(C)** The KM survival curve distribution of 5-gene signature in TCGA data sets.

the five models. The results showed that the C-index of our model was the highest (**Figure 10I**), indicating that the overall performance of our model is the best among the five models.

Clinical Expression and Prognostic Value of the Five Genes in Pan-Cancer

The results showed that mRNA levels of ATG10, IL18RAP, PRKCD, SLC11A1, and SPP1 differed between tumor tissues and normal tissues in most of the 33 cancer types. ATG10, SPP1, and PRKCD were overexpressed in HCC (**Figures 11A,C,D**), while IL18RAP and SLC11A1 had low expression (**Figures 11B,E**). Furthermore, IL18RAP had low expression in most tumors, while SPP1 was overexpressed.

We measured the protein expression of the 5 genes in 4 pairs of LIHC and normal samples. Compared with the normal samples, ATG10, PRKCD, and SPP1 were highly expressed in the tumor samples. Expression of IL18RAP and SLC11A1 were relatively higher in the normal samples than LIHC samples. The results were almost consistent with the database (**Figure 11F**).

The overall survival analysis showed that ATG10, PRKCD, SPP1, and SLC11A1 were highly expressed in HCC with poor prognosis (**Figures 12A,C,D,E**). Besides, SPP1 was a high-risk gene in most tumors. The high expression of IL18RAP has a good prognosis in HCC and it is a protective gene in most tumors (**Figure 12B**).

The Relation Between Five Genes and Tumor Microenvironment in Pan-Cancer

The results showed that IL18RAP, PRKCD, SLC11A1, and SPP1 were positively correlated with the immune score in most tumors, including LIHC. This indicates that the immune score of most tumors increased with the increase of these genes' expression. IL18RAP was strongly correlated with immune score in most tumors (**Figure 13A**). IL18RAP, SLC11A1, and SPP1 were positively correlated with StromalScore in most tumors, including LIHC. This indicates that StromalScore increased with the increase of IL18RAP, SLC11A1, SPP1 gene expression (**Figure 13B**). ATG10 was weakly correlated with immune core and StromalScore in tumors (**Figure 13C**). IL18RAP was highly correlated with B cell, macrophages, Neutrophil, dendritic cell, CD8 cell, and CD4 cell (**Figure 13D**). Both SPP1 and PRKCD were positively correlated with various types of immune cells, and were most strongly correlated with macrophages (**Figures 13E,F**).

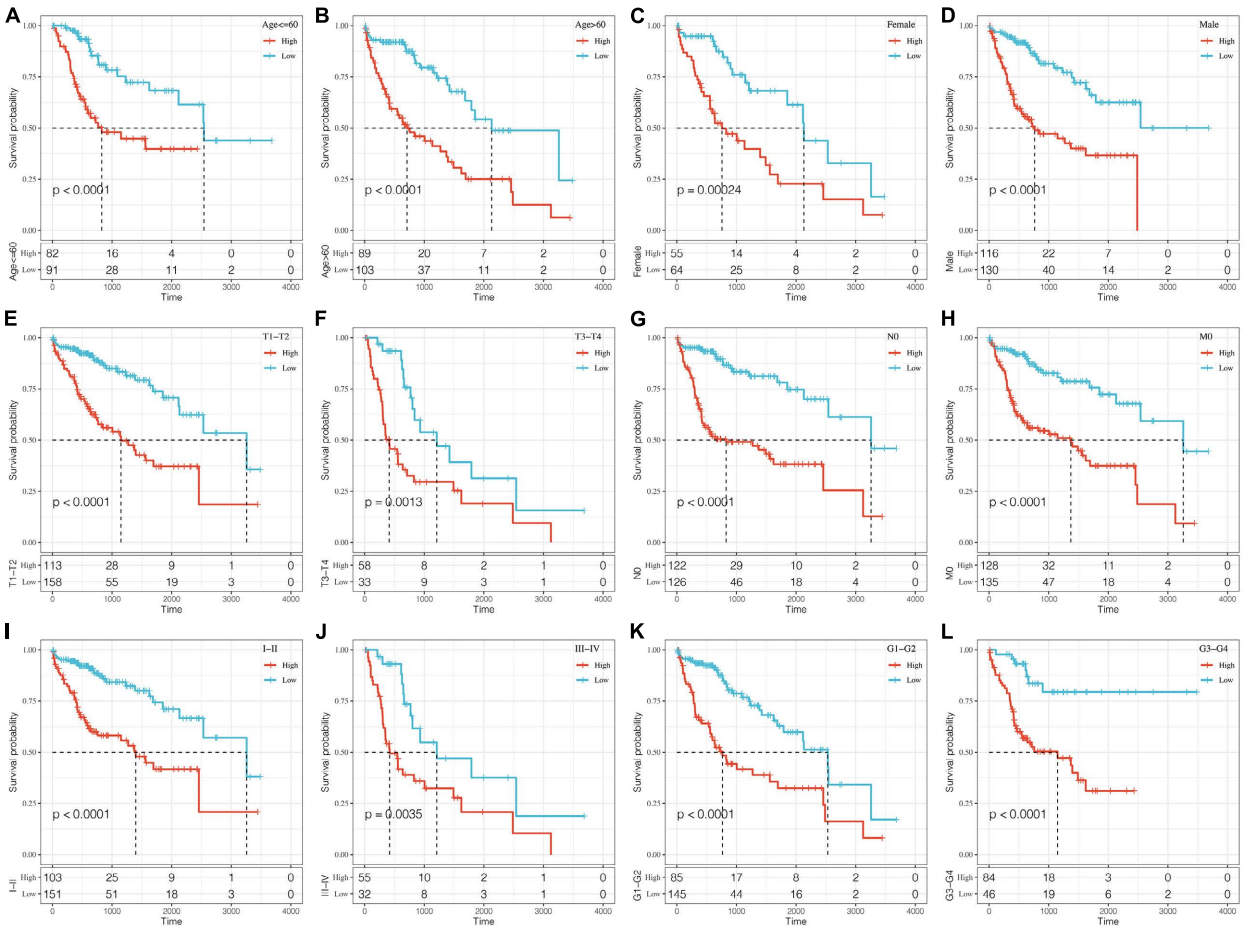


FIGURE 6 | Prognostic analysis of clinical subgroups based on RiskScore. The horizontal axis represents survival time, and the vertical axis represents survival probability. Blue represents low expression group; red represents high expression group. **(A–L)** Based on Riskscore, survival curves were analyzed for Age ≤ 60, Age > 60, Female, Male, T1+T2, T3+T4, N0, M0, Stage I+II, Stage III+IV, Grade 1+2, and Grade 3+4 groups, respectively.

SLC11A1 also showed a high positive correlation with macrophages (**Figure 13G**).

Potential Mechanism of Five Genes Involved in HCC

Studies of more than 10,000 tumor samples from 33 cancer types on the TCGA database have identified six immune subtypes: C1 (wound healing), C2 (INF-r dominant), C3 (inflammation), C4 (lymphocyte depletion), C5 (immunologically silent), and C6 (TGF-BETA predominates) (Thorsson et al., 2018). Since the C5 subtype is immunologically silent, we analyzed the association of the five genes with C1, C2, C3, C4, and C6 subtypes.

We further analyzed whether there were significant differences in the expression of these five genes among different clinical characteristics, including immune subtypes, clinical stages, tumor grades, and T stages of HCC. The results showed that levels of ATG10, IL18RAP, PRKCD, SLC11A1, and SPP1 varied among different immune subtypes and T stages (**Figures 14A,D**). Levels of PRKCD and SPP1 also varied among different clinical stages and tumor grades (**Figures 14B,C**). Pathway enrichment analysis showed that ATG10 was mainly enriched in HALLMARK_GLYCOLYSIS and G2M_CHECKPOINT. IL18RAP was mainly enriched in EPITHELIAL_MESENCHYMAL_TRANSITION, TNFA_SIGNALING_VIA_NFKB pathway. SPP1 was mainly enriched in HALLMARK_E2F_TARGETS, MTORC1_SIGNALING pathway. PRKCD was mainly enriched in HALLMARK_ALLOGRAFT_REJECTION, HALLMARK_MITOTIC_SPINDLE pathway. SLC11A1 was mainly enriched in EPITHELIAL_MESENCHYMAL_TRANSITION, P53_PATHWAY (**Figures 14E–I**).

DISCUSSION

Hepatocellular carcinoma is the most common type of primary liver cancer and the sixth most common cancer worldwide (Desai et al., 2019). Considered the second leading cause of cancer deaths, it causes more than 700,000 deaths worldwide each year and thus poses a serious threat to human health (Halegoua-De Marzio and Hann, 2014). The liver is a quintessential immune tolerance organ due to its unique immune microenvironment. By blocking the immune escape of tumor cells and killing

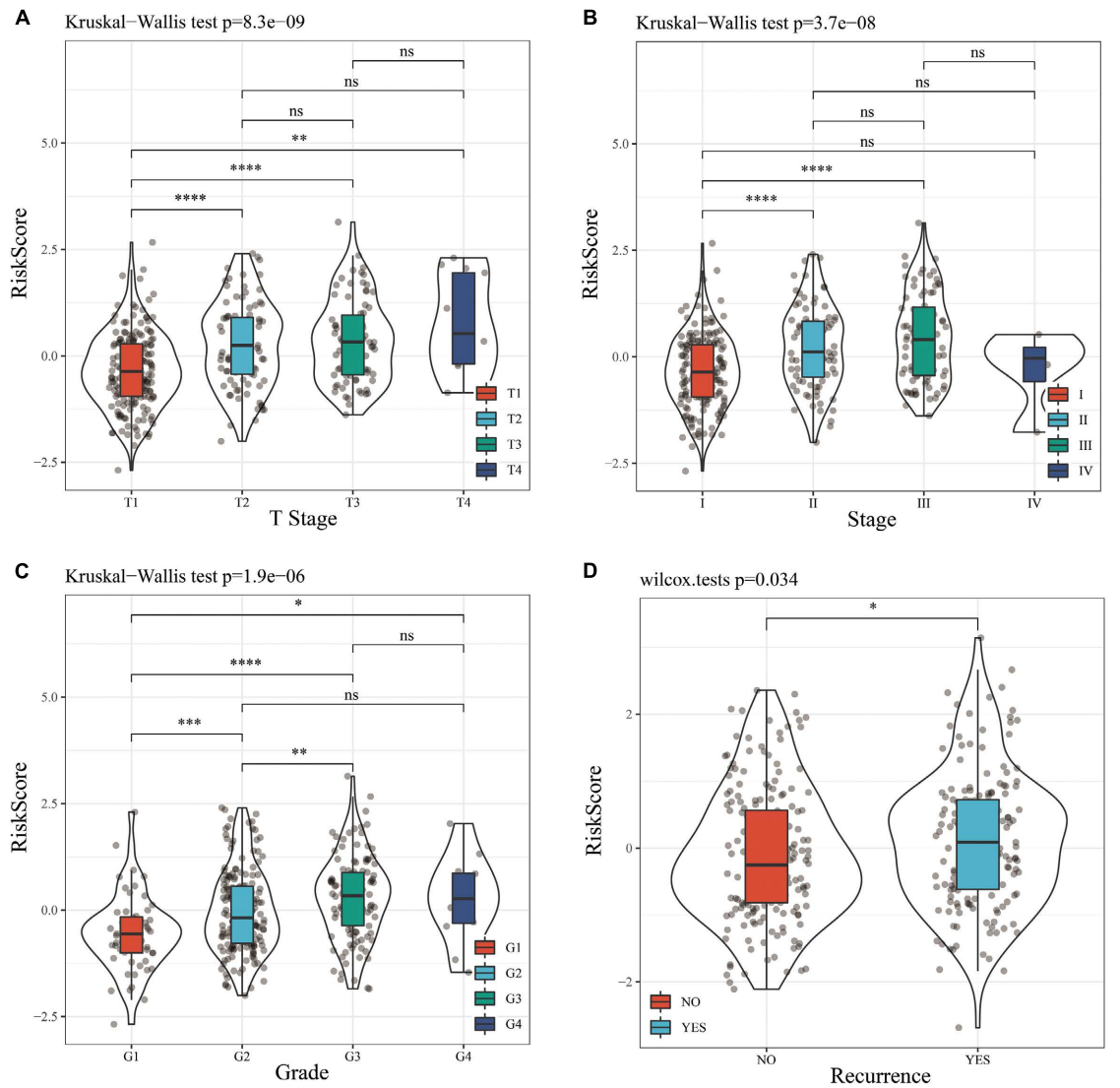


FIGURE 7 | (A) Comparison of RiskScore among T Stage grouping samples. **(B)** Comparison of RiskScore among stage grouping samples. **(C)** Comparison of RiskScore among grade grouping samples. **(D)** Comparison of RiskScore between grouped samples with or without recurrence. * means *P* value < 0.05; ** means *P* value < 0.01; *** means *P* value < 0.005; **** means *P* value < 0.001; *ns*, no significant.

liver cancer cells, immunotherapy may be the most promising treatment for completely killing HCC cells. The response to immunotherapy usually depends on the interaction between tumor cells and their surrounding immune microenvironment. Comprehensive and systematic analysis of immune-related genes in HCC is therefore helpful for guiding immunotherapy and prognosis for HCC patients.

In this study, we first analyzed the differentially expressed immune genes of 730 immune-related genes in tumor and para-cancer samples using four data sets: GSE22058, GSE25097, GSE64041, and GSE36376. The results showed that 64 immune-related genes were involved in the development and progression of HCC through immune pathways such as NF–kappa B signaling pathway, TNF signaling pathway, and Cytokine-cytokine receptor interaction. The over-activation of NF-kappa B has been found to be closely related to the pathogenesis of HCC (Dai et al., 2017). PIGU (phosphatidylinositol glycan anchor biosynthesis class U) can decelerate the malignant progression of HCC by activating the NF-kappa B signaling pathway and promoting immune escape (Wei et al., 2020). TNF-α is an important component of the inflammatory microenvironment of HCC: it can promote the expression of B7-H1 in HCC cells induced by IFN-γ, thereby activating adaptive immune tolerance. Using RankAggreg analysis, we selected the top 15 most closely associated immune genes, including the chemokine families CXCL14, CXCL12, and CXCL2. CXCL14 is responsible for the recruitment and maturation of immune cells and promotes the movement of epithelial cells, which helps to establish

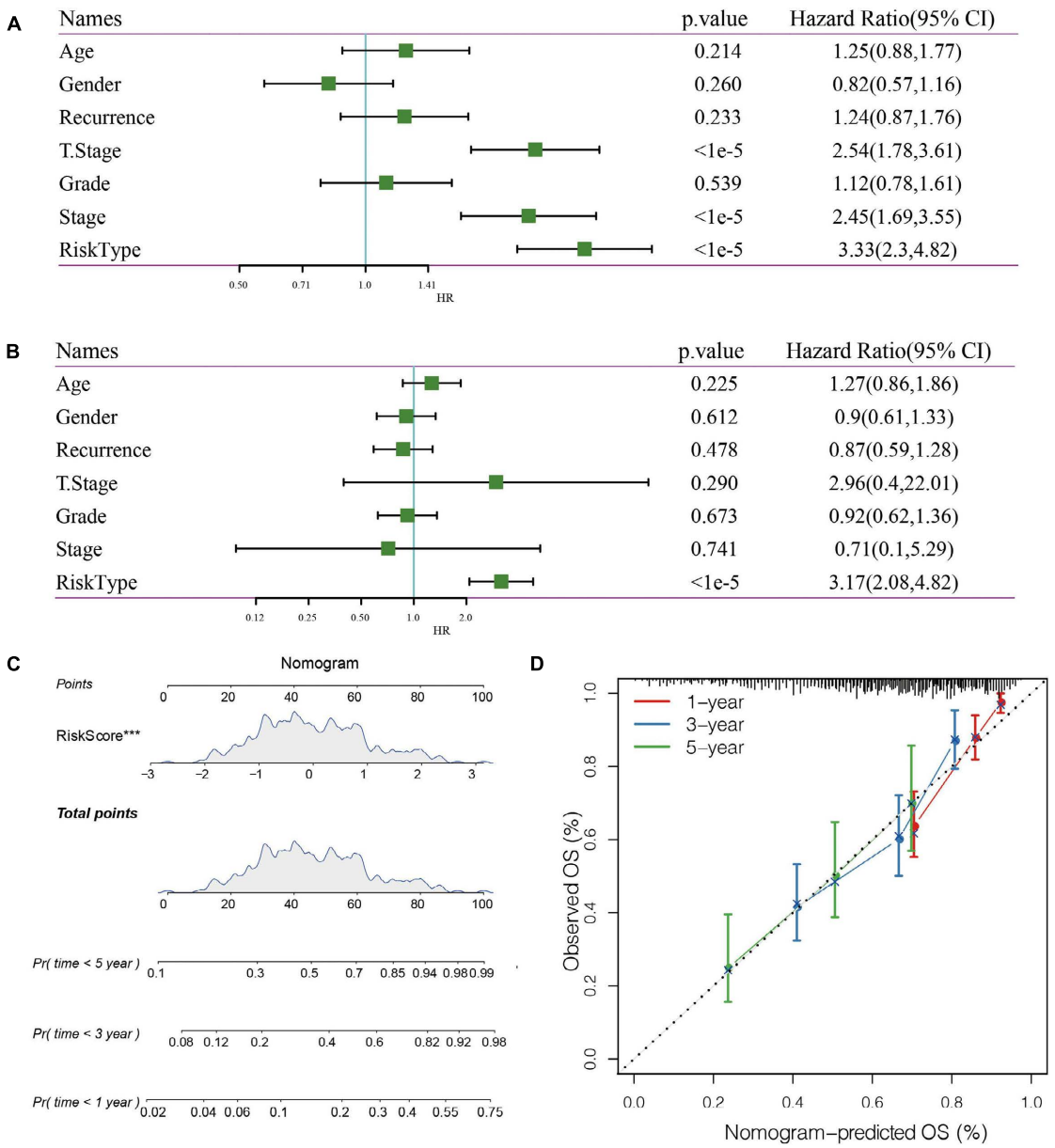


FIGURE 8 | (A) Univariate Cox analysis. **(B)** Multivariate Cox analysis. **(C)** Nomogram model. **(D)**: 1-, 3-, and 5- year correction curve of the model.

immune surveillance in normal epithelium (Westrich et al., 2020). In addition, CXCL14 suppresses human papillomavirus-associated head and neck cancer through up-regulation of MHC-1 expression and antigen-specific CD8T cell response (Westrich et al., 2019). DNA methyltransferase 1 (DNMT1) can impair the homing ability of cytotoxic T cells to tumor cells by down-regulating CXCL12 (Li B. et al., 2018). High expression of CXCL2 enhances neutrophil recruitment and activation in an autocrine and/or paracrine manner (Li et al., 2016).

Lasso regression analysis and AIC analysis were used to further compress the target gene and construct five immune gene prognostic models: ATG10, IL18RAP, PRKCD, SLC11A1, and SPP1. To explore the potential mechanism of five genes in the malignant progression of HCC, we further analyzed the expression and prognosis value of five genes in pan-cancer. The results suggested that IL18RAP and SLC11A1 genes had low expression in HCC, while SLC11A1 with high expression had poor prognosis. IL18RAP is a member of the interleukin receptor family and facilitates IL-8-influenced signal transduction by producing IFN-γ (Cheung et al., 2005). In renal cell carcinoma, the high expression of IL18RAP suggests a poor prognosis (Yamada et al., 2018). The IL18RAP polymorphism may be associated with malignant progression of esophageal carcinoma (Zhu et al., 2016), suggesting that IL18RAP may

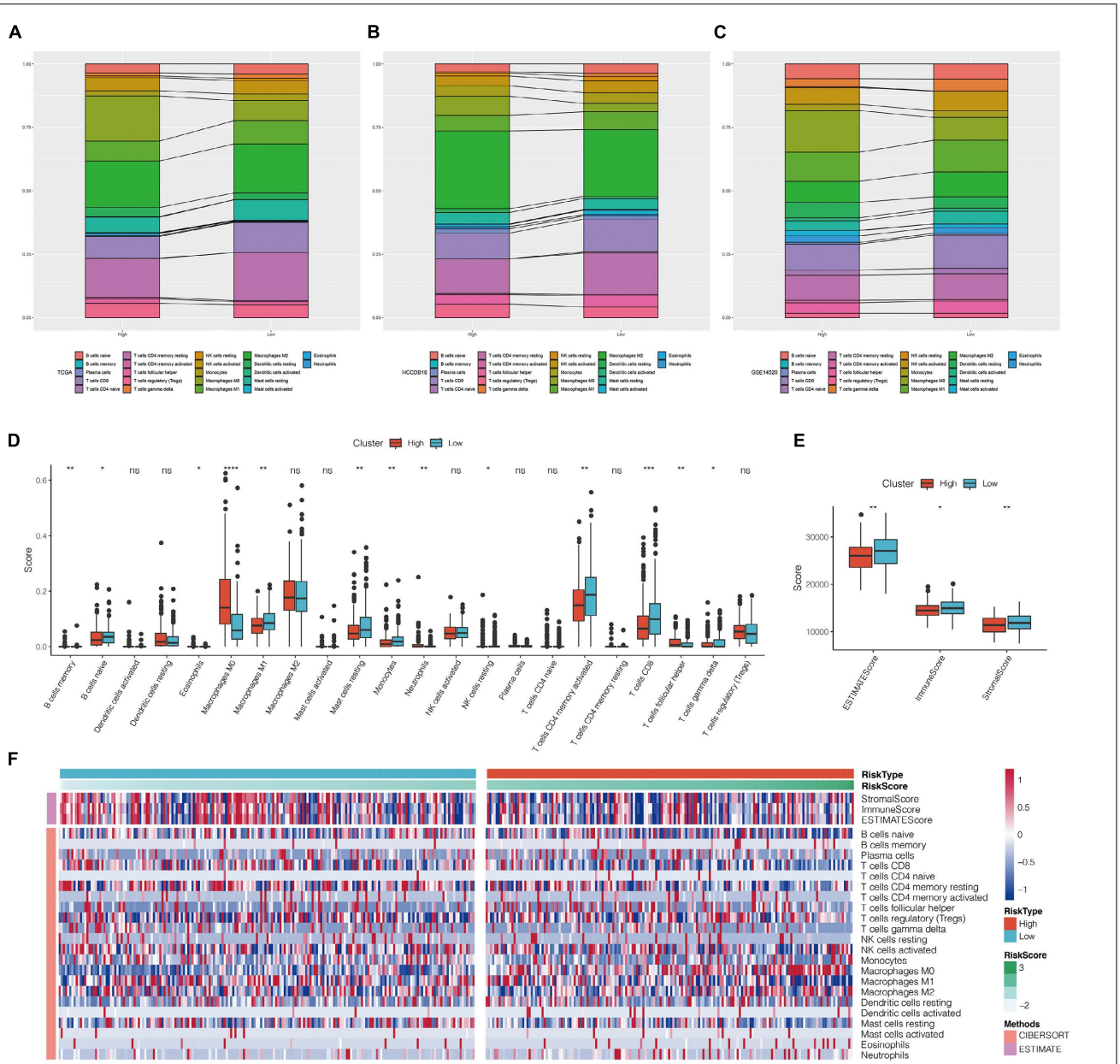


FIGURE 9 | (A–C) Landscape of immune cell infiltration score in high-risk and low-risk groups of TCGA, HCCDB18, and GSE14520 data sets. **(D)** Comparison of CIBERSORT immune score between high-risk and low-risk groups in TCGA data set. **(E)** Comparison of estimate immune score between high-risk and low-risk groups in TCGA data set. **(F)** Heat map of correlation between CIBERSORT immune score and estimate immune score in high-risk and low-risk groups. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.005$, and $^{****}P < 0.001$; *ns*, no significant.

have tissue specificity in different types of cancer. ATG10 is an autophagic E2-like enzyme which has been found to be overexpressed in colorectal cancer (Jo et al., 2012) and lung cancer (Xie et al., 2016). The overexpression of ATG10 is also associated with lymphatic invasion and lymph node metastasis in colorectal cancer (Jo et al., 2012). ATG10 promotes the proliferation and migration of lung and colon cancer cells (Xie et al., 2016; Jo et al., 2017). Functional variation of ATG10 rs10514231 may be associated with malignant progression of HCC (Shen and Lin, 2019). PRKCD is a member of serine and threonine-specific protein kinase C family. It is not only a tumor suppressor, but also promotes the cell cycle. PRKCD was inhibited in HCC, and activation of PRKCD could decrease the viability of HCC cells (Nambotin et al., 2011). SLC11A1 is a phagocyte membrane protein expressed in monocytes (Bauler et al., 2017). As a pro-inflammatory factor, SLC11A1 is closely related to the occurrence and development of many inflammatory diseases and susceptibility to infectious diseases

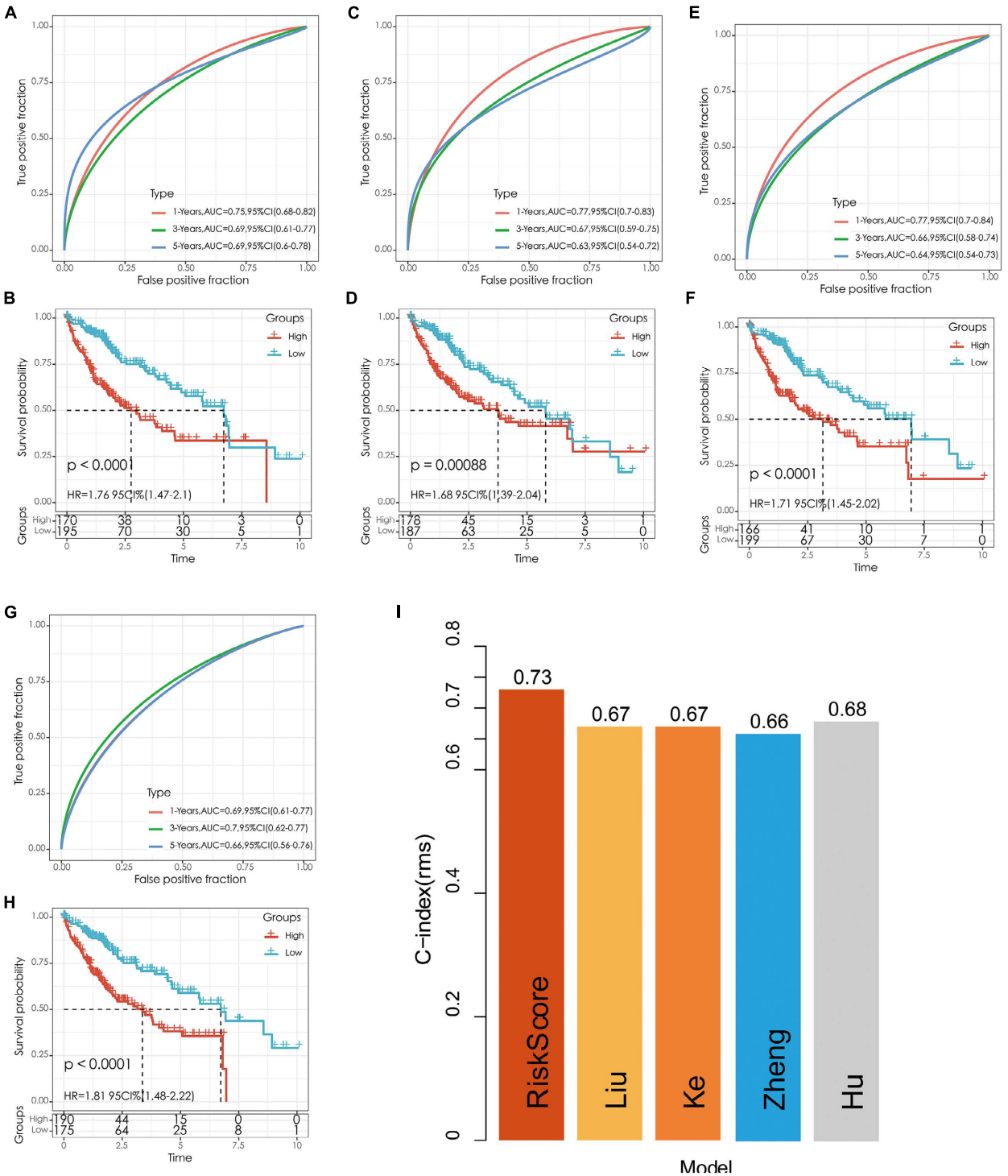


FIGURE 10 | (A,B) The ROC curve of Hu's risk model and the KM curve of high-risk and low-risk groups. **(C,D)** The ROC curve of Liu's risk model and the KM curve of high-risk and low-risk groups. **(E,F)** The ROC curve of Ke's risk model and the KM curve of high-risk and low-risk groups. **(G,H)** The ROC curve of Zheng's risk model and the KM curve of high-risk and low-risk groups. **(I)** C-index comparison of different prognostic risk models.

(Braliou et al., 2019; Xu and Yang, 2020). SLC11A1 has been shown to be associated with esophageal cancer susceptibility in South African populations (Zaahl et al., 2005). SPP1, also known as osteopontin, acts as a cytokine to promote the expression of interferon and IL-12 (Wu et al., 2019). SPP1 is thought to be an oncogene in tumors, and the high expression

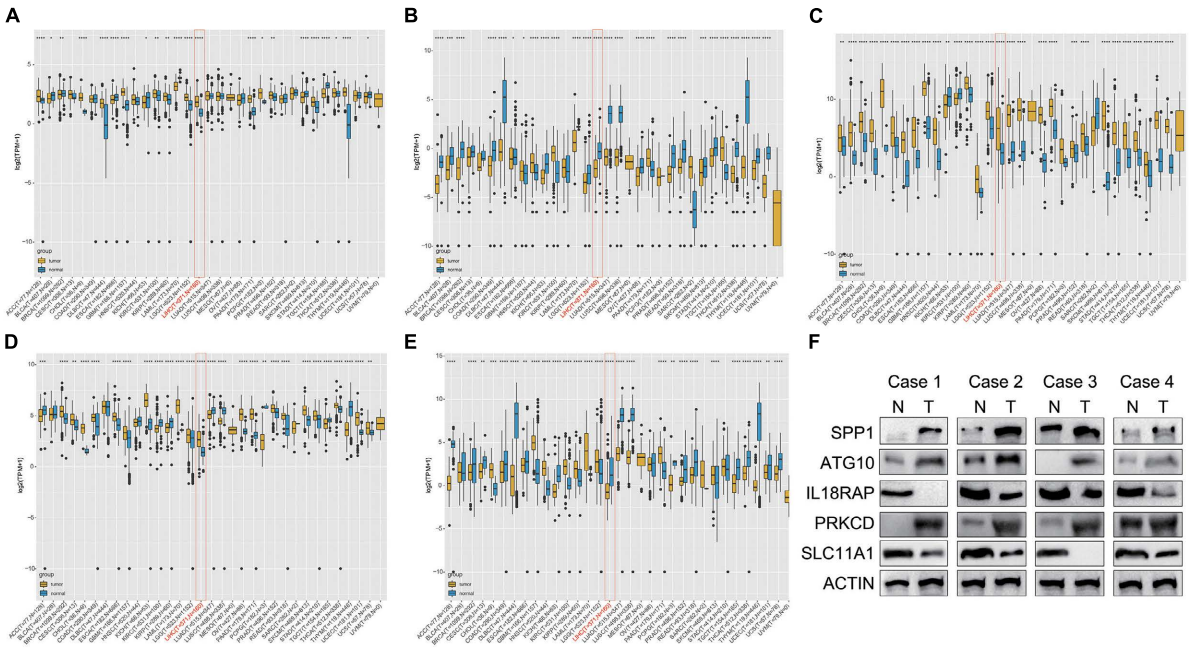


FIGURE 11 | Differential box diagram of 5-gene expression in pan-cancer. **(A)** ATG10, **(B)** IL18RAP, **(C)** SPP1, **(D)** PRKCD, **(E)** SLC11A1, and **(F)** the protein expression of the 5 genes in 4 pairs of LIHC and normal samples. *$P < 0.05$, **$P < 0.01$, ***$P < 0.005$, and ****$P < 0.001$.

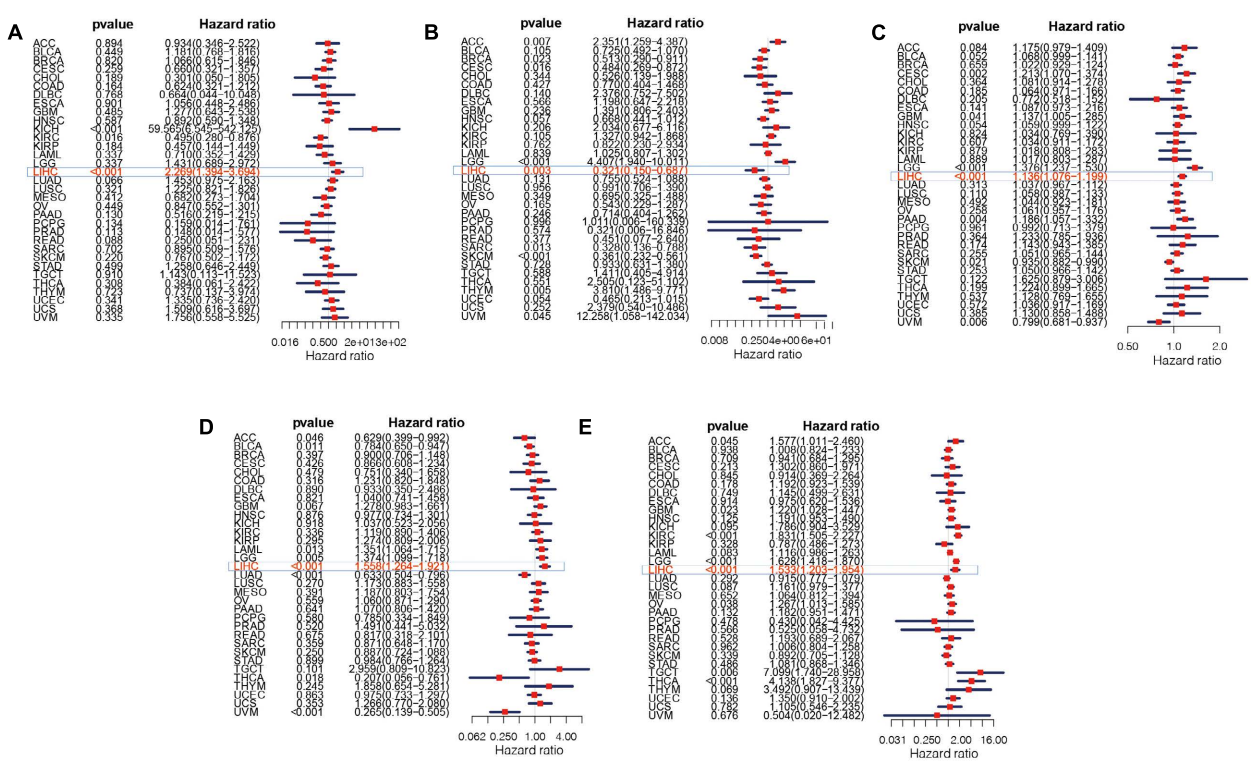


FIGURE 12 | Survival curve of five genes in pan-cancer. **(A)** ATG10, **(B)** IL18RAP, **(C)** SPP1, **(D)** PRKCD, and **(E)** SLC11A1.

of SPP1 in HCC suggests a poor prognosis (Chen et al., 2010). The polymorphism of SPP1 may be one of the genetic factors of HBV clearance and HCC (Shin et al., 2007). When identifying M1 macrophage-related modules related to the prognosis of thyroid cancer by WGCNA, it was found that the 4-gene signature including SLC11A1 and SPP1 could be

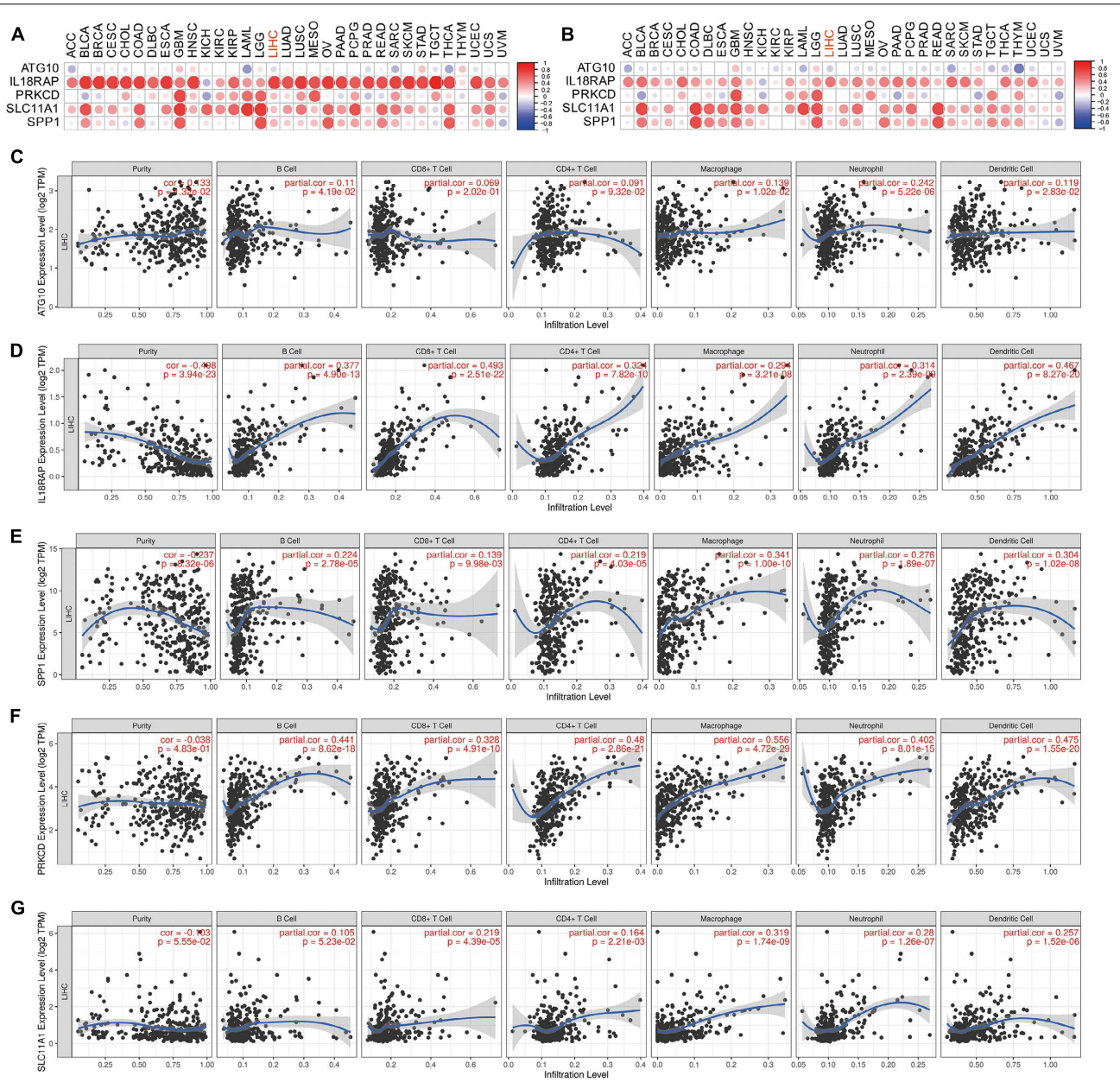


FIGURE 13 | The correlation between five genes and tumor microenvironment. **(A)** The correlation between genes and immune core. **(B)** The correlation between genes and StromalScore. **(C–G)** The correlation between ATG10, IL18RAP, SPP1, PRKCD, SLC11A1, and immune cells.

used to predict the prognosis of patients with thyroid cancer (Zhuang et al., 2020). By analyzing the correlation between five genes and tumor microenvironment, we found that SPP1, PRKCD, and SLC11A1 were highly positively correlated with macrophages. The results further confirmed our conjecture that SLC11A1 and SPP1 may promote the malignant progression of HCC by regulating M1 macrophages to M2 macrophages. The results of enrichment analysis showed that both IL18RAP and SLC11A1 were closely related to EMT signal pathway, which suggested that IL18RAP and SLC11A1 might promote the metastasis of HCC by regulating HCC cells. ATG10 gene is closely related to glycolysis signal pathway, and PRKCD gene is enriched in graft rejection-related pathway. Our study provides a new direction for the study of the mechanism of these five genes involved in the malignant progression of HCC. M1 macrophages can exert their anti-tumor and immune-enhancing effects by secreting inflammatory factors, chemokines, effector molecules, and TNF- molecules. Under the action of Lmur4 and IL-13, M1 macrophages can be polarized into M2 macrophages to inhibit immune response. The high infiltration of M1 macrophages is usually related to the increase of tumor survival (Edin et al., 2012; Jackute et al., 2018).

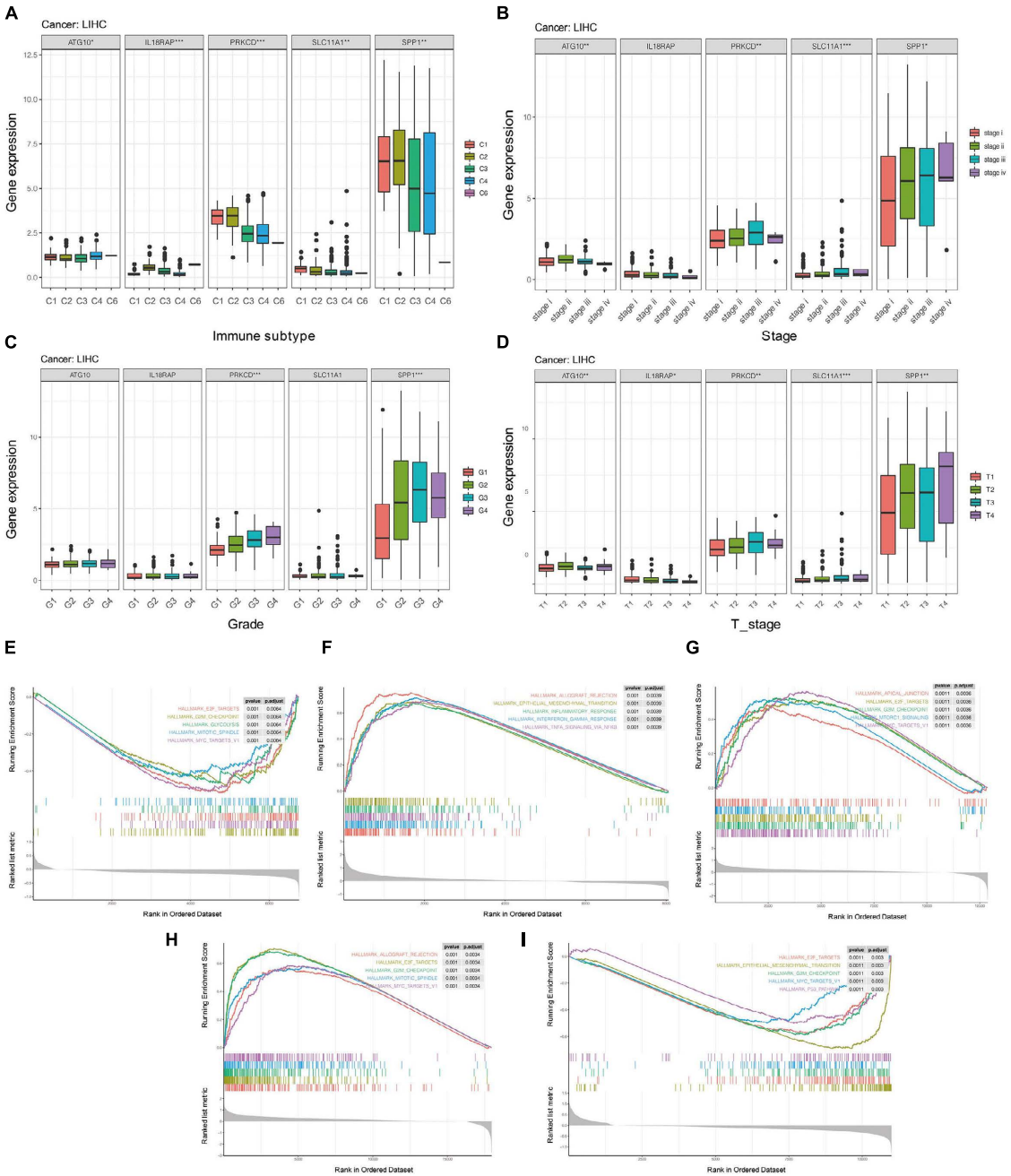


FIGURE 14 | The potential mechanism of five genes in HCC. **(A)** Correlation between five genes and immune subtypes of pan-cancer. **(B)** Correlation between five genes and clinical stage. **(C)** Correlation between five genes and clinical grade. **(D)** Correlation between five genes and T stage. **(E–I)** GSEA enrichment analysis of ATG10, IL18RAP, SPP1, PRKCD, and SLC11A1. *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$.

The 5-gene signature is robust in both internal and external data sets. Further analysis indicates that the high infiltration of Mast cell resting is associated with increased survival time in lung adenocarcinoma (Wang et al., 2020). The higher the CD8 level of T cells, the better the prognosis of HCC patients (Gabrielson et al., 2016). All the above studies confirmed that the high-risk score group with low Macrophages M1, Mast cell resting, and T cells CD8 scores had poorer survival outcomes, which confirmed our analysis. Additionally, the changes in immune cells provided insights into the molecular mechanism of the 5-gene signature in HCC.

However, first of all, as our study is a retrospective study based on public databases, lacking clinical validation of the prognostic value of the signature.

Secondly, based on the results of pathway enrichment analysis, prospective verification of the underlying biological mechanism of core genes is required.

And the population ethnicities in the TCGA database are mainly confined to White people and Black people, and extrapolating our findings to other ethnic groups needs to be substantiated. Finally, we will also explore the possibility of containing more variables to further improve prognostic performance.

CONCLUSION

In conclusion, this study systematically described the immune features of HCC based on 730 immune-related genes, identified the genes associated with prognosis, and constructed a 5-gene prognostic risk model. Thus, it can provide a good prognosis evaluation for the HCC samples. Our study further elucidates the relationship between immune microenvironment and prognosis of patients with HCC and provides new insights for immunotherapy of HCC.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The patients and their families in this study were fully informed, and informed consent was obtained from the participants. This study was approved by the Ethics Committee of Hainan General Hospital.

AUTHOR CONTRIBUTIONS

DX, YW, and JW designed the study, performed data analysis, and wrote the manuscript. YZ, ZL, and YC performed data collection. JZ supervised the manuscript. All authors read and approved the manuscript.

FUNDING

This research was supported by Natural Science Foundation of Hainan Province Project (820MS130) and Hainan Provincial Science and Technology Program (Clinical Medical Research Center: LCYX202101).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.686664/full#supplementary-material

Supplementary Figure 1 | (A) Venn diagram of differentially expressed immune genes. **(B–G)** The difference in expression of C6, EGR1, C8B, CFP, CXCL2, and CCL19 genes in tumor and paracancerous tissues, respectively. $^{**}P < 0.01$ and $^{****}P < 0.0001$.

Supplementary Figure 2 | The KM curve of the five genes.

Supplementary Figure 3 | (A) RiskScore, survival time, and 5-gene expression in GSE14520. **(B)** ROC curve and AUC of 5-gene signature classification. **(C)** KM survival curve distribution of 5-gene signature in independently validated data set GSE14520.

Supplementary Figure 4 | (A) RiskScore, survival time, and 5-gene expression in HCCDB18. **(B)** ROC curve and AUC of 5-gene signature classification. **(C)** KM survival curve distribution of 5-gene signature distribution in HCCDB18.

REFERENCES

Bauler, T. J., Starr, T., Nagy, T. A., Sridhar, S., Scott, D., Winkler, C. W., et al. (2017). *Salmonella* meningitis associated with monocyte infiltration in mice. *Am. J. Pathol.* 187, 187–199. doi: 10.1016/j.ajpath.2016.09.002

Braliou, G. G., Kontou, P. I., Boleti, H., and Bagos, P. G. (2019). Susceptibility to leishmaniasis is affected by host SLC11A1 gene polymorphisms: a systematic review and meta-analysis. *Parasitol. Res.* 118, 2329–2342. doi: 10.1007/s00436-019-06374-y

Chen, R. X., Xia, Y. H., Cui, J. F., Xue, T. C., and Ye, S. L. (2010). Osteopontin, a single marker for predicting the prognosis of patients with tumor-node-metastasis stage I hepatocellular carcinoma after surgical resection. *J. Gastroenterol. Hepatol.* 25, 1435–1442. doi: 10.1111/j.1440-1746.2010.06277.x

Cheung, H., Chen, N. J., Cao, Z., Ono, N., Ohashi, P. S., and Yeh, W. C. (2005). Accessory protein-like is essential for IL-18-mediated signaling. *J. Immunol.* 174, 5351–5357. doi: 10.4049/jimmunol.174.9.5351

Dai, W., Wu, J., Zhang, S., Shi, B., Xu, X., Wang, D., et al. (2017). Genes directly regulated by NF-kappaB in human hepatocellular carcinoma HepG2. *Int. J. Biochem. Cell Biol.* 89, 157–170. doi: 10.1016/j.biocel.2017.05.031

Desai, A., Sandhu, S., Lai, J. P., and Sandhu, D. S. (2019). Hepatocellular carcinoma in non-cirrhotic liver: a comprehensive review. *World J Hepatol* 11, 1–18. doi: 10.4254/wjh.v11.i1.1

Edin, S., Wikberg, M. L., Dahlin, A. M., Rutegard, J., Oberg, A., Oldenborg, P. A., et al. (2012). The distribution of macrophages with a M1 or M2 phenotype in relation to prognosis and the molecular characteristics of colorectal cancer. *PLoS One* 7:e47045. doi: 10.1371/journal.pone.0047045

El-Serag, H. B., and Rudolph, K. L. (2007). Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology* 132, 2557–2576. doi: 10.1053/j.gastro.2007.04.061

Forner, A., Reig, M., and Bruix, J. (2018). Hepatocellular carcinoma. *Lancet* 391, 1301–1314. doi: 10.1016/S0140-6736(18)30010-2

Gabrielson, A., Wu, Y., Wang, H., Jiang, J., Kallakury, B., Gatalica, Z., et al. (2016). Intratumoral CD3 and CD8 T-cell densities associated with relapse-free survival in HCC. *Cancer Immunol. Res.* 4, 419–430. doi: 10.1158/2326-6066.CIR-15-0110

Halegoua-De Marzio, D., and Hann, H. W. (2014). Prevention of hepatocellular carcinoma and its recurrence with anti-hepatitis B viral therapy. *Minerva Gastroenterol. Dietol.* 60, 191–200.

Hu, B., Yang, X. B., and Sang, X. T. (2020). Development of an immune-related prognostic index associated with hepatocellular carcinoma. *Aging (Albany NY)* 12, 5010–5030. doi: 10.18632/aging.102926

Jackute, J., Zemaitis, M., Pranys, D., Sitkauskiene, B., Miliauskas, S., Vaitkiene, S., et al. (2018). Distribution of M1 and M2 macrophages in tumor islets and stroma in relation to prognosis of non-small cell lung cancer. *BMC Immunol.* 19:3. doi: 10.1186/s12865-018-0241-4

Jo, Y. K., Kim, S. C., Park, I. J., Park, S. J., Jin, D. H., Hong, S. W., et al. (2012). Increased expression of ATG10 in colorectal cancer is associated with lymphovascular invasion and lymph node metastasis. *PLoS One* 7:e52705. doi: 10.1371/journal.pone.0052705

Jo, Y. K., Roh, S. A., Lee, H., Park, N. Y., Choi, E. S., Oh, J. H., et al. (2017). Polypyrimidine tract-binding protein 1-mediated down-regulation of ATG10 facilitates metastasis of colorectal cancer cells. *Cancer Lett.* 385, 21–27. doi: 10.1016/j.canlet.2016.11.002

Ke, K., Chen, G., Cai, Z., Huang, Y., Zhao, B., Wang, Y., et al. (2018). Evaluation and prediction of hepatocellular carcinoma prognosis based on molecular classification. *Cancer Manag. Res.* 10, 5291–5302. doi: 10.2147/CMAR.S178579

Kukreja, S. L., Löfberg, J., and Brenner, M. J. (2006). A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proc. Vol.* 39, 814–819. doi: 10.3182/20060329-3-AU-2901.00128

Li, B., Wang, Z., Wu, H., Xue, M., Lin, P., Wang, S., et al. (2018). Epigenetic regulation of CXCL12 plays a critical role in mediating tumor progression and the immune response in osteosarcoma. *Cancer Res.* 78, 3938–3953. doi: 10.1158/0008-5472.CAN-17-3801

Li, J. L., Lim, C. H., Tay, F. W., Goh, C. C., Devi, S., Malleret, B., et al. (2016). Neutrophils self-regulate immune complex-mediated cutaneous inflammation through CXCL2. *J. Invest Dermatol.* 136, 416–424. doi: 10.1038/JID.2015.410

Li, Y., He, X., Zhang, X., Xu, Y., Wu, Y., and Xu, X. (2021). Immune-related microRNA signature for predicting prognosis and the immune microenvironment in hepatocellular carcinoma. *Life Sci.* 265:118799. doi: 10.1016/j.lfs.2020.118799

Liu, G. M., Zeng, H. D., Zhang, C. Y., and Xu, J. W. (2019). Identification of a six-gene signature predicting overall survival for hepatocellular carcinoma. *Cancer Cell Int.* 19:138. doi: 10.1186/s12935-019-0858-2

Nambotin, S. B., Lefrancois, L., Sainsily, X., Berthillon, P., Kim, M., Wands, J. R., et al. (2011). Pharmacological inhibition of Frizzled-7 displays anti-tumor properties in hepatocellular carcinoma. *J. Hepatol.* 54, 288–299. doi: 10.1016/j.jhep.2010.06.033

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi: 10.1093/nar/gkv007

Sangro, B., Gomez-Martin, C., de la Mata, M., Inarrairaegui, M., Garralda, E., and Barrera, P. (2013). A clinical trial of CTLA-4 blockade with tremelimumab in patients with hepatocellular carcinoma and chronic hepatitis C. *J. Hepatol.* 59, 81–88. doi: 10.1016/j.jhep.2013.02.022

Shen, M., and Lin, L. (2019). Functional variants of autophagy-related genes are associated with the development of hepatocellular carcinoma. *Life Sci.* 235:116675. doi: 10.1016/j.lfs.2019.116675

Shin, H. D., Park, B. L., Cheong, H. S., Yoon, J. H., Kim, Y. J., and Lee, H. S. (2007). SPP1 polymorphisms associated with HBV clearance and HCC occurrence. *Int. J. Epidemiol.* 36, 1001–1008. doi: 10.1093/ije/dym093

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., et al. (2018). The immune landscape of cancer. *Immunity* 48, 812.e14–830.e14. doi: 10.1016/j.immuni.2018.03.023

Tian, M. X., Liu, W. R., Wang, H., Zhou, Y. F., Jin, L., Jiang, X. F., et al. (2019). Tissue-infiltrating lymphocytes signature predicts survival in patients with early/intermediate stage hepatocellular carcinoma. *BMC Med.* 17:106. doi: 10.1186/s12916-019-1341-6

Wang, C., Tang, X., Wang, J., and Xu, Y. (2020). Patterns of immune infiltration in lung adenocarcinoma revealed a prognosis-associated microRNA-mast cells network. *Hum. Cell* 33, 205–219. doi: 10.1007/s13577-019-00300-1

Wei, X., Yang, W., Zhang, F., Cheng, F., Rao, J., and Lu, L. (2020). PIGU promotes hepatocellular carcinoma progression through activating NF-kappaB pathway and increasing immune escape. *Life Sci.* 260:118476. doi: 10.1016/j.lfs.2020.118476

Westrich, J. A., Vermeer, D. W., Colbert, P. L., Spanos, W. C., and Pyeon, D. (2020). The multifarious roles of the chemokine CXCL14 in cancer progression and immune responses. *Mol. Carcinog* 59, 794–806. doi: 10.1002/mc.23188

Westrich, J. A., Vermeer, D. W., Silva, A., Bonney, S., Berger, J. N., Cicchini, L., et al. (2019). CXCL14 suppresses human papillomavirus-associated head and neck cancer through antigen-specific CD8(+) T-cell responses by upregulating MHC-I expression. *Oncogene* 38, 7166–7180. doi: 10.1038/s41388-019-0911-6

Wu, Q., Zhang, B., Sun, Y., Xu, R., Hu, X., Ren, S., et al. (2019). Identification of novel biomarkers and candidate small molecule drugs in non-small-cell lung cancer by integrated microarray analysis. *Onco Targets Ther.* 12, 3545–3563. doi: 10.2147/OTT.S198621

Xie, K., Liang, C., Li, Q., Yan, C., Wang, C., Gu, Y., et al. (2016). Role of ATG10 expression quantitative trait loci in non-small cell lung cancer survival. *Int. J. Cancer* 139, 1564–1573. doi: 10.1002/ijc.30205

Xu, J., and Yang, Y. (2020). Potential genes and pathways along with immune cells infiltration in the progression of atherosclerosis identified via microarray gene expression dataset re-analysis. *Vascular* 28, 643–654. doi: 10.1177/1708538120922700

Yamada, Y., Arai, T., Kojima, S., Sugawara, S., Kato, M., Okato, A., et al. (2018). Regulation of antitumor miR-144-5p targets oncogenes: direct regulation of syndecan-3 and its clinical significance. *Cancer Sci.* 109, 2919–2936. doi: 10.1111/cas.13722

Yarchoan, M., Agarwal, P., Villanueva, A., Rao, S., Dawson, L. A., Llovet, J. M., et al. (2019). Recent developments and therapeutic strategies against hepatocellular carcinoma. *Cancer Res.* 79, 4326–4330. doi: 10.1158/0008-5472.CAN-19-0803

Zaahl, M. G., Warnich, L., Victor, T. C., and Kotze, M. J. (2005). Association of functional polymorphisms of SLC11A1 with risk of esophageal cancer in the South African Colored population. *Cancer Genet. Cytogenet.* 159, 48–52. doi: 10.1016/j.cancergencyto.2004.09.017

Zhang, Y., Zhang, L., Xu, Y., Wu, X., Zhou, Y., and Mo, J. (2020). Immune-related long noncoding RNA signature for predicting survival and immune checkpoint blockade in hepatocellular carcinoma. *J. Cell Physiol.* 235, 9304–9316. doi: 10.1002/jcp.29730

Zhao, E., Chen, S., and Dang, Y. (2020). Development and external validation of a novel immune checkpoint-related gene signature for prediction of overall survival in hepatocellular carcinoma. *Front. Mol. Biosci.* 7:620765. doi: 10.3389/fmolb.2020.620765

Zheng, Y., Liu, Y., Zhao, S., Zheng, Z., Shen, C., An, L., et al. (2018). Large-scale analysis reveals a novel risk score to predict overall survival in hepatocellular carcinoma. *Cancer Manag. Res.* 10, 6079–6096. doi: 10.2147/CMAR.S181396

Zhu, J., Liu, C., Teng, X., Yin, J., Zheng, L., Wang, L., et al. (2016). Association of the interleukin-18 receptor 1 and interleukin-18 receptor accessory protein polymorphisms with the risk of esophageal cancer. *Biomed. Rep.* 4, 227–235. doi: 10.3892/br.2015.552

Zhuang, G., Zeng, Y., Tang, Q., He, Q., and Luo, G. (2020). Identifying M1 macrophage-related genes through a co-expression network to construct a four-gene risk-scoring model for predicting thyroid cancer prognosis. *Front. Genet.* 11:591079. doi: 10.3389/fgene.2020.591079

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Xu, Wang, Wu, Zhang, Liu, Chen and Zheng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 27 September 2021
doi: 10.3389/fcell.2021.755511

Large Scale Identification of Osteosarcoma Pathogenic Genes by Multiple Extreme Learning Machine

Zhipeng Zhao[1†], Jijun Shi[2†], Guang Zhao[3], Yanjun Gao[3], Zhigang Jiang[4] and Fusheng Yuan[3]*

[1] Department of Basic Medical Sciences, Taizhou University, Taizhou, China, [2] Department of Orthopedics, Songyuan Central Hospital, Songyuan, China, [3] Department of Orthopedics, The Fourth Affiliated Hospital of China Medical University, Shenyang, China, [4] Department of Hand Surgery, Changchun Central Hospital, Changchun, China

At present, the main treatment methods of osteosarcoma are chemotherapy and surgery. Its 5-year survival rate has not been significantly improved in the past decades. Osteosarcoma has extremely complex multigenomic heterogeneity and lacks universally applicable signal blocking targets. Osteosarcoma is often found in adolescents or children under the age of 20, so it is very important to explore its genetic pathogenic factors. We used known osteosarcoma-related genes and computer algorithms to find more osteosarcoma pathogenic genes, laying the foundation for the treatment of osteosarcoma immune microenvironment-related treatments, so as to carry out further explorations on these genes. It is a traditional method to identify osteosarcoma related genes by collecting clinical samples, measuring gene expressions by RNA-seq technology and comparing differentially expressed gene. The high cost and time consumption make it difficult to carry out research on a large scale. In this paper, we developed a novel method "RELM" which fuses multiple extreme learning machines (ELM) to identify osteosarcoma pathogenic genes. The AUC and AUPR of RELM are 0.91 and 0.88, respectively, in 10-cross validation, which illustrates the reliability of RELM.

Keywords: Index Term-osteosarcoma, pathogenic genes, fuses multiple extreme learning machine, machine learning, large scale identification

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Ningyi Zhang,
Harbin Institute of Technology, China
Hong Ju,
Heilongjiang Vocational College of Biology Science and Technology, China

***Correspondence:**
Fusheng Yuan
yuanfs07@mails.jlu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *09 August 2021*
Accepted: *02 September 2021*
Published: *27 September 2021*

Citation:
Zhao Z, Shi J, Zhao G, Gao Y, Jiang Z and Yuan F (2021) Large Scale Identification of Osteosarcoma Pathogenic Genes by Multiple Extreme Learning Machine. Front. Cell Dev. Biol. 9:755511. doi: 10.3389/fcell.2021.755511

INTRODUCTION

Osteosarcoma is the most common malignant bone tumor in clinic (Marko et al., 2016), which is mostly seen in children and adolescents. Although surgery combined with neoadjuvant chemotherapy significantly improves the 5-year survival rate of patients with local tumors (Yang et al., 2018), most patients with osteosarcoma will metastasize, and the 5-year survival rate of patients with metastatic osteosarcoma is only 20 ∼ 30% (Murakami et al., 2017). At present, osteosarcoma is still the second leading cause of cancer-related death in adolescents (Chen et al., 2020). Considering the complex intra - and inter tumor heterogeneity, a suitable specific target for osteosarcoma has not been found. However, based on previous studies on the heterogeneity of other tumors, the immune microenvironment may have relatively low heterogeneity and become a more appropriate direction of intervention (Koirala et al., 2016). Therefore, identifying genes

related to osteosarcoma immune microenvironment may provide a robust and effective target for clinical application (Mirabello et al., 2020).

According to the age at which osteosarcoma occurs suddenly increases with the onset of puberty, and its largest growth site is shown to be related to the rapid proliferation of bones, it indicates that osteosarcoma is significantly related to the rapid growth of bones (Ho et al., 2017). At the same time, exposure to alkylating agents may also promote the development of osteosarcoma (Zhang et al., 2021). In addition, radiotherapy is one of the few identified environmental risk factors for osteosarcoma. Studies have shown that increasing the radiation dose of primary cancer is linearly related to the risk of secondary osteosarcoma. Another study based on American adults also found that radiotherapy is significantly associated with an increased risk of osteosarcoma diagnosis in the future (Wu et al., 2012).

Whole-exome and whole-genome sequencing analysis of the germline DNA of patients with osteosarcoma showed that the prevalence of pathogenic variants in genes associated with known cancer susceptibility syndromes was higher than expected (Gianferante et al., 2017). Chromosomal abnormalities, pathogenic variants of tumor suppressor genes, transcription factors and growth factors, and abnormalities of WWOX and miRNA all play an important role in the occurrence and development of osteosarcoma (Lin et al., 2017). The frequency of osteosarcoma in individuals with mutations in the RB1 gene is higher than that in the population. Studies have shown that there is an interaction between primary inheritance and genes in the pathogenesis of the disease (Spritz, 2007). A 2016 study found that among individuals with pathogenic mutations in the germline tumor suppressor gene TP53, the cumulative incidence of osteosarcoma reached 5–11% (Mai et al., 2016). Transforming growth factor β (TGF-β) protein affects cell growth and metabolism, and the expression of TGF-β1 is significantly increased in highly malignant osteosarcoma. Insulin growth factors IGF-I and IGF-II can bind to the corresponding receptors to play a role, and they are overexpressed in osteosarcoma. The overexpression of CCN3 in osteosarcoma is related to its poor prognosis. Parathyroid hormone (PTH), parathyroid hormone related peptide (PTHrP) and parathyroid hormone receptor (PTHR1) have been shown to be related to the progression and metastasis of osteosarcoma (Berdiaki et al., 2010). Various molecular changes and genomes closely related to the occurrence and progress of osteosarcoma have been identified. These changes include gene amplification, deletion and germline mutation, overexpression and RTK activation, abnormal cell proliferation, metastasis, apoptosis, drug tolerance genes and miRNAs (Saraf et al., 2018). Osteosarcoma is characterized by complex and unbalanced karyotypes and abnormal gene expression profiles. Abnormalities of chromosome structure and value can be detected in most osteosarcoma (Isakoff et al., 2015). Common chromosome numerical abnormalities include germline mutation, deletion, polyploidy, aneuploidy, duplication and unbalanced ectopic errors (Morrow and Khanna, 2015). TP53 tumor gene and retinoblastoma tumor suppressor gene RB1 are the most prominent genes of germline mutation (Oliveira et al., 2005). They are the key detection sites of mitosis and the root cause of chromosome instability. Most osteosarcoma contains inactivation of both p53 and Rb pathways (Levine and Fleischli, 2000). In essence, the main causes of osteosarcoma are the inactivation of tumor suppressor gene expression and the abnormal doubling of oncogenes (Orr and Compton, 2013). Common oncogenes, such as avian cell homolog Myc, purine / pyrimidine exonuclease 1 (APEX1), action associated vascular endothelial growth factor A (VEGFA) and RecQ protein analog 4 (RecQL4). These amplified genes are closely related to the biological processes of osteosarcoma cell proliferation, growth and angiogenesis. Liu et al. (2019) identified 125 genes which are related to osteosarcoma and can be used to predict survival of osteosarcoma. Deng et al. (2021) used univariate, Lasso, and machine learning algorithm-iterative Lasso Cox regression analyses to predict survival of osteosarcoma by lncRNAs.

At present, there are two common biological methods for discovering disease-related genes. First, collect disease samples and health samples, respectively, conduct RNA-seq sequencing to obtain the expression of genes in different health states, and then obtain the genes significantly differentially expressed in disease and health populations through differential expression analysis (Zhao et al., 2021b). Second, through genome-wide association analysis, collect a large number of disease and healthy people, sequence the whole genome, and then compare the sequences to obtain sites with significant differences in mutation frequency (Peng and Zhao, 2020; Zhao et al., 2020c). However, both of them need a large number of samples to support in order to ensure the accuracy, which results in a large consumption of time and money (Bhakta and Tsukahara, 2020). With the continuous accumulation of biological data and the continuous improvement of calculation methods, bioinformatics experts find biological laws through calculation methods, and then infer more biological conclusions (Chen et al., 2019; Liu et al., 2020; Zhao et al., 2020a). The calculation methods have identified disease-related genes and drugs on a large scale (Tianyi et al., 2020; Zhao et al., 2020b). Although some conclusions are not completely accurate, they greatly reduce the scope of research and save time and money (Wu et al., 2021). Moreover, the models constructed by deep learning and machine learning can be used for reference by other research problems (Zhao et al., 2021a). Therefore, we developed a machine learning method to identify osteosarcoma-related genes in this paper. Using the idea of random forest for reference, we fused multiple Extreme Learning Machines (ELM) to build a model through the known osteosarcoma related genes to predict more genes potentially associated with osteosarcoma.

MATERIALS AND METHODS

Workflow

Firstly, we obtained 2,339 genes which are reported to be related to osteosarcoma in DisGeNET (Piñero et al., 2020). Then, we constructed gene interaction network based on these genes. More genes are included in this network since many genes can interact with these 2,339 genes. We extracted the features of this network by random walk and used Random Extreme Learning Machine

(RELM) to identify osteosarcoma-related genes. The way of constructing RELM is to build multiple ELM models and the output of each model is attached with weight, and the final result is obtained by voting. The whole workflow is shown in **Figure 1**.

Extreme Learning Machine

The calculation process of single hidden layer neural network is as follows:

1. The input value is multiplied by the weight value
2. Add bias value
3. Calculation of activation function
4. Repeat steps 1 to 3 for each layer
5. Calculate output value
6. Error back propagation
7. Repeat steps 1 to 6.

Extreme learning machines improves it by removing step 4 and replacing step 6 with a primary matrix inverse operation and removing step 7.

The process of ELM is to construct the formula (1):

$$f_L(x) = \sum_{i=1}^{L} \beta_i g_i\ (x) = \sum_{i=1}^{L} \beta_i g\ (w_i * x_j + b_j), j = 1, \ldots, N \quad (1)$$

L is the number of hidden units. N is the number of training samples. β_i is the weight between ith hidden layer and output. w_i is the weight between input and output. $g(x)$ is activation function. b is bias and x is the input. Since ELM only has one hidden layer, i is 1 in our model.

The calculation process of the extreme learning machine is very similar to the standard back-propagation neural network, but the weight matrix between the hidden layer and the output is a pseudo-inverse matrix. The above formula can be abbreviated as:

$$T = H\beta \quad (2)$$

$$H = \begin{bmatrix} g(w_1 * x_1 + b_1) & \ldots & g(w_L * x_L + b_L) \\ \vdots & & \vdots \\ g(w_1 * x_N + b_1) & \ldots & g(w_L * x_N + b_L) \end{bmatrix}_{N \times L} \quad (3)$$

m is the number of outputs; H is the hidden layer output matrix; T is the target matrix of the training set.

Random Extreme Learning Machine (RELM)

Extreme learning machines is a special artificial neural network with only one hidden layer, which causes its accuracy to be low. However, the calculation speed of ELM is extremely fast. Therefore, we can use this advantage to build multiple ELM models and use weighted voting to improve accuracy.

Random extreme learning machine draws on the idea of random forest (RF), regards ELM as a simple decision tree, and trains multiple ELMs to form an ELM forest to achieve the goal of improving accuracy.

The idea of RELM is to randomly extract the multi-dimensional features of genes, and then randomly extract the training set to form a simple ELM. Through repeated extraction with replacement, new ELMs are continuously trained. After

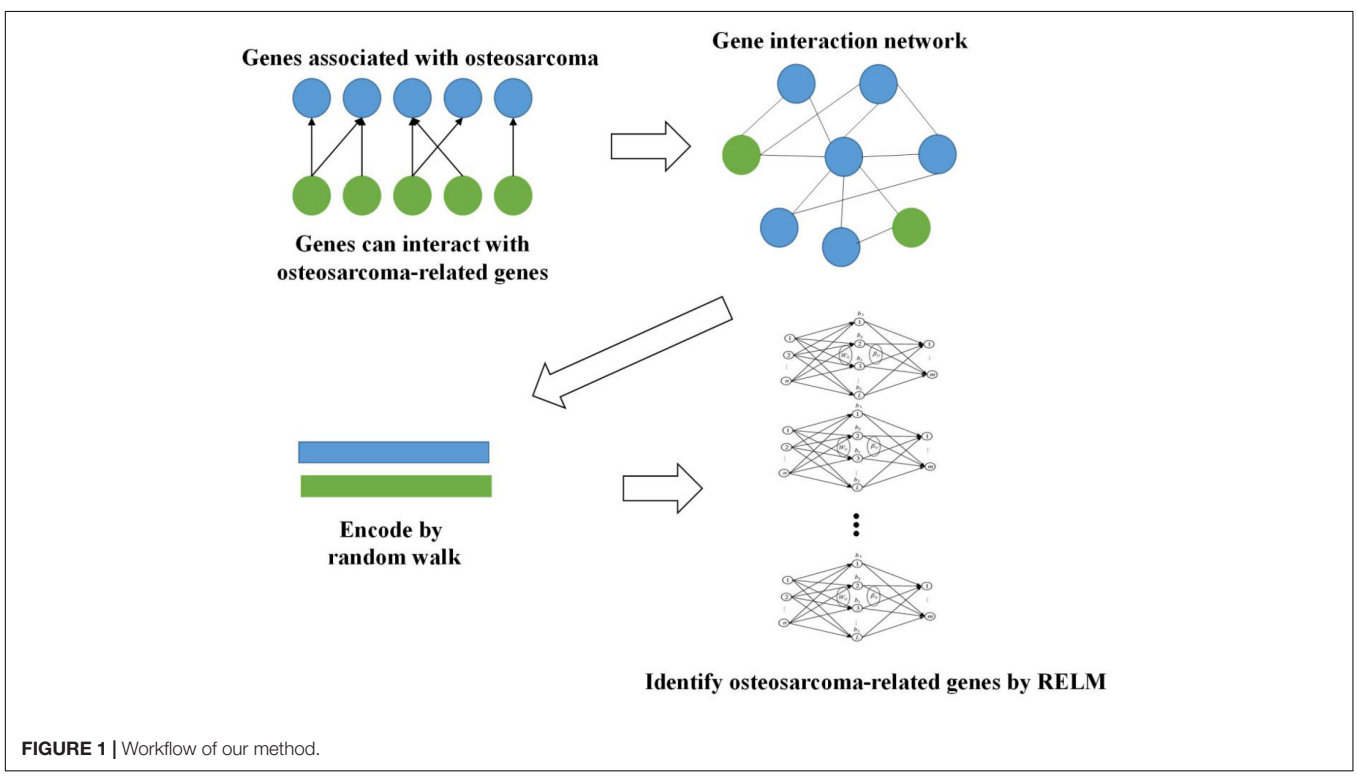


FIGURE 1 | Workflow of our method.

getting enough ELM models, the final result is obtained by weighting and averaging the output results of the 500 models.

The number of features for each ELM model is selected as (Zhao et al., 2017):

$$n = \sqrt{N} \quad (4)$$

N is the whole dimension of features. n is the number of features for each ELM model.

In the meanwhile, we randomly selected samples for each ELM model too. After each modeling, we will also put the sample back. We choose one-tenth of the samples for modeling each time.

RESULTS

Selection of Extreme Learning Machine Model Number

We should construct multiple ELM models to obtain RELM, but the number of ELM models is not sure. Therefore, we tried 10, 20, 50, 100, 200 ELM models and used 10-cross validation to obtain the final number.

The AUC curves of 10, 20, 50, 100, 200 ELM models are shown in **Figure 2**. The AUC values of these models are 0.66, 0.72, 0.82, 0.92, 0.92, respectively. The AUC of 100 models and 200 models are similar.

The PR curves of 10, 20, 50, 100, 200 ELM models are shown in **Figure 3**. The AUPR values of these models are 0.46, 0.54, 0.72, 0.88, 0.88, respectively. The AUPR of 100 models and 200 models are similar too. Therefore, we chose 100 ELM models to construct RELM.

Performance of Random Extreme Learning Machine

Because the unknown genes are far more than known osteosarcoma-related genes, we randomly selected negative samples to build RELM model. For each time, the number of negative samples is as same as positive samples. We repeated to select negative samples 5 times and did 10-cross validation for each time. The AUC and AUPR is shown as **Figure 4**.

The mean AUC is 0.889 and standard deviation is 0.009. The mean AUPR is 0.887 and standard deviation is 0.011.

In order to further explore the advantages of RELM, we compared RELM with ELM, RSVM, RANN. RSVM is to replace

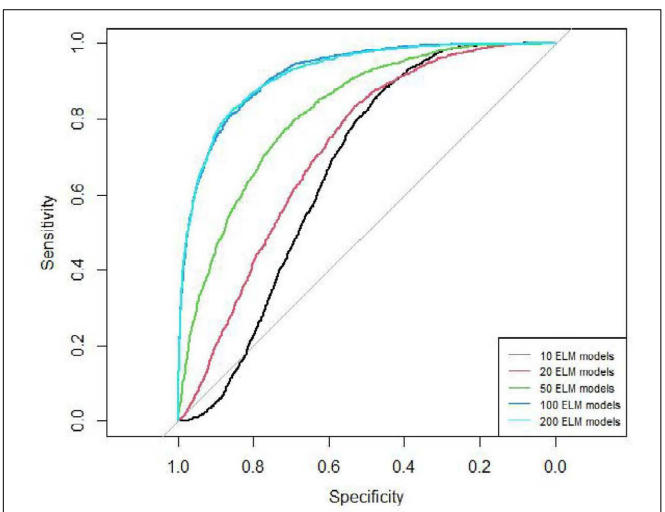


FIGURE 2 | ROC curves of different ELM model number.

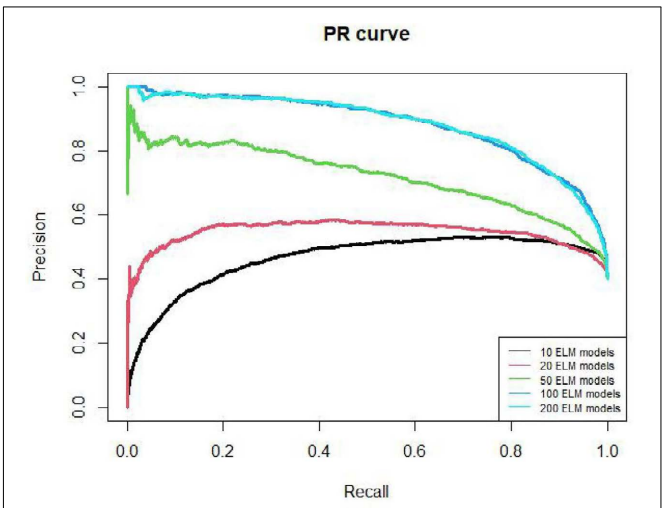


FIGURE 3 | PR curves of different ELM model number.

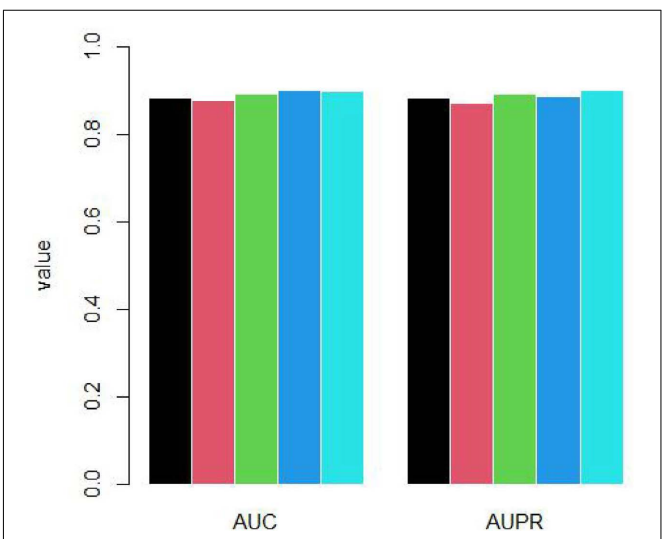


FIGURE 4 | Five times 10-cross validation by randomly sampling.

TABLE 1 | Comparison result.

Method	AUC	AUPR
RELM	0.889	0.887
ELM	0.761	0.684
RSVM	0.865	0.841
RANN	0.876	0.849

the ELM of RELM by SVM and RANN is to replace ELM of RELM with ANN. The experiments results are shown in **Table 1**.

As we can see in **Table 1**, RELM performed best among these method. SVM is more suitable for small sample modeling and ANN needs large sample set to build a precise model. Therefore, these two methods are not suitable for our case.

CONCLUSION

Whole-exome and whole-genome sequencing analysis of the germline DNA of patients with osteosarcoma showed that the prevalence of pathogenic variants in genes associated with known cancer susceptibility syndromes was higher than expected. Osteosarcoma is highly aggressive and progresses rapidly. In all age groups, as many as 25% of patients have metastasized at the time of diagnosis, so its early diagnosis is necessary for the long-term prognosis of patients. At present, the diagnosis of osteosarcoma is still based on the patient's clinical manifestations, imaging examinations and biopsy. Gene therapy includes tumor suppressor gene therapy, antisense gene therapy, suicide gene therapy and combined gene therapy. Although the research of gene therapy has made great progress and it has good therapeutic prospects, the clinical application of gene therapy still has a long way to go. In recent years, with continuous research on the key genes of osteosarcoma, its application value as a gene therapy target has gradually revealed.

To identify osteosarcoma-related genes in large scale, in this paper, we developed an ELM-based method for identifying osteosarcoma-related genes. 100 ELM models have been constructed to build a final RELM model. By constantly randomly selecting negative sets, we performed five times of 10-cross validation. The accuracy of RELM is stable and high in all experiments.

Overall, we purposed a reliable method for identifying osteosarcoma-related genes in large-scale. This method could help understand the pathogenesis of osteosarcoma and develop drug targets.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: gene-disease associations: https://www.disgenet.org/ and gene interaction: http://www.inetbio.org/humannet.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

ZZ, JS, and FY conceived and designed this study. ZZ, JS, GZ, YG, and ZJ analyzed the data. ZZ, JS, and GZ wrote the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by the Natural Science Foundation of China (NSFC) (No. 82060458).

REFERENCES

Berdiaki, A., Datsis, G. A., Nikitovic, D., Tsatsakis, A., Katonis, P., Karamanos, N. K., et al. (2010). Parathyroid hormone (PTH) peptides through the regulation of hyaluronan metabolism affect osteosarcoma cell migration. *IUBMB Life* 62, 377–386. doi: 10.1002/iub.320

Bhakta, S., and Tsukahara, T. (2020). Artificial RNA editing with ADAR for gene therapy. *Curr. Gene Ther.* 20, 44–54. doi: 10.2174/1566523220666200516170137

Chen, L., Wang, M., Lin, Z., Yao, M., Wang, W., Cheng, S., et al. (2020). Mild microwave ablation combined with HSP90 and TGF-β1 inhibitors enhances the therapeutic effect on osteosarcoma. *Mol. Med. Rep.* 22, 906–914. doi: 10.3892/mmr.2020.11173

Chen, X. G., Shi, W. W., and Deng, L. (2019). Prediction of disease comorbidity using hetesim scores based on multiple heterogeneous networks. *Curr. Gene Ther.* 19, 232–241. doi: 10.2174/1566523219666190917155959

Deng, Y., Yuan, W., Ren, E., Wu, Z., Zhang, G., and Xie, Q. (2021). A four-methylated LncRNA signature predicts survival of osteosarcoma patients based on machine learning. *Genomics* 113, 785–794. doi: 10.1016/j.ygeno.2020.10.010

Gianferante, D. M., Mirabello, L., and Savage, S. A. (2017). Germline and somatic genetics of osteosarcoma—connecting aetiology, biology and therapy. *Nat. Rev. Endocrinol.* 13, 480–491. doi: 10.1038/nrendo.2017.16

Ho, X. D., Phung, P., Le, V. Q., Nguyen, V. H., Reimann, E., Prans, E., et al. (2017). Whole transcriptome analysis identifies differentially regulated networks between osteosarcoma and normal bone samples. *Exp. Biol. Med.* 242, 1802–1811. doi: 10.1177/1535370217736512

Isakoff, M. S., Bielack, S. S., Meltzer, P., and Gorlick, R. (2015). Osteosarcoma: current treatment and a collaborative pathway to success. *J. Clin. Oncol.* 33, 3029–3035. doi: 10.1200/jco.2014.59.4895

Koirala, P., Roth, M. E., Gill, J., Piperdi, S., Chinai, J. M., Geller, D. S., et al. (2016). Immune infiltration and PD-L1 expression in the tumor microenvironment are prognostic in osteosarcoma. *Sci. Rep.* 6:30093.

Levine, R., and Fleischli, M. (2000). Inactivation of p53 and retinoblastoma family pathways in canine osteosarcoma cell lines. *Vet. Pathol.* 37, 54–61. doi: 10.1354/vp.37-1-54

Lin, W., Zhu, X., Yang, S., Chen, X., Wang, L., Huang, Z., et al. (2017). MicroRNA-203 inhibits proliferation and invasion, and promotes apoptosis of osteosarcoma cells by targeting Runt-related transcription factor 2. *Biomed. Pharmacother.* 91, 1075–1084. doi: 10.1016/j.biopha.2017.05.034

Liu, F., Xing, L., Zhang, X., and Zhang, X. (2019). A four-pseudogene classifier identified by machine learning serves as a novel prognostic marker for survival of osteosarcoma. *Genes* 10:414. doi: 10.3390/genes10060414

Liu, Y. B., Zhang, X., and Yang, L. (2020). Genetic engineering of AAV capsid gene for gene therapy application. *Curr. Gene Ther.* 20, 321–332. doi: 10.2174/1566523220666200930105521

Mai, P. L., Best, A. F., Peters, J. A., DeCastro, R. M., Khincha, P. P., Loud, J. T., et al. (2016). Risks of first and subsequent cancers among TP53 mutation carriers in the National cancer institute Li-Fraumeni syndrome cohort. *Cancer* 122, 3673–3681. doi: 10.1002/cncr.30248

Marko, T. A., Diessner, B. J., and Spector, L. G. (2016). Prevalence of metastasis at diagnosis of osteosarcoma: an international comparison. *Pediatr. Blood Cancer* 63, 1006–1011. doi: 10.1002/pbc.25963

Mirabello, L., Zhu, B., Koster, R., Karlins, E., Dean, M., Yeager, M., et al. (2020). Frequency of pathogenic germline variants in cancer-susceptibility genes in patients with osteosarcoma. *JAMA Oncol.* 6, 724–734. doi: 10.1001/jamaoncol.2020.0197

Morrow, J. J., and Khanna, C. (2015). Osteosarcoma genetics and epigenetics: emerging biology and candidate therapies. *Crit. Rev. Oncogen.* 20, 173–197. doi: 10.1615/critrevoncog.2015013713

Murakami, T., Igarashi, K., Kawaguchi, K., Kiyuna, T., Zhang, Y., Zhao, M., et al. (2017). Tumor-targeting *Salmonella typhimurium* A1-R regresses an osteosarcoma in a patient-derived xenograft model resistant to a molecular-targeting drug. *Oncotarget* 8, 8035–8042. doi: 10.18632/oncotarget.14040

Oliveira, A. M., Ross, J. S., and Fletcher, J. A. (2005). Tumor suppressor genes in breast cancer: the gatekeepers and the caretakers. *Pathol. Patterns Rev.* 124(Suppl. 1), S16–S28. doi: 10.1309/5XW3L8LU445QWGQR

Orr, B., and Compton, D. A. (2013). A double-edged sword: how oncogenes and tumor suppressor genes can contribute to chromosomal instability. *Front. Oncol.* 3:164. doi: 10.3389/fonc.2013.00164

Peng, J., and Zhao, T. (2020). Reduction in TOM1 expression exacerbates Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* 117, 3915–3916. doi: 10.1073/pnas.1917589117

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855. doi: 10.1093/nar/gkz1021

Saraf, A. J., Fenger, J. M., and Roberts, R. D. (2018). Osteosarcoma: accelerating progress makes for a hopeful future. *Front. Oncol.* 8:4. doi: 10.3389/fonc.2018.00004

Spritz, R. A. (2007). The genetics of generalized vitiligo and associated autoimmune diseases. *Pigment Cell Res.* 20, 271–278. doi: 10.1111/j.1600-0749.2007.00384.x

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinformat.* 22:bbaa044. doi: 10.1093/bib/bbaa044

Wu, L. C., Kleinerman, R. A., Curtis, R. E., Savage, S. A., and de González, A. B. (2012). Patterns of bone sarcomas as a second malignancy in relation to radiotherapy in adulthood and histologic type. *Cancer Epidemiol. Prevent. Biomark.* 21, 1993–1999. doi: 10.1158/1055-9965.epi-12-0810

Wu, Q., Nasoz, F., Jung, J., Bhattarai, B., Han, M. V., Greenes, R. A., et al. (2021). Machine learning approaches for the prediction of bone mineral density by using genomic and phenotypic data of 5130 older men. *Sci. Rep.* 11:4482. doi: 10.1038/s41598-021-83828-3

Yang, Y., Han, L., He, Z., Li, X., Yang, S., Yang, J., et al. (2018). Advances in limb salvage treatment of osteosarcoma. *J. Bone Oncol.* 10, 36–40.

Zhang, Z., Wu, X., Han, Q., and Huang, Z. (2021). Downregulation of long non-coding RNA UCA1 represses tumorigenesis and metastasis of osteosarcoma via miR-513b-5p/E2F5 axis. *Anti Cancer Drugs* 32, 602–613. doi: 10.1097/cad.0000000000001034

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 22:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020c). MRTFB regulates the expression of NOMO1 in colon. *Proc. Natl. Acad. Sci. U.S.A.* 117, 7568–7569. doi: 10.1073/pnas.2000499117

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021a). Prediction and collection of protein–metabolite interactions. *Brief. Bioinform.* 22:bbab014. doi: 10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Zhao, T., Zhang, N., Zhang, Y., Ren, J., Xu, P., Liu, Z., et al. (2017). A novel method to identify pre-microRNA in various species knowledge base on various species. *J. Biomed. Semant.* 8(Suppl. 1):30. doi: 10.1186/s13326-017-0143-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Zhao, Shi, Zhao, Gao, Jiang and Yuan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 29 September 2021
doi: 10.3389/fcell.2021.703415

Prognostic Value and Immunological Role of MORF4-Related Gene-Binding Protein in Human Cancers

Dongqi Chai[1†], Lilong Zhang[1†], Yongjun Guan[1†], Jingping Yuan[2], Man Li[1] and Weixing Wang[1*]*

[1] Department of Hepatobiliary and Laparoscopic Surgery, Renmin Hospital of Wuhan University, Wuhan, China, [2] Department of Pathology, Renmin Hospital of Wuhan University, Wuhan, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Ziheng Wang,
Affiliated Hospital of Nantong University, China
Shilin Xia,
Dalian Medical University, China

***Correspondence:**
Man Li
liman192@whu.edu.cn
Weixing Wang
sate.llite@163.com

[†] These authors have contributed equally to this work and share the first authorship

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *30 April 2021*
Accepted: *02 September 2021*
Published: *29 September 2021*

Citation:
Chai D, Zhang L, Guan Y, Yuan J, Li M and Wang W (2021) Prognostic Value and Immunological Role of MORF4-Related Gene-Binding Protein in Human Cancers. Front. Cell Dev. Biol. 9:703415. doi: 10.3389/fcell.2021.703415

MORF4-related gene-binding protein (MRGBP) is the subunit of the NuA4 histone acetyltransferase complex which is involved in transcriptional activation of select genes principally by acetylation of nucleosomal histones H4 and H2A. Much of the research indicated an oncogenic role of MRGBP in the development of cancers. However, it is still unknown the role MRGBP plays in human cancers, which deserves further exploration. In this research, the expression profile, prognostic value of MRGBP, and the relationship between MRGBP and immune infiltration were explored in 33 types of cancer. The differences in MRGBP expression in tumor and normal tissues were explored using data from The Cancer Genome Atlas, Gene Expression Omnibus and ONCOMINE. Analysis of the association between MRGBP and prognosis using Kaplan-Meier survival curve and COX analysis. The data of Tumor mutational burden (TMB), microsatellite instability (MSI) from TCGA. The relationship Between MRGBP expression and immunity was analyzed using the ESTIMATE algorithm and CIBERSORT. Furthermore, we explored MRGBP expression and the relationship between MRGBP expression and macrophage infiltration using immunohistochemical analysis in lower grade glioma (LGG). Our results revealed that MRGBP was highly expressed in most cancer tissues compared with normal tissues. Tumors with increased MRGBP expression had a high clinicopathologic stage and poor prognosis. The expression of MRGBP was closely related to the TMB, MSI. We also found a significant negative correlation between MRGBP expression and stromal scores and immune scores in various types of cancer. Furthermore, MRGBP expression was associated with a variety of immune cells including B cells, NK cells, T cells, and macrophages. LGG and LIHC was selected as representative cancer types for further study, the results of immunohistochemistry indicated that the protein levels of MRGBP were significantly elevated in tumor tissues. Moreover, our LIHC data analysis showed that patients with high MRGBP expression were associated with short survival rates and MRGBP was a risk factor to determine OS. Immunohistochemistry

also confirmed that M0 macrophage infiltration in the MRGBP-high group significantly increased. In conclusion, these results reveal that MRGBP can serve as a potential prognostic biomarker and it plays an important role in tumor immune infiltration in various tumors, especially in LGG and LIHC.

Keywords: MRGBP, cancer, prognosis, immune infiltration, macrophage

INTRODUCTION

MORF4-related gene-binding protein (MRGBP), also known as chromosome 20 open reading frame 20 (C20orf20), encodes a subunit of the TRRAP/TIP60 histone acetyltransferase complex and binds directly to MRG15 and MRGX proteins (Cai et al., 2003). The level of MRGBP expression was relatively low in normal tissues except testis. However, in most colorectal cancers, MRGBP expression is increased and it promotes colorectal carcinogenesis by providing an advantage in cell proliferation and cancer cell division (Carvalho et al., 2009; Yamaguchi et al., 2010, 2011). Furthermore, patients with higher TNM stage and T classification exhibit higher levels of MRGBP expression (Ding et al., 2017). It has also been shown that MRGBP accelerates androgen receptor-mediated transactivation and promotes the growth of androgen receptor-positive prostate cancer cells (Ito et al., 2018). Another tumor, cutaneous squamous cell carcinoma, in which MRGBP is thought to be overexpressed. When MRGBP is inhibited it induces apoptosis and reduces tumor growth (Watt et al., 2011). Also, it is associated with the progression of lung and cervical cancers (Scotto et al., 2008; Dai et al., 2019). MRGBP is involved in a variety of tumor processes, and its expression is significantly increased in a variety of tumors, suggesting that MRGBP may be used as a diagnostic biomarker and therapeutic target for tumors. However, the role of MRGBP in tumors has only been studied in a few tumors, and the link between MRGBP and various cancers deserves to be investigated.

Immune checkpoint proteins generate co-stimulatory or inhibitory signals in the immune response and modulate the host immune response under normal conditions. Recent studies have focused on the immune checkpoint PD-1 and its ligand PD-L1 signaling axis, which are highly expressed in tumors and can bind to PD-1 on the surface of T cells, limiting T cell activation and inducing a depleted state leading to tumor immune escape, Activation of T cell immune response by anti-CTLA-4 monoclonal antibody clears Treg cells (Lingel and Brunner-Weinzierl, 2019; Xin Yu et al., 2020). Recent studies have found that immunotherapy that blocks these receptors has shown good efficacy in treating several cancers such as melanoma, non-small cell lung cancer, kidney cancer, and bladder cancer (Topalian et al., 2012; Luke et al., 2017; Rizvi et al., 2018). The combination of targeted blockade of PD-1 and CTLA-4 is also a synergistic treatment option with good efficacy. The tumor microenvironment (TME) is composed of tumor cells, stromal cells, immune cells and their secreted factors. Immune cells play a dual role in inhibiting tumor progression and assisting tumor immune evasion (Junttila and de Sauvage, 2013). Immunotherapy for tumors is to target immunosuppressive cells in the immune microenvironment. However, there are a significant number of patients remain unresponsive to currently available immunotherapies which may be caused by the heterogeneity of the immune microenvironment.

In this study, our data were derived from a combination of public databases and web tools, and R language was used to analyze the association between MRGBP expression and patient prognosis and its potential role in tumor immunity.

MATERIALS AND METHODS

Expression Analysis of MORF4-Related Gene-Binding Protein

Data from the Genotype-Tissue Expression (GTEx) portal and the Cancer Cell Line Encyclopedia (CCLE) database were used to analyze MRGBP gene expression in normal tissues and tumor cell lines. Differential MRGBP expression between tumor and adjacent normal tissue was assessed by setting the threshold fold change to 1.5 and placing a p-value cutoff of 0.05 in the ONCOMINE database. The UCSC Xena was chosen to obtained RNA sequences associated with 33 cancers. We extracted and integrated MRGBP expression levels by Perl software for analysis of pan-cancer. The Wilcox. test method was used to analyze the differences in mRNA expression in different cancer types. The box plot is generated with the R-package "gg pubr." We combined the GTEx database and The Cancer Genome Atlas (TCGA) database to further explore the differences between cancer and normal tissue. Validation of differential expression of MRGBP between normal and tumor tissues in the GEO database using GSE16011. The box plot is implemented by the R software package "ggplot2."

Correlation of the MORF4-Related Gene-Binding Protein Expression to Clinical Phenotype and Prognosis

Next, we explored the association between MRGBP expression and clinical outcomes through survival information obtained from the TCGA database. The relationship between mRNA expression levels and patient survival rate was investigated by overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI). The relationship between MRGBP expression and survival outcome in different cancer types we performed by Kaplan-Meier, log-rank test and COX analysis, KM curves, and forest plots were plotted using the R program (v4.0.3). Subsequently, we used "limma" and "ggpubr" in R to perform a correlation analysis of clinicopathological characteristics. $p < 0.05$ were considered significant. The correlation of MRGBP levels to patients' OS

and PFS was further analyzed using Kaplan-Meier Plotter, which has data from 3 databases including Gene Expression Omnibus (GEO), European Genome-phenome Archive (EGA), and TCGA. Validation of the prognostic value of MRGBP in LGG using the Chinese Glioma Genome Altas (CGGA) and GEO dataset (GSE 4412). The survival curves were implemented by the R software package "survival" and "survminer."

Mutation Analysis of MORF4-Related Gene-Binding Protein

We perform mutation analysis of MRGBP in the CBio Cancer Genome Portal which is an open platform that provides visualization and analysis of gene copy number alterations and mutations in various cancers (Cerami et al., 2012). Catalog of Somatic Mutations in Cancer is a detailed and comprehensive resource for exploring the effect of somatic mutations in human cancer (Tate et al., 2019). Therefore, we investigated MRGBP mutations in various cancer types using COSMIC.

Relationship Between MORF4-Related Gene-Binding Protein Expression and Tumor Mutation Burden Tumor Microsatellite Instability and Some Specific Genes

We calculate the number of mutations in 33 types of cancer by counting the TMB and dividing it by the total length of the exons according to the Perl script. We used the somatic mutation data obtained from TCGA to identify the MSI scores for each tumor sample. We performed correlation analysis between cancer gene expression and TMB or MSI using the Spearman method, and the results were visualized by the "fmsb" in the R program.

Data for 5 MMR genes (MLH1, MSH2, MSH6, PMS2, and EPCAM), 5 methyltransferases genes (DNMT1, TRDMT1, DNMT3A, DNMT3B, DNMT3L), and 13 m6A-related genes were all downloaded from TCGA. We used Pearson correlation analysis to assess the relationship between MRGBP expression and the level of mutations in the genes mentioned above.

Relationship Between MORF4-Related Gene-Binding Protein Expression and Immunity

The percentage of stromal and immune cells in TME is expressed by the mesenchymal score and the immune score. ESTIMATE score uses available gene expression profiles to calculate the extent of mesenchymal or immune cell infiltration in the tumor, which indirectly represents tumor purity. In this research, the relationship between MRGBP levels and these two scores were calculated using the ESTIMATE algorithm in the R program we downloaded. CIBERSORT is a tool that can be used to calculate the relative fraction of specific immune cells based on gene expression data. Hence, we use it to evaluate the immune cell infiltration in each tumor. Moreover, the relationship between MRGBP and immune-related genes was analyzed using the "limma" and visualized using the "reshape2" and "RColorBreyer" in R software.

Enrichment Analysis of MORF4-Related Gene-Binding Protein-Related Gene

We acquired MRGBP binding proteins and constructed a PPI network using the STRING website (Szklarczyk et al., 2015). GEPIA2 was used to obtain the top 5 target genes associated with MRGBP and Pearson correlation analysis between MRGBP and target genes was then performed using the "Correlation Analysis" module in GEPIA2. Besides, the heatmap which showed the correlation between the top 5 target gene and 33 cancer types were obtained from TIMER2. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets were downloaded from the Gene Set Enrichment Analysis (GSEA) website. Both KEGG and GO enrichment pathway as well as functional annotation analysis of MRGBP were conducted using "clusterProfiler," "org.Hs.eg.db," "enrichplot," "DOSE," "colorspace," "stringi," and "ggplot2" R software package. The functional status of MRGBP in multiple cancer types was estimated by CancerSEA (filtered by correlation strength > 0.3 and false discovery rate < 0.05).

Immunohistochemical Analysis

We collected tumors and paracancerous tissues from LGG, COAD, and READ at the People's Hospital of Wuhan University, China, and 20 pairs of each tumor type were collected to verify the differential expression of MRGBP. In addition, we collected 10 tumor tissue specimens from LGG, totaling 30 samples, to verify the relationship between MRGBP expression and macrophage infiltration. The LIHC tumor tissues and adjacent normal tissues of 125 patients were collected from resected specimens at Tongji Hospital of Huazhong University of Science and Technology between 2012 and 2016. Tissue microarray plates containing 125 LIHC cases were constructed from paraffin-embedded LIHC tissues. Among them, 20 pairs of tissues without normal structure were excluded, and a total of 105 cases of LIHC were included. The association between MRGBP expression and macrophage infiltration in LIHC was validated using 8 MRGBP high and 8 low expressing tumor tissue samples, which were selected based on the results of tissue microarrays. For IHC, following deparaffinization, hydration, and epitope retrieval, the activity of endogenous peroxidase in the slices was inhibited for 15 min by 3% hydrogen peroxide. Then, slides were incubated overnight at 4°C with the primary antibody MRGBP (1:200, Santacruz, sc-32757) and CD68 (1:200, Abcam, ab955) in a humidity box, and subsequently placed in secondary antibody. Finally, the slides were visualized by diaminobenzidine and counterstained with hematoxylin. Immunohistochemical sections were observed by an Olympus BX63 microscope and the quantitative analysis of slides also used Image J software. Data were expressed as the mean ± standard deviation. Statistical measurements were performed using the SPSS 21.0 statistical software (SPSS Inc., Chicago, United States).

See **Supplementary Table 1** for details of the database sites used in this study. **Table 1** provided the full cancer type name corresponding to each abbreviation listed in the legend and the text. $P < 0.05$ in this article was considered statistically significant.

TABLE 1 | Pan-cancer data primary from TCGA database.

Abbreviation	Full name
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute myeloid leukemia
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular germ cell tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UVM	Uveal melanoma

RESULTS

The Differential Expression Level of MORF4-Related Gene-Binding Protein in Human Cancers

We studied 31 types of normal tissues from GTEx databases to identify MRGBP mRNA levels. Our results indicated that MRGBP expression was commonly elevated in the bone marrow, ovary, spleen and testis (**Supplementary Figure 1A**). Subsequently, we investigated MRGBP expression in 39 kinds of cancer cell lines. According to CCLE analysis results, MRGBP showed significantly different expression levels among these cell lines (**Supplementary Figure 1B**).

We next evaluated the differences in MRGBP expression between the tumor and normal tissues. First, we used the Oncomine database and found that it was highly expressed in most tumors compared to normal tissues (**Figure 1A**). However, MRGBP was less expressed in some tumors such as brain and central nervous system cancer, breast cancer, leukemia, and other cancer. Subsequently, we used the TCGA database to further verify. Collectively, MRGBP was significantly increased in most cancers including BLCA, BRCA, CECS, CHOL, COAD, ESCA, GBM, HNSC, KICH, KIRC,KIRP, LIHC, LUAD, LUSC, PRAD, READ, SARC, STAD, THCA, and UCEC compared with paired adjacent normal tissues (**Figure 1B**). These results indicated that MRGBP was commonly upregulated in most tumors and was statistically significant.

We combined data from GTEx and TCGA to analyze the differential expression of MRGBP in 27 normal and tumor tissues due to the small number of normal samples in the TCGA database. The results showed that MRGBP level in tumor tissues was significantly higher than normal tissues in most tumors, which was consistent with previous results. Notably, MRGBP expression was also shown to be increased in tumors such as ACC and LGG when adding normal tissue samples (**Figure 1C**).

Multifaceted Prognostic Value of MORF4-Related Gene-Binding Protein in Cancers

Subsequently, we analyzed the correlation between the level of MRGBP expression and prognostic value in pan-cancer by using TCGA databases. Cox regression analysis of MRGBP-related survival (OS, DSS, DFI, and PFI) indicated that high mRNA expression was a detrimental prognostic factor in ACC (OS: $p = 0.003$; DSS: $p = 0.005$; PFI: $p < 0.001$), HNSC (OS: $p = 0.043$) KIRC (OS: $p < 0.001$; DSS: $p < 0.001$; PFI: $p < 0.001$), KIRP (OS: $p = 0.009$; DSS: $p = 0.011$), LAML (OS: $p = 0.02$)LGG (OS: $p < 0.001$; DSS: $p < 0.001$; DFI: $p = 0.007$; PFI: $p < 0.001$), LIHC (OS: $p < 0.001$; DSS: $p = 0.011$; DFI: $p = 0.018$; PFI: $p = 0.006$), MESO (OS: $p < 0.001$; DSS: $p < 0.001$; DFI: $p = 0.036$; PFI: $p = 0.004$), PCPG (PFI: $p = 0.01$), PRAD (DSS: $p = 0.037$; DFI: $p < 0.001$; PFI: $p < 0.001$), SARC (OS: $p = 0.004$; DSS: $p = 0.002$, DFI: $p = 0.003$; PFI: $p = 0.013$), UCEC (OS: $p = 0.005$; DSS: $p < 0.001$; DFI: $p = 0.004$; PFI: $p < 0.001$), and UVM (OS: $p < 0.001$; DSS: $p < 0.001$; PFI: $p < 0.001$) (**Figures 2A**, **3A**, **4A**, **5A**).

The Kaplan-Meier cumulative curves displayed upregulated MRGBP expression was associated with poor prognosis in ACC (DFI: $p = 0.041$; PFI: $p = 0.009$), KIRC (OS: $p < 0.001$; DSS: $p < 0.001$; PFI: $p < 0.001$), KIRP (OS: $p = 0.018$; DSS: $p = 0.002$), LGG (OS: $p < 0.001$; DSS: $p < 0.001$;PFI: $p < 0.001$), LIHC (OS: $p = 0.004$; DSS: $p = 0.037$; DFI: $p = 0.006$; PFI = 0.006), LUAD (DFI: $p = 0.006$) MESO (OS: $p < 0.001$; DSS: $p = 0.003$; DFI: $p = 0.019$; PFI: $p = 0.002$), THCA (DSS: $p = 0.04$), PCPG (OS: $p = 0.09$; DSS: $p = 0.027$; PFI: $p = 0.01$), PRAD (DFI: $p = 0.002$; PFI: $p < 0.001$), SARC (DSS: $p = 0.048$, DFI: $p = 0.041$), UCEC (OS: $p = 0.002$; DSS: $p < 0.001$; DFI: $p = 0.003$; PFI: $p < 0.001$), and UVM (OS: $p < 0.001$; DSS: $p < 0.001$; PFI: $p < 0.001$) (**Figures 2–5**). However, high levels of MRGBP expression had longer survival time in patient with DLBC (OS: $p = 0.016$; DSS: $p = 0.04$) (**Figures 2B,3B**), while Cox regression analysis showed no statistical significance (**Figures 2A**, **3A**).

Next we used Kaplan-Meier Plotter, which has data from GEO, EGA, and TCGA, to further analyze MRGBP-related survival

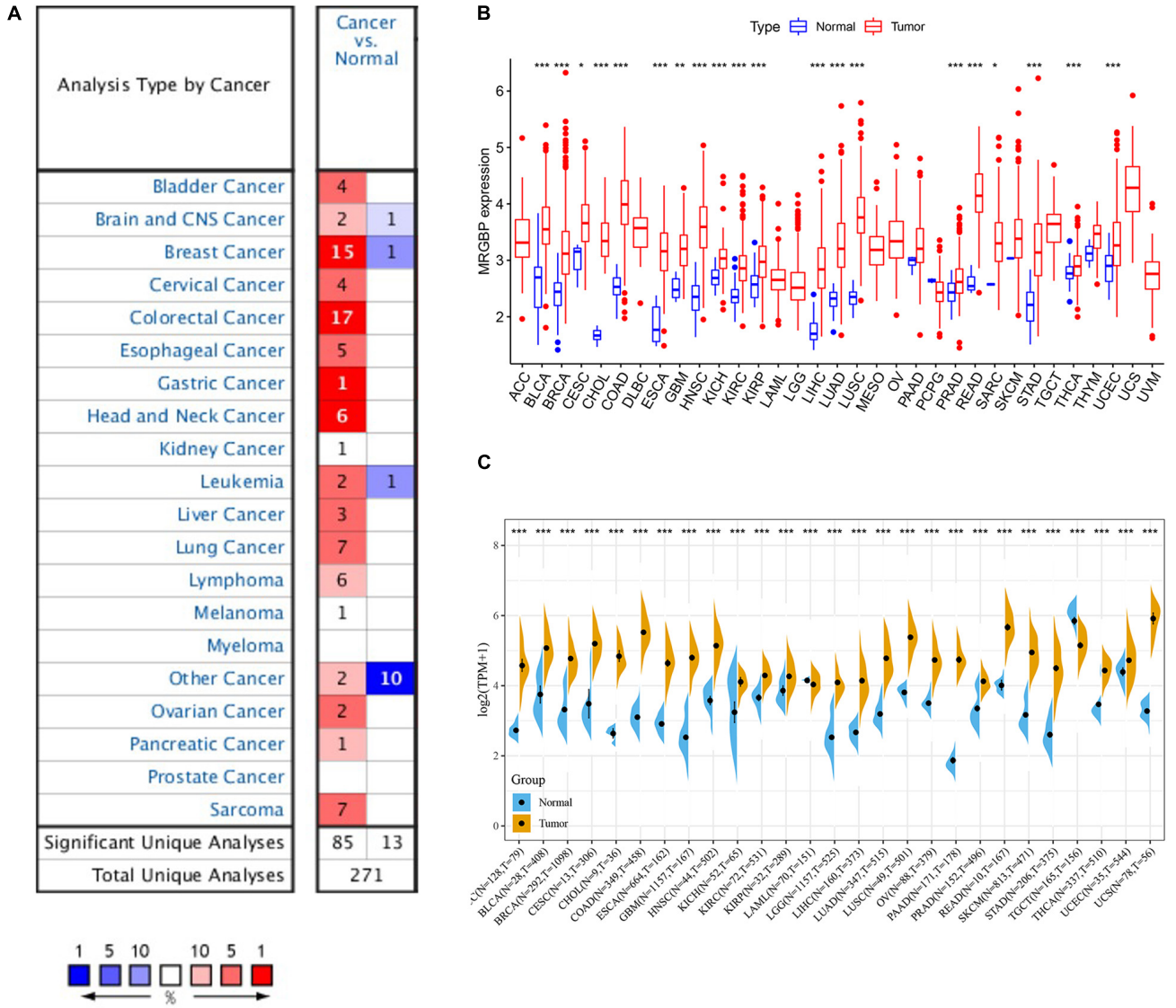


FIGURE 1 | mRNA expression of MRGBP gene in various cancers and its corresponding normal tissues. **(A)** Oncomine dataset. **(B)** TCGA dataset. **(C)** TCGA and GTEx dataset (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$).

including OS and RFS. The results shown in the **Supplementary Figure 2** were mostly consistent with those obtained above, and MRGBP was a deleterious prognostic factor in BLCA (OS: HR = 1.66, logrank $P = 0.0075$), HNSC (OS: HR = 1.58, logrank $P = 0.0015$), KIRC (OS: HR = 2.13, logrank $P < 0.001$), KIRP (OS: HR = 2.46, logrank $P = 0.0025$; RFS: HR = 2.17, logrank $P = 0.046$), LIHC (OS: HR = 2.1, logrank $P < 0.001$;RFS: HR = 1.87, logrank $P < 0.001$), LUAD (OS: HR = 1.52, logrank $P = 0.0071$), SARC (OS: HR = 1.99, logrank $P < 0.001$; RFS: HR = 1.99, logrank $P = 0.0062$), UCEC (OS: HR = 2.21, logrank $P < 0.001$; RFS: HR = 2.33, logrank $P = 0.0011$). However, for ESCA, a high expression of MRGBP was found to be beneficial for overall survival (OS: HR = 0.07, logrank $P < 0.001$) but detrimental for relapse free survival (RFS: HR = 8.35, logrank $P = 0.03$).

Association Between MORF4-Related Gene-Binding Protein Expression and Clinical Characteristics in Various Cancers

Next, we explored the association between the expression of mRNA and clinical feature in 33 cancers. Expression of MRGBP in different age groups showed significant differences in ESCA, KIRC, LGG, LIHC, LUAD, OV, SARC, STAD, and UCEC. Among them, less MRGBP expression was shown in patients over 65 years of age in ESCA, LIHC, and LUAD, and more expressed in other above cancers (**Supplementary Figure 3**). Moreover, MRGBP expression was significantly in different clinical stages in 10 cancers types including ACC, BRCA, COAD, ESCA, HNSC, KIRC, LIHC, LUSC, TGCT, and UVW

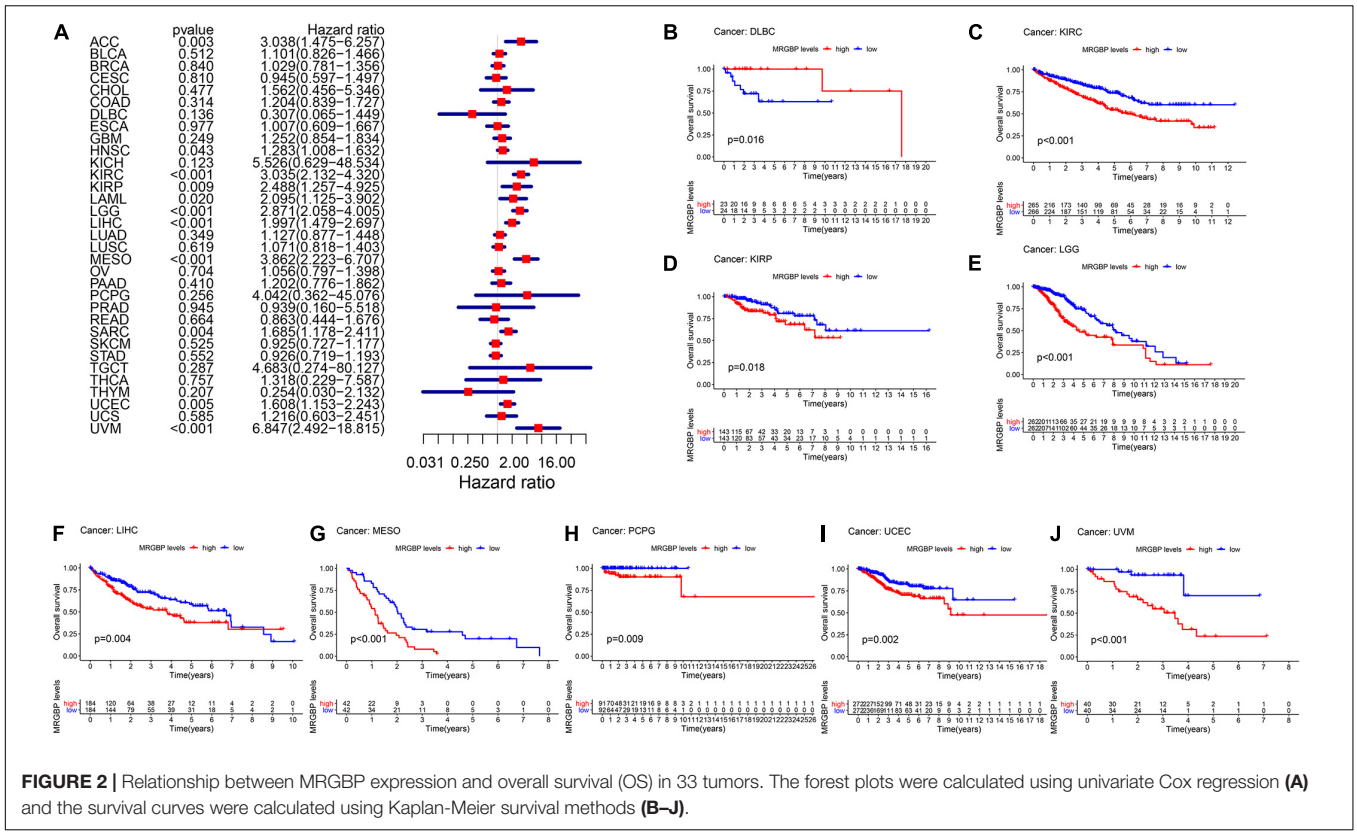


FIGURE 2 | Relationship between MRGBP expression and overall survival (OS) in 33 tumors. The forest plots were calculated using univariate Cox regression **(A)** and the survival curves were calculated using Kaplan-Meier survival methods **(B–J)**.

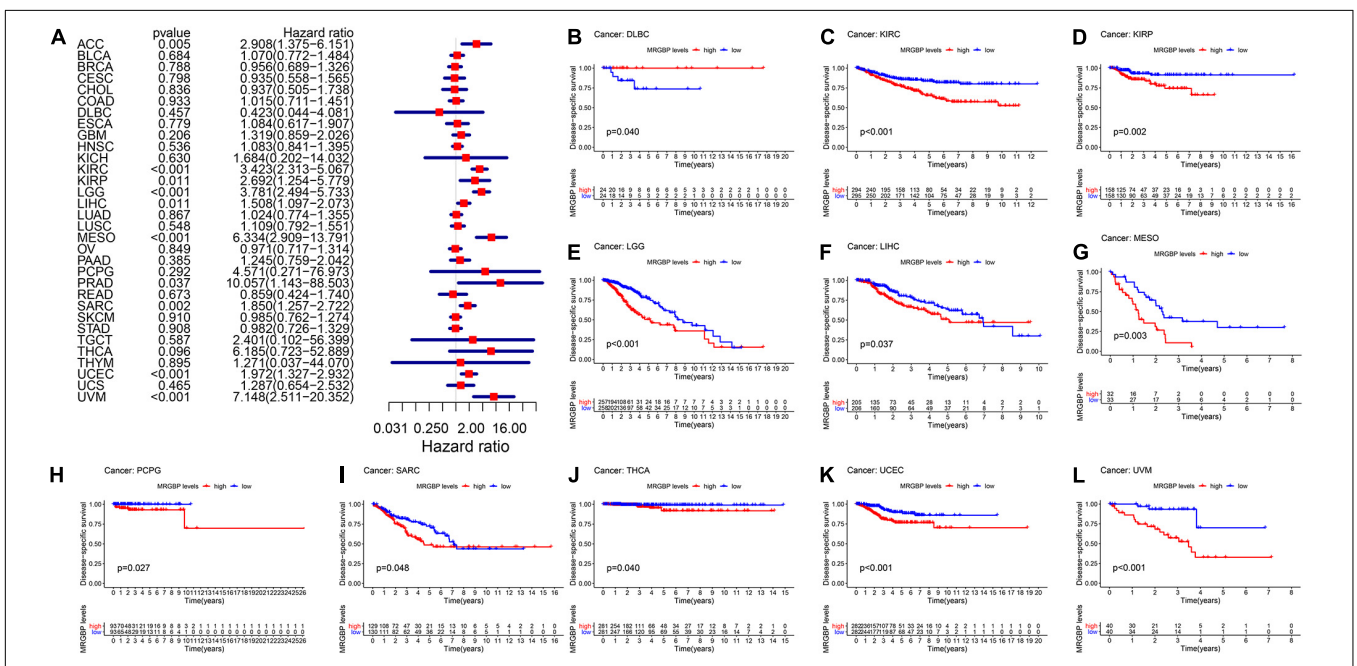


FIGURE 3 | Relationship between MRGBP expression and disease-specific survival (DSS) in 33 tumors. The forest plots were calculated using univariate Cox regression **(A)** and the survival curves were calculated using Kaplan-Meier survival methods **(B–L)**.

(**Supplementary Figure 4**). It appears that MRGBP expression is higher in patients with stage III and IV than in patients with stage I and II. In addition, the expression of MRGBP in different molecular subtypes of COAD, GBM, HNSC, KIRP, LGG, LUSC, OV, PCPG, PRAD, READ, SKCM, STAD ACC, BRCA, and UCEC was significantly different (**Supplementary Figure 5**).

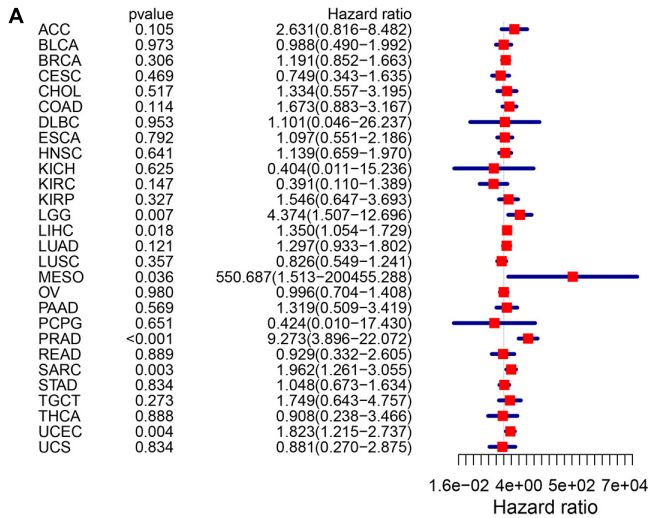


FIGURE 4 | Relationship between MRGBP expression and disease-free interval (DFI) in 33 tumors. The forest plots were calculated using univariate Cox regression **(A)** and the survival curves were calculated using Kaplan-Meier survival methods **(B–H)**.

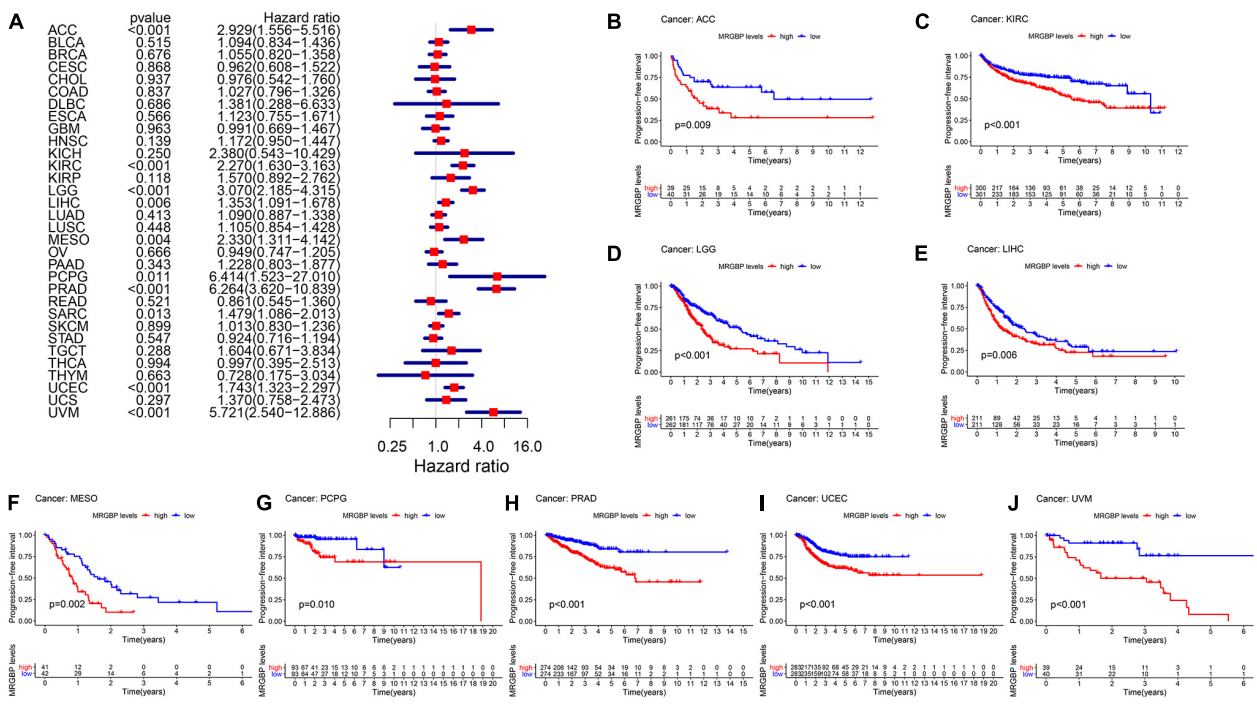


FIGURE 5 | Relationship between MRGBP expression and progression-free interval (PFI) in 33 tumors. The forest plots were calculated using univariate Cox regression **(A)** and the survival curves were calculated using Kaplan-Meier survival methods **(B–J)**.

MORF4-Related Gene-Binding Protein Copy Number Alterations and Mutations in Multiple Cancers

We investigated the copy number alterations of MRGBP using the cBioPortal database. In all cancers, we found that the mutation frequency was relatively higher in colorectal adenocarcinoma and ovarian serous cystadenocarcinoma compared with other cancers. Amplification was the highest type of alteration, followed by mutation and deep deletion (**Figures 6A,B**). We identified 28 missense sites and 1 truncation site in MRGBP situated between amino acids 0 and 204 through the cBioPortal database (**Figure 6C**). Also, we obtained detailed and comprehensive mutational information on MRGBP in multiple tumors including missense mutations, synonymous mutations, shift deletions, and other mutation types using COSMIC (**Supplementary Figure 6**). Missense mutations were evident in most cancers such a large intestine cancer, lung cancer, and skin cancer and synergistic mutations were relatively uncommon. Other tumors with smaller sample sizes of mutations also showed different types of mutations. Among MRGBP coding chain mutations, C > T and G > A types were the most common, and other types of mutations were rarely found in different cancers.

Correlation of MORF4-Related Gene-Binding Protein Expression With Tumor Mutation Burden, Tumor Microsatellite Instability, and Some Specific Genes in Various Cancers

TMB is the number of somatic mutations in the tumor genome after removal of germline mutations, and high TMB shows better immunotherapeutic effects (Choucair et al., 2020). Defective DNA mismatch repair causes errors in tumor cell DNA replication leading to high MSI, which is associated with the effectiveness of immunotherapy in patients (van Velzen et al., 2020). As shown in **Figure 7A**, MRGBP expression significantly related to TMB in UCEC, STAD, SKCM, SARC, PRAD, PAAD, OV, MESO, LUSC, LUAD, LGG, KIRC, HNSC, THYM, COAD, CESC, BRCA, and ACC ($P < 0.05$) and most of them were positively correlated with TMB except THYM, COAD, CESC, BRCA, and ACC. We further verified whether the expression of MRGBP was associated with MSI in various cancers. In general, mRNA level was positively associated with MSI in 15 types of cancer including UCEC, THCA, STAD, SKCM, SARC, PRAD, PAAD, LUSC, LUAD, LIHC, KICH, HNSC, THYM, DLBC, BRCA, and BLCA. Only a few tumors such as LAML and COAD are negatively correlated with MRGBP expression (**Figure 7B**).

Mismatch repair (MMR) is one of the most important DNA repair processes and loss of function of DNA mismatch repair genes can cause the accumulation of mismatches during DNA replication, leading to tumorigenesis (Baretti and Le, 2018). Therefore, we explored the correlations between the expression of five MMR genes (EPCAM, PMS2, MSH6, MSH2, and MLH1) and MRGBP levels. Our results showed that the level of MMR gene was positively related to MRGBP expression in most tumors (**Figure 7C**). Methylation of DNA is an important epigenetic modification and abnormal methylation can lead to tumor development and somatic cell mutations. As shown in **Figure 7D**, in most tumors, MRGBP expression levels were positively associated with the expression of DNMT1, TRDMT1, DNMT3A, DNMT3B, and DNMT3L, especially DNMT3B. N6-methyladenosine (m6A), one of the most common RNA modifications, plays an important role in almost all important biological processes, including tumorigenesis and tumor progression. The modification level of m6A is dynamically regulated by a methyltransferase (writers), binding protein (readers), and demethylase (erasers). The results indicated that MRGBP expression was significantly related to the expression of these genes (**Figure 7E**).

Association Between MORF4-Related Gene-Binding Protein Expression and Tumor Microenvironment in Human Cancers

Immune and stromal cells are the two major types of non-tumor components in the tumor microenvironment (TME). The degree of infiltration of them in TME has been reported to be of great value in the diagnosis of tumors and the response to immunotherapy (Wu and Dai, 2017). We then studied the relationship between the expression of MRGBP and TME in 33 cancers using the ESTIMATE algorithm. Our results indicated that MRGBP expression was significantly and negatively correlated with stromal scores and immune scores in COAD, GBM, HNSC, KIRP, LUAD, LUSC, PAAD, STAD, and UCEC. It suggested that the expression level of MRGBP was increased while the content of stromal or immune cells was decreased. We showed the six tumors with the highest correlation with stromal scores and immune scores (**Figure 8**) and the others are shown in **Supplementary Figure 7**.

Association Between MORF4-Related Gene-Binding Protein Expression and Levels of Tumor Immune Cell Infiltration in Different Types of Cancer

Subsequently, the correlation between MRGBP expression levels and the levels of 21 immune cell infiltrates in human cancers was studied. The results showed a significant correlation between the level of immune cell infiltration and MRGBP expression in almost all tumors. Among them, BRCA ($n = 12$), HNSC ($n = 5$), KIRC ($n = 8$), LUAD ($n = 7$), PRAD ($n = 5$), THCA ($n = 5$), THYM ($n = 5$), and UCEC ($n = 6$), which had the highest correlation coefficients between their immune cell infiltration levels and MRGBP levels, were used to further analyze (**Figure 9**).

Our results showed a negative correlation between MRGBP expression and the level of infiltrated naive B cells in KIRC, BRCA, and THCA but in terms of memory B cells, activated and resting NK cells, their infiltration levels were positively related to the MRGBP expression except for the infiltration of activated NK cells in THYM. For infiltrating resting and activated dendritic cells and monocytes, their levels were negatively correlated

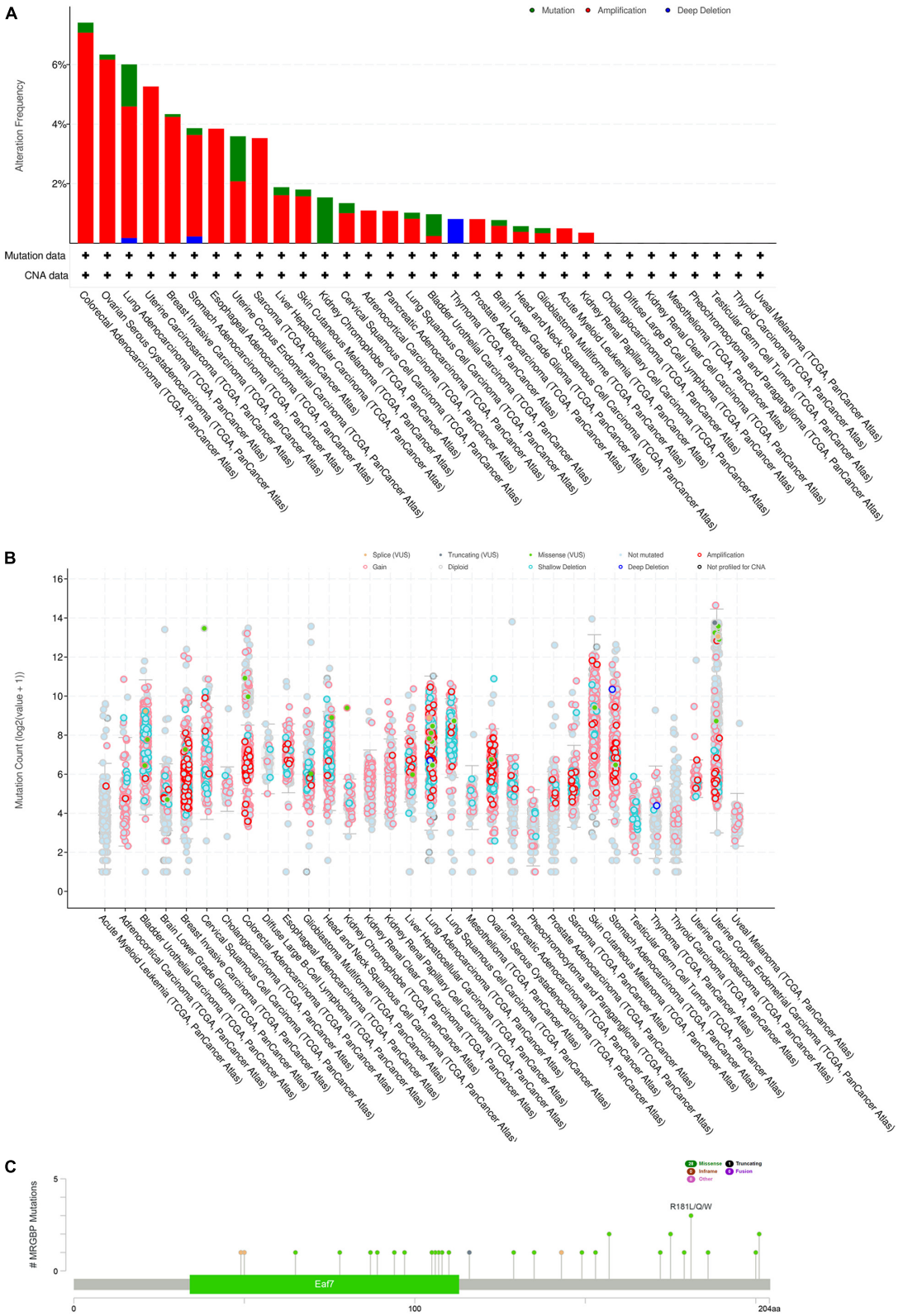


FIGURE 6 | MRGBP Copy Number Alterations and Mutations. **(A)** MRGBP mutation level from the cBioPortal database. **(B)** MRGBP mutation frequency in multiple TCGA pan-cancer studies according to the cBioPortal database. **(C)** Mutation diagram of MRGBP in different cancer types across protein domains from the cBioPortal database.

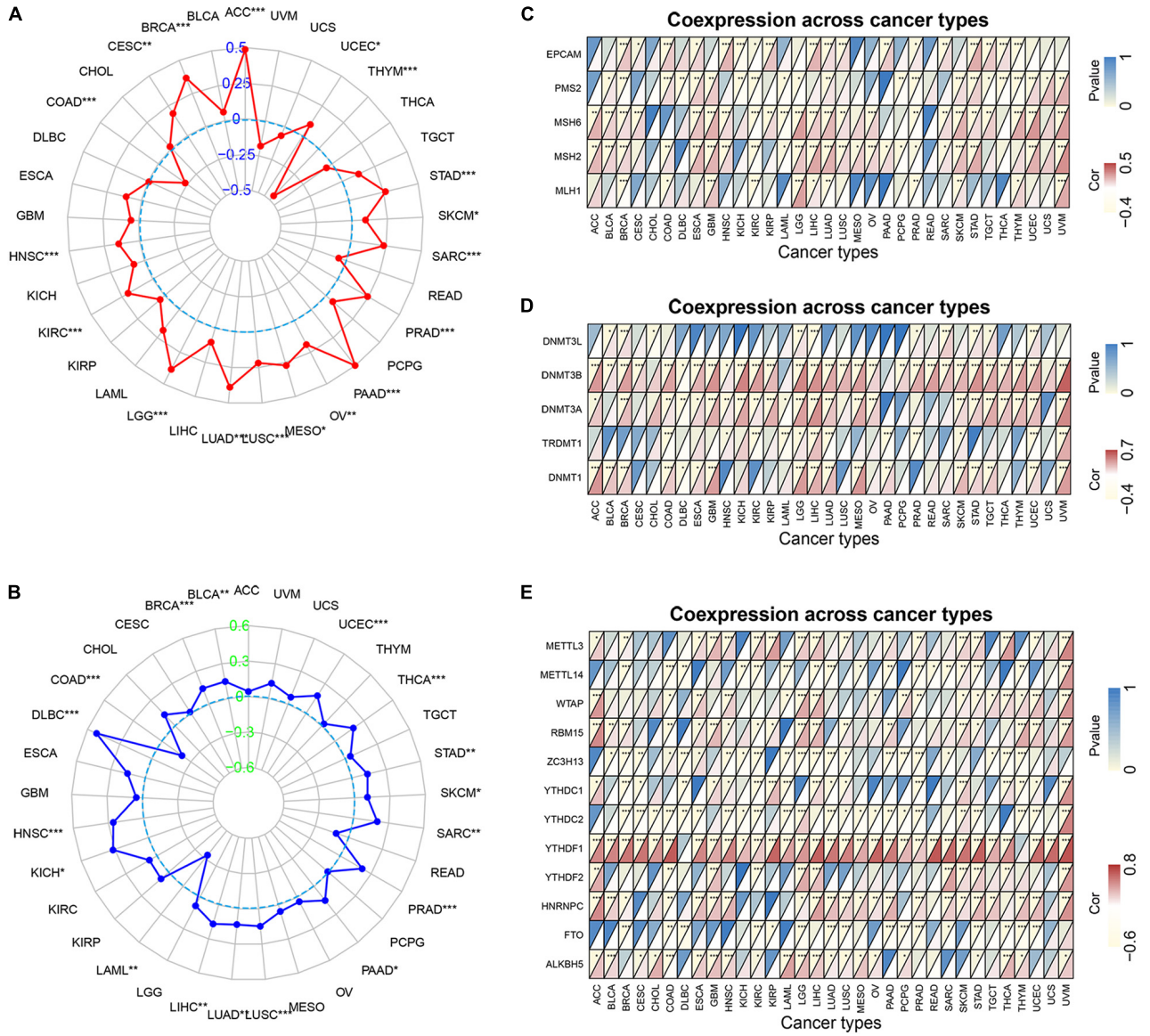


FIGURE 7 | Relationship between MRGBP expression and tumor mutational burden **(A)**, microsatellite instability **(B)**, MMR genes **(C)**, DNA methyltransferases genes **(D)** and mRNA m6A related genes **(E)** in various tumors in TCGA database. The lower triangle in each tile indicates coefficients calculated by Pearson's correlation test, and the upper triangle indicates $\log_{10}$-transformed *P*-value (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$).

with MRGBP expression. Moreover, MRGBP expression levels appeared to be correlated with macrophage polarization. As shown in **Figure 9**, MRGBP expression was positively correlated with infiltrating M1 macrophages but negatively correlated with M2 macrophages. Also, there were different correlations between MRGBP expression levels and different infiltrating T cell subpopulations. The results showed a positive correlation between MRGBP expression and the levels of infiltrating activated memory CD4 + T cells, CD8 + T cells, follicular helper T cells, and regulatory T cells (except in UCEC); however, it was negatively related to the levels of resting memory CD4 T cells. MRGBP expression was negatively correlated with the levels of infiltrating resting mast cells in BRCA, HNSC, KIRC, LUAD, and THYM, but positively associated with infiltrating activated mast cells in HNSC. The relationship between MRGBP expression and infiltrating immune cells in other types of cancer is included in **Supplementary Figure 8**.

Co-expression of Immune-Related Genes With MORF4-Related Gene-Binding Protein in Pan-Cancers

Co-expression analyses were used to explore the relationship between MRGBP expression and immune-related genes including immune checkpoints, MHC, chemokine, and chemokine receptor proteins genes in 33 tumors. As is shown in

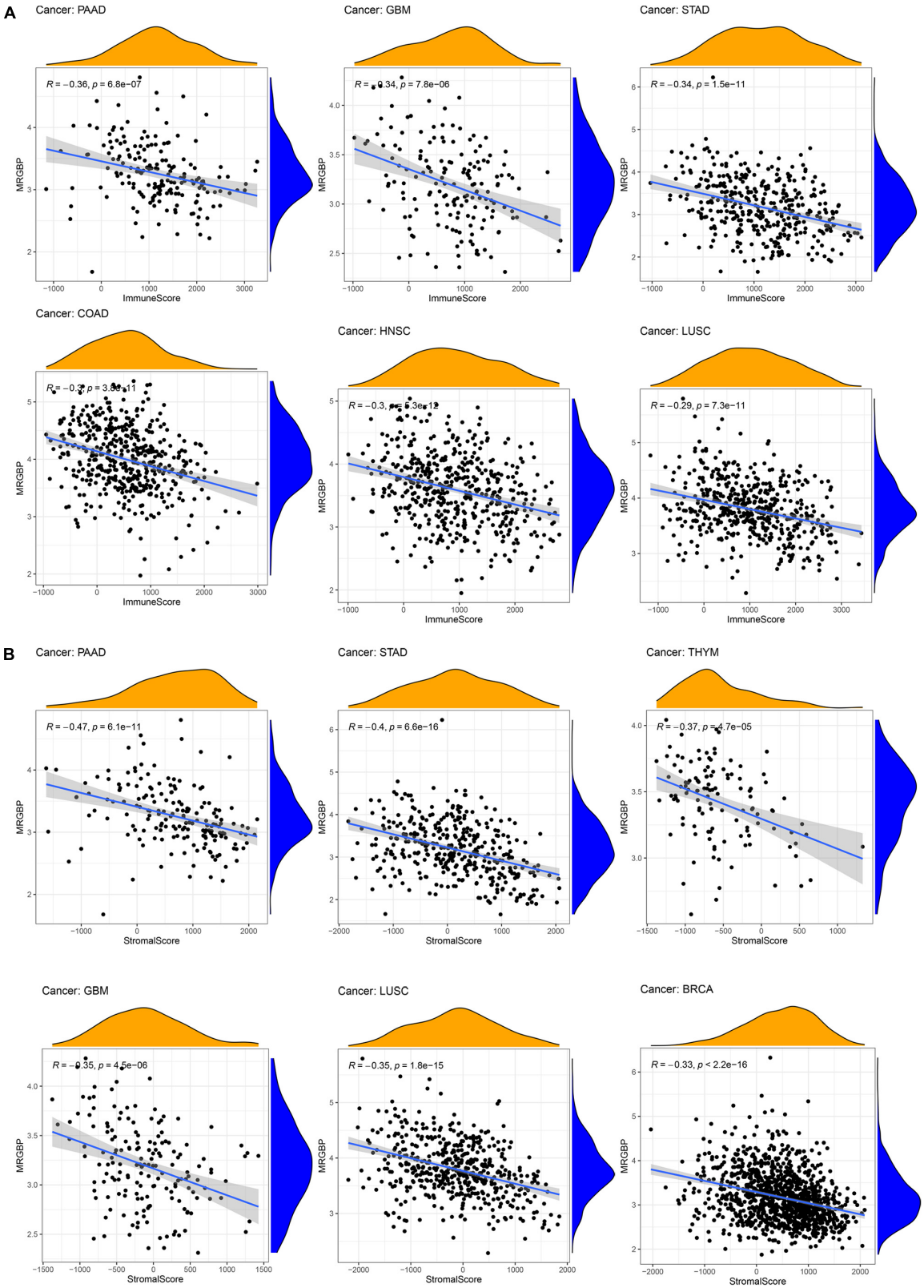


FIGURE 8 | Association between MRGBP expression and immune scores **(A)** in PAAD, GBM, STAD, COAD, HNSC, and LUSC and stromal scores **(B)** in PAAD, STAD, THYM, GBM, LUSC, and BRCA.

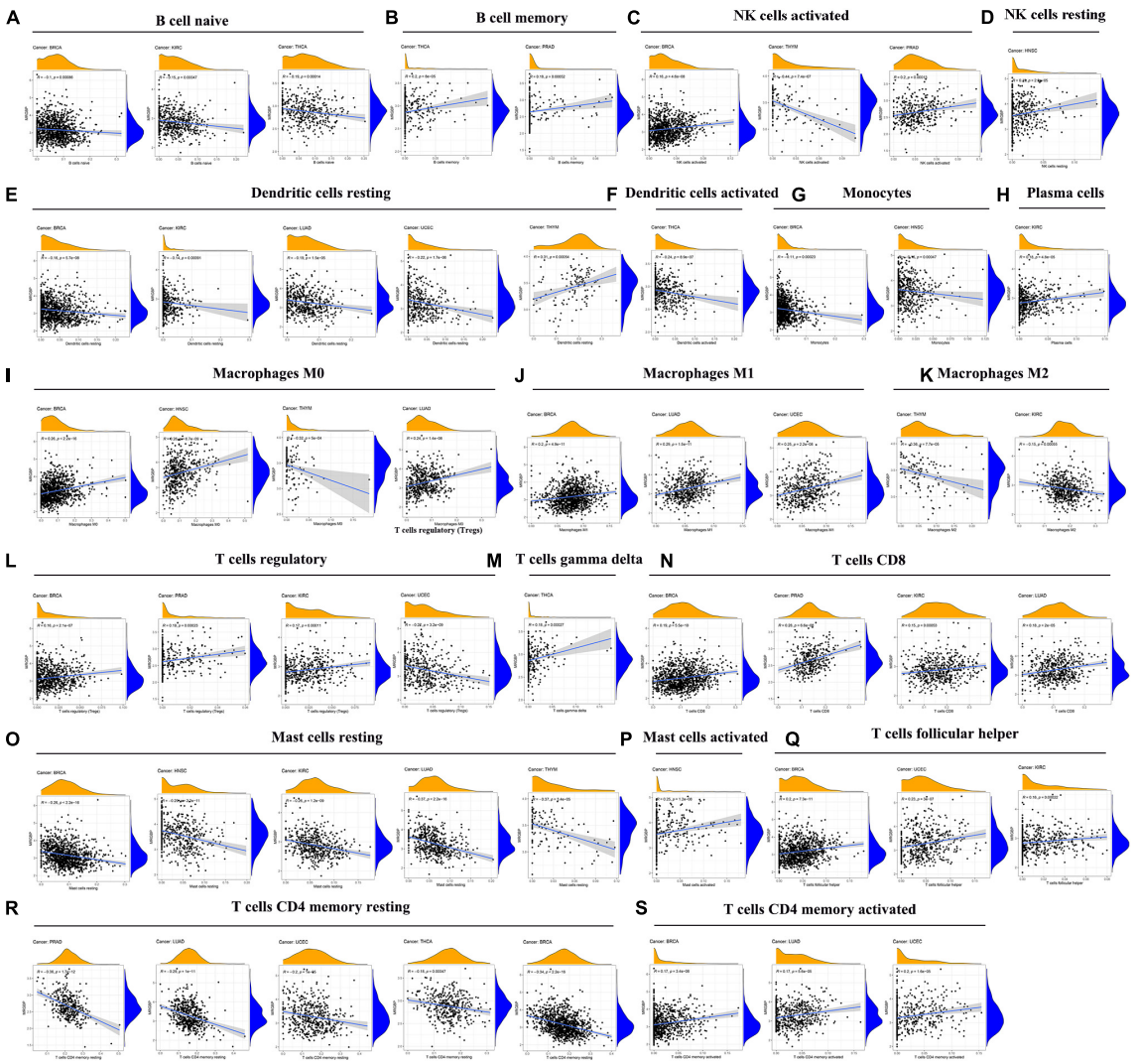


FIGURE 9 | Association between MRGBP expression and the level of tumor infiltration of different immune cells in BRCA, HNSC, KIRC, LUAD, PRAD, THCA, THYM, and UCEC.

Figure 10, we observed that most of the immune-related genes were associated with MRGBP expression and the majority were negatively related to the expression of MRGBP in different types of tumor except UVM. This suggests an interaction between MRGBP and immune checkpoints. Also, we investigated the expression difference between responders and non-responders according to TISIDB, which showed MRGBP expression was not significantly different between immunotherapy responders and non-responders (**Supplementary Table 2**).

Enrichment Analysis of MORF4-Related Gene-Binding Protein-Related Partners

We screened MRGBP binding proteins and MRGBP expression-related genes for a series of pathway enrichment analyses to investigate the molecular mechanisms of MRGBP genes in tumorigenesis and tumor progression. As shown in the PPI network (**Figure 11A**), we obtained a total of 30 proteins that bind predominantly to MRGBP and these are supported by experimental evidence in STING. Also, we explored the correlation between MRGBP and the top 100 genes that related to MRGBP expression using the GEPIA2 tool, the results showed that MRGBP had the most significant correlation with PSMA7 ($R = 0.74$; $p < 0.001$), ADRM1 ($R = 0.69$; $p < 0.001$), TPD52L2 ($R = 0.66$; $p < 0.001$), AURKA ($R = 0.64$; $p < 0.001$) and DDX27 ($R = 0.63$; $p < 0.001$) (**Figure 11B**). Heatmap data also indicated a positive correlation between MRGBP and the above five genes in most cancer types (**Figure 11C**).

We next performed GO function annotation and KEGG pathway analysis using data from GSEA. The results of GO functional annotation showed MRGBP related genes were involved in the regulation of the cell cycle such as organelle fission, nuclear division, mitotic

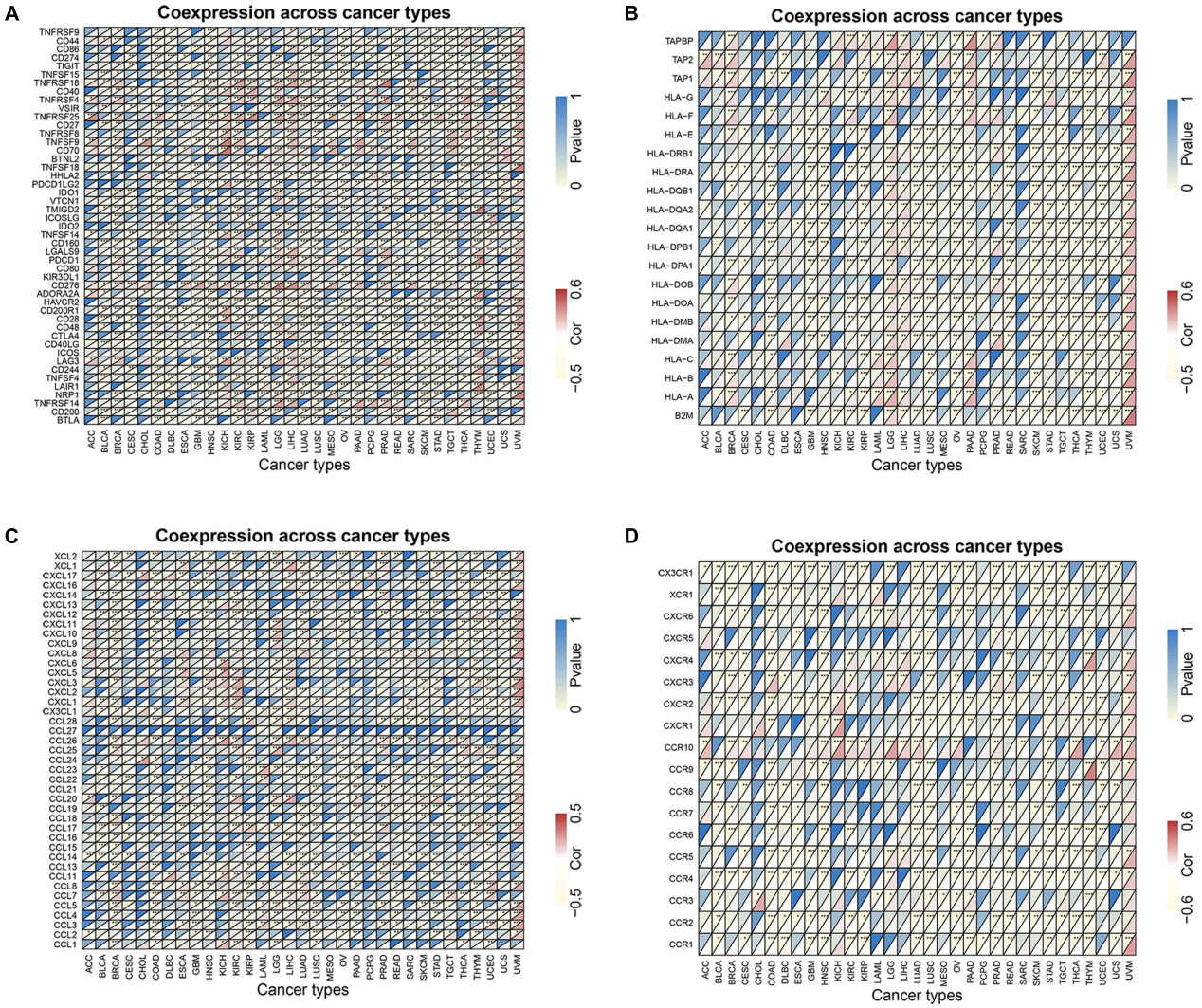


FIGURE 10 | Co-expression of immune-related genes with MRGBP in pan-cancers. **(A)** Co-expression of immune checkpoint related genes with MRGBP. **(B)** Co-expression of MHC genes with MRGBP. **(C)** Co-expression of chemokine genes with MRGBP. **(D)** Co-expression of chemokine receptor genes with MRGBP.

nuclear division, and histone modification. According to KEGG pathway enrichment analysis, MRGBP-related genes were positively correlated with organelle fission, nuclear division, mitotic nuclear division, histone modification which were consistent with the results of GO analysis (**Figures 11D,E**).

To better understand the potential mechanisms of MRGBP in cancer, we estimated the functional status of MRGBP in the CancerSEA database. As shown in **Supplementary Figure 9**, MRGBP has been studied at the single cell level in 13 types of tumors. We found that MRGBP was positively associated with DNA repair (cor = 0.479, P = 0.007), while negatively correlated with angiogenesis (cor = −0.363, P = 0.049) in acute lymphoblastic leukemia. MRGBP was negatively associated with invasion (cor = −0.433, $P < 0.001$) in ovarian cancer. MRGBP was positively associated with inflammation (cor = 0.329, P = 0.012) and quiescence (cor = 0.323, P = 0.013), while negatively correlated with DNA repair (cor = −0.406, P = 0.002) in colorectal cancer.

Preliminary Verification of MORF4-Related Gene-Binding Protein Signature in Lower Grade Glioma and LIHC

Notably, in LGG and LIHC, high expression of MRGBP was significantly associated with poor prognosis including OS, DSS, DFI, and PFI ($p < 0.001$), suggesting that elevated MRGBP has significant predictive power for prognosis. Hence, we identified LGG and LIHC as representative cancer types for subsequent analysis. At first, we evaluated the difference in MRGBP expression in LGG and normal tissues using IHC. The results showed MRGBP was significantly upregulated (**Figures 12A,B**). We also verified the expression level of MRGBP in the GEO

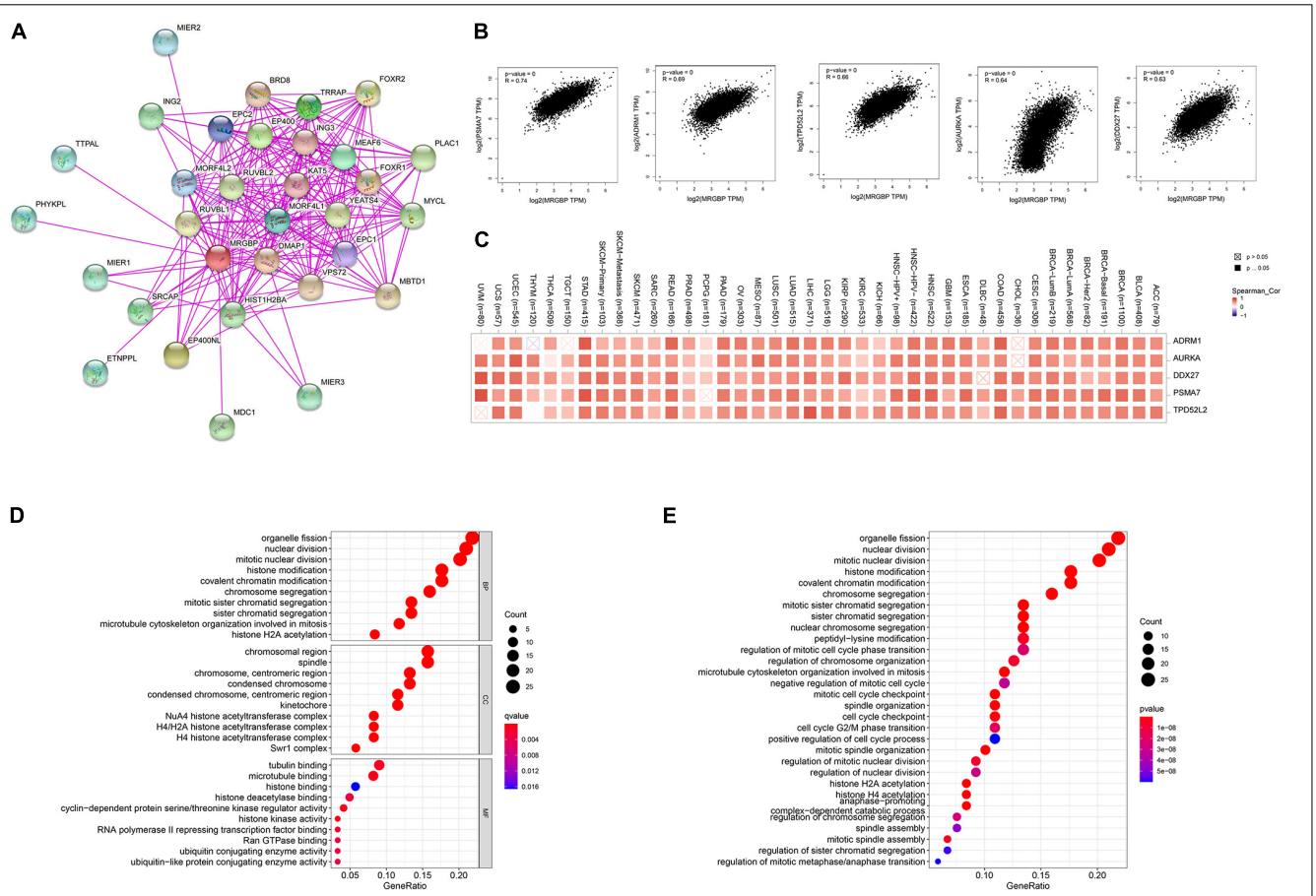


FIGURE 11 | Enrichment analysis of MRGBP-related partners. **(A)** The PPI network of MRGBP is constructed by STRING database. **(B)** Correlation analysis between MRGBP and main interacting genes in TCGA using GEPIA2 tool. **(C)** Corresponding heatmap between the MRGBP expression and the main interacting genes in 33 cancer types. **(D)** The barplot and bubble of GO enrichment analysis. **(E)** The barplot and bubble of KEGG enrichment analysis. KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; BP, biological process; CC, cellular component; MF, molecular function.

database (**GSE16011**) and the results were consistent with our findings (**Figure 12C**). Also, the results of IHC showed MRGBP was highly expressed in COAD and READ compared with normal tissues (**Supplementary Figure 10A**). Moreover, we investigated the prognostic value of MRGBP in LGG using the Chinese Glioma Genome Altas (CGGA) and GEO dataset (**GSE 4412**). The results of CGGA showed a high expression of MRGBP was related to a detrimental effect on overall survival (**Supplementary Figure 10B**). Similarly, increased expression of MRGBP predicted poor overall survival according to the GEO dataset (**Figure 12D**), which is generally consistent with our survival analysis of the TCGA database. Next, we conducted a univariate independent prognostic analysis using the GEO dataset, the expression level of MRGBP, age, gender, and grade were risk factors used to determine OS (**Figure 12E**) and it is consistent with the results from TCGA (**Supplementary Figure 10C**) except gender.

Next, we investigated the correlation between MRGBP and LIHC by detecting MRGBP protein expression in microarrays of liver tumor tissues containing 105 LIHC tissues and adjacent normal liver tissues by IHC staining. The results showed that MRGBP was significantly overexpressed in LIHC compared with adjacent normal tissues ($p < 0.001$) (**Figures 13A,C** and **Supplementary Table 3**). 90 patients containing clinical features and survival information were selected for further study of the association between MRGBP expression and clinical features and prognosis (**Supplementary Table 3**). We found that MRGBP expression was higher in stage III and IV patients compared to stage I and II ($p = 0.04$) (**Figure 13D**). In female, age $\geq$ 60, AFP $\geq$ 400 and Child-pugh B subgroups, MRGBP tended to be highly expressed although there was no statistically significant difference ($p > 0.05$) (**Supplementary Figure 11**). The survival curves revealed that patients with high expression of MRGBP had an unfavorable OS ($p = 0.0015$) (**Figure 13E**). Moreover, the results of univariate and multivariate COX analyses also identified MRGBP as a risk factor for OS (**Figure 13B**). The prognostic value of MRGBP in subgroups of LIHC patients with different clinical characteristics was further explored (**Table 2**). MRGBP correlated with poor prognosis for OS in male patients and patients with age < 60, AFP less than 400, earlier clinical stage and child-pugh A.

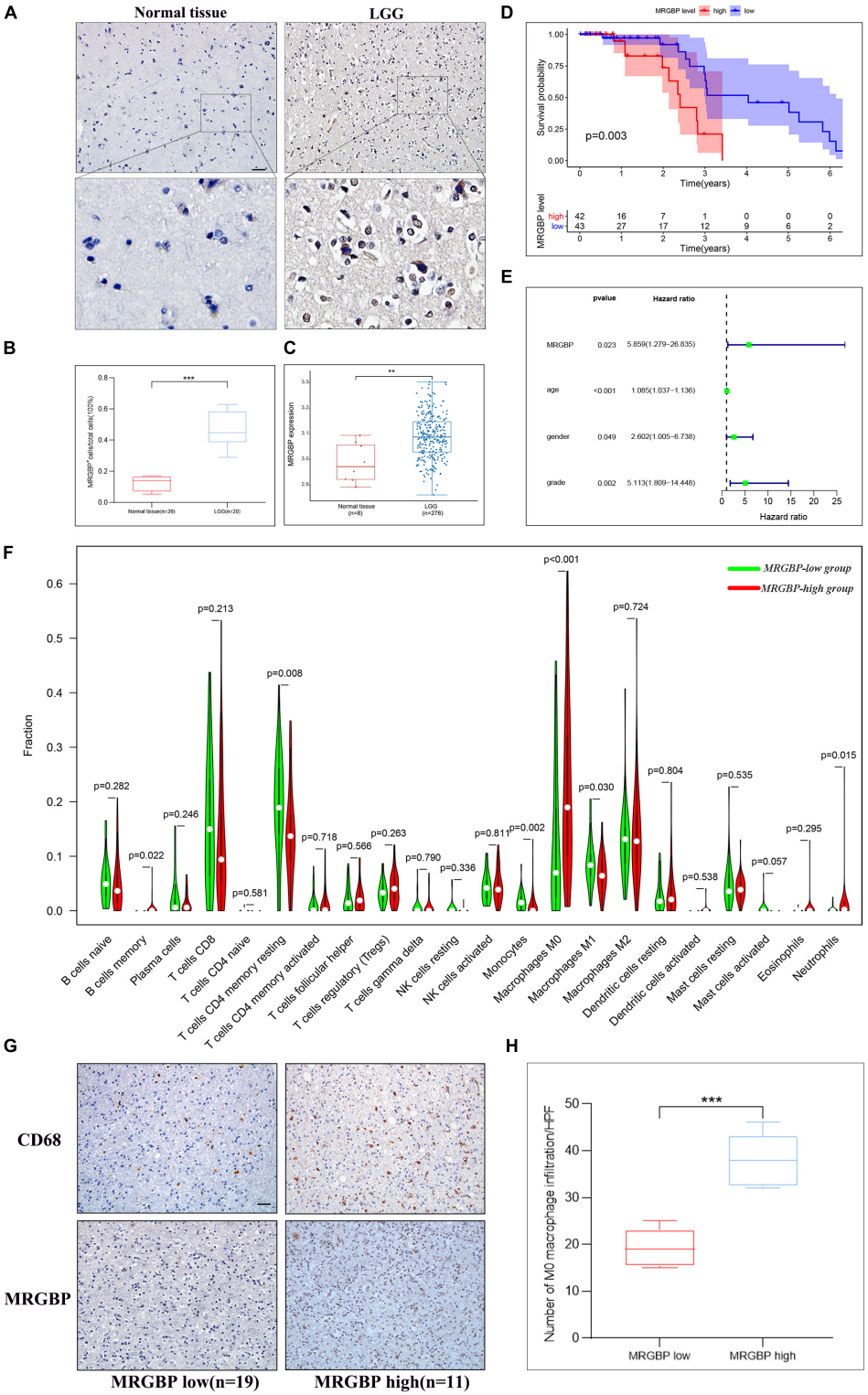


FIGURE 12 | Preliminary verification of MRGBP signature in LGG. **(A)** Representative images of MRGBP immunohistochemistry in LGG. The picture below shows a 16× magnification. scale bars = 50 μm. **(B)** Quantification analysis showed higher MRGBP expression in LGG. **(C)** Validation of MRGBP expression by GSE16011. **(D)** Association between MRGBP expression and prognosis in GSE4412. **(E)** Univariate independent prognostic analysis in GSE4412. **(F)** Relative abundance of tumor-infiltrating immune cells in the MRGBP low expression group vs. MRGBP high expression group. **(G)** Representative images of CD68 immunohistochemistry in the MRGBP low expression group and MRGBP high expression group. scale bars = 50 μm. **(H)** Quantification analysis showed higher levels of M0 macrophage infiltration in the MRGBP high expression group (***$P < 0.001$, **$P < 0.01$).

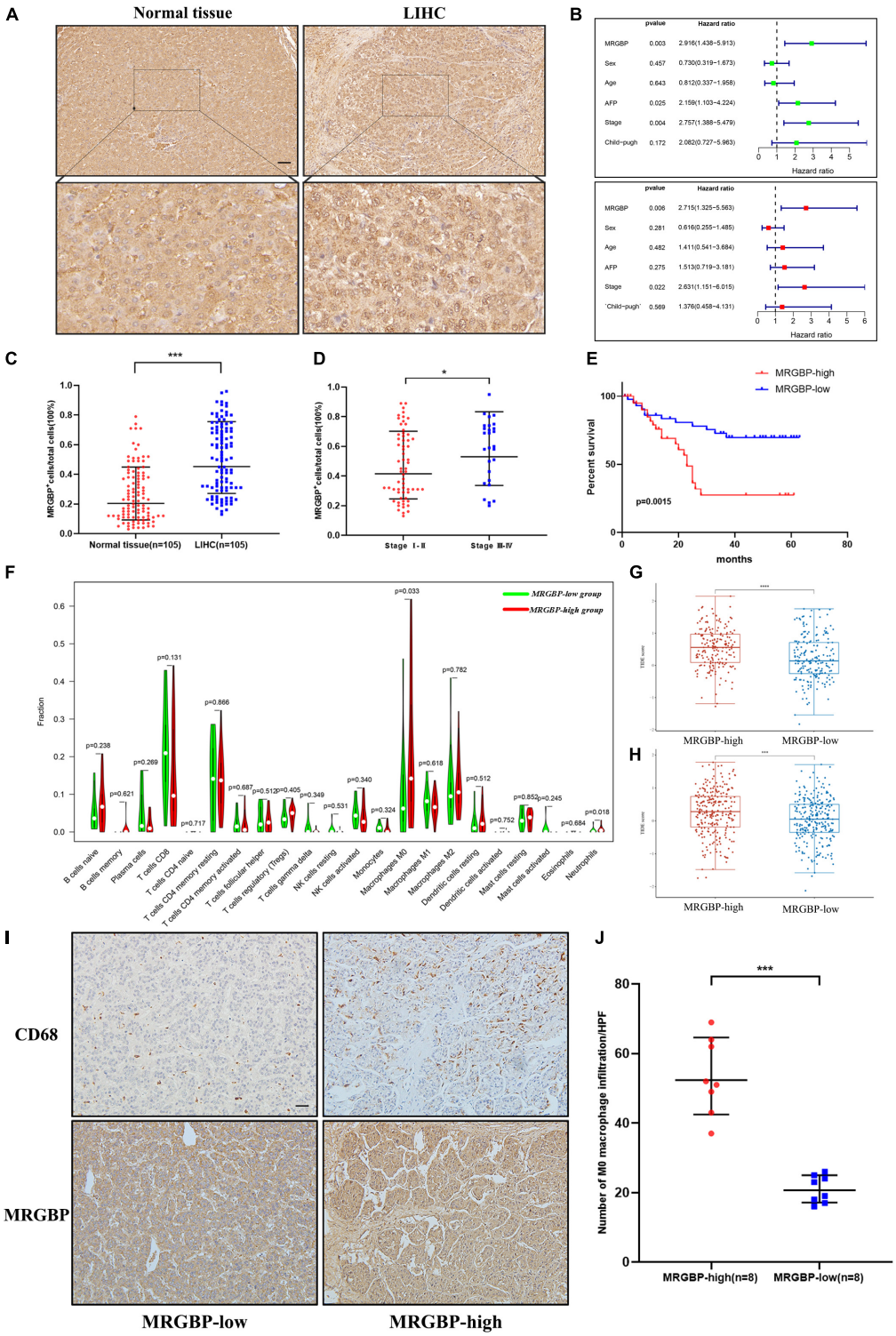


FIGURE 13 | Preliminary verification of MRGBP signature in LIHC. **(A)** Representative images of MRGBP immunohistochemistry in LIHC. The picture below shows the enlarged images. scale bars = 50 μm. **(B)** Univariate and multivariate independent prognostic analysis of our LIHC data. **(C)** Quantification analysis showed higher MRGBP expression in LIHC. **(D)**. Quantification analysis showed higher MRGBP expression in stage III-IV patients. **(E)** Association between MRGBP expression and prognosis. **(F)** Relative abundance of tumor-infiltrating immune cells in the MRGBP low expression group vs. MRGBP high expression group. **(G,H)** TIDE scores in the high MRGBP and low MRGBP groups at LIHC and LGG. **(I)** Representative images of CD68 immunohistochemistry in the MRGBP low expression group and MRGBP high expression group. scale bars = 50 μm. **(J)** Quantification analysis showed higher levels of M0 macrophage infiltration in the MRGBP high expression group (****$P < 0.0001$, ***$P < 0.001$, *$P < 0.05$).

TABLE 2 | Univariate survival analysis of MRGBP expression in subgroups with different clinical parameters.

Clinical parameters	n	Hazard ratio (95% CI)	*p*-value
Sex			
Male	75	3.688 (1.639–8.297)	**0.002**
Female	15	1.028 (0.213–4.969)	0.973
Age			
<60	70	3.806 (1.688–8.584)	**0.001**
≥60	20	0.908 (0.162–5.077)	0.912
AFP			
<400	48	4.586 (1.524–13.799)	**0.007**
≥400	42	1.959 (0.781–4.918)	0.152
Stage			
I–II	63	2.730 (1.129–6.602)	**0.026**
III–IV	27	2.518 (0.750–8.452)	0.135
Child-pugh			
A	82	2.893 (1.377–6.079)	**0.005**
B	8	1.674 (0.168–16.681)	0.660

Bold values represent a statistical difference (p <0.05).

Furthermore, we evaluated the abundance of TILs in the LGG and LIHC microenvironment using the TCGA gene expression data. The results showed that the levels of M0 macrophages were significantly higher in the MRGBP-high group in both LGG and LIHC, while the numbers of M1 macrophage, and resting memory CD4 T cells were significantly lower in the MRGBP-high group than MRGBP-low group in LGG. Lower levels of CD8 T cells infiltration were seen in patients in the MRGBP-high group of LGG and LIHC, although no significance was demonstrated (**Figures 12F**, **13F**). Subsequently, TIDE algorithm (Jiang et al., 2018) was used to predict the potential ICB response. Our results showed elevated TIDE scores in MRGBP-high group in LIHC (**Figure 13G**) and LGG (**Figure 13H**), suggesting poor efficacy of immune checkpoint blockade therapy in the patients with high expression of MRGBP. The relationship between MRGBP expression and M0 macrophage infiltration level was next analyzed. We used serial paraffin sections from the same case to verify the relationship between CD68 and MRGBP expression levels. Representative images are shown in **Figures 12G**, **13I** and it showed that in the group with high MRGBP expression levels, CD68 staining was stronger compared to the MRGBP-low group. MRGBP levels were positively correlated with the number of M0 macrophages (**Figures 12H**, **13J**).

DISCUSSION

MRGBP, a component of the NuA4 histone acetyltransferase complex, is involved in the transcriptional activation of selected genes mainly through acetylation of nucleosomal histones H4 and H2A. MRGBP is upregulated in several types of cancers such as colorectal, cervical, prostate, pancreatic, and cutaneous squamous cell carcinomas and has been proved to increase replication, induce apoptosis, reduce growth and promote aggressiveness. There was also evidence that upregulation of MRGBP was associated with worse survival. All this evidence indicated that MRGBP was associated with tumor development as well as patient prognosis. However, no studies were exploring the role of MRGBP in a variety of cancers. In this report, we performed a multifaceted analysis of MRGBP in human cancers including its expression levels in tumors, normal tissues, and cell lines, the relationship between MRGBP expression and patient prognosis, TMB, and MSI. Also, we explored the correlation of MRGBP expression with TME, immune cell infiltration levels, and immune-related genes. Finally, we conducted a preliminary validation of our analysis results in the LGG.

Several studies have shown that MRGBP expression was increased in most cancer cell lines such as PDAC cell lines and the proliferative, migratory and invasive capacities of ASPC-1 and Mia PaCa-2 were significantly reduced (Ding et al., 2017). Another study showed that MRGBP expression was higher in most prostate cancer cell lines than in normal prostate cell lines and was able to enhance their aggressiveness (Ito et al., 2014). Similarly, elevated MRGBP expression was observed in colorectal cancer cell lines (Yamaguchi et al., 2011). Next, we analyzed the expression levels of MRGBP in a variety of normal as well as tumor tissues using multiple databases. The results showed that gene levels of MRGBP were significantly higher in 20 tumors compared to normal tissue. MRGBP expression in colorectal, prostate, and lung cancers is consistent with their previous studies (Yamaguchi et al., 2010; Ito et al., 2018; Dai et al., 2019) and it was also confirmed by our immunohistochemical results in LGG (**Figure 12A**). These findings suggest that MRGBP may play an important role in the development of the above tumors and it may be a promising diagnostic biomarker in tumors. Some studies analyzed the changed genes in intestinal and diffuse gastric cancer in the TCGA ribonucleic acid sequence data and found that MRGBP was one of the most changed genes (Tanabe et al., 2020). Our results also showed that MRGBP expression was significantly different among different molecular subtypes in most types of tumors.

Also, we analyzed the relationship between the MRGBP expression and prognostic value in 33 types of cancer using TCGA databases and Kaplan-Meier plotter. Based on both Cox and KM survival analysis of TCGA, high MRGBP expression served as a detrimental prognostic factor in some types of cancer including DLBC, KIRC, KIRP, LGG, LIHC, MESO, PCPG, UCEC, UVM. Similarly, in pancreatic cancer, elevated expression of MRGBP was previously reported as positively associated with poor prognosis (Ding et al., 2017). Also, analysis of altered genes in 37 colorectal adenomas and 31 adenocarcinomas showed that MRGBP expression was significantly higher in carcinomas compared to adenomas (Carvalho et al., 2009). This revealed MRGBP played an important role in the progression of adenoma to cancer and it might have potential clinical applications as a highly specific biomarker for colorectal cancer. According to the Kaplan-Meier plotter, we found that elevated MRGBP expression was related to worse prognosis in all tumors with statistically different, especially in KIRC, KIRP, LIHC, and UCEC which is consistent with the previous results analyzed with TCGA. We also performed a preliminary validation in LGG using the GEO and CGGA database. The results were consistent with previous studies in TCGA. The results of our

LIHC data analysis also indicated that patients with high MRGBP expression were associated with shorter survival times and that MRGBP expression levels were an independent prognostic factor. These results suggested that MRGBP could be used as a potential prognostic biomarker for multiple tumors, especially for LGG and LIHC.

Furthermore, MRGBP expression in relation to age was observed in certain types of cancers. MRGBP was lowly expressed in older patients (age > 65) with ECSC, LIHC, and LUAD, while in KIRC, LGG, OV, SARC, STAD, and UCEC, lower MRGBP expression was associated with young patients. These results have implications for the selection of immunotherapy regimens for patients in different age groups. In most cancers, MRGBP expression was associated with tumor stage and was significantly higher in patients with advanced stages. MRGBP expression was significantly higher in stage III and IV tumors. Our tissue microarray results confirmed a higher number of MRGBP-positive cells in stage III and IV patients.

TMB is a newly discovered quantifiable clinical metric that holds promise for predicting tumor response to immunotherapy (Samstein et al., 2019). High TMB has been reported to predict the efficacy of tumor immunotherapy in patients with lung cancer (Devarakonda et al., 2018), melanoma (Alexandrov et al., 2013), and colorectal cancers (Lee et al., 2019). Patients with high TMB had significantly longer overall survival compared to patients with low TMB after immune checkpoint blockade treatment (Samstein et al., 2019). MMR, an important DNA repair mechanism, plays an important role in maintaining genomic stability. Mutations or deletions in MMR genes lead to defects in gene function and microsatellite instability (MSI) (Boland and Goel, 2010). Lee et al. (2016) reported that the PD-1 monoclonal antibody pembrolizumab improved survival in patients with high MSI cancers compared to those with low MSI cancers. Similarly, colorectal cancer patients with high MSI gain better benefit from targeted monoclonal antibody nivolumab therapy (Overman et al., 2017). Our results showed that MRGBP expression was positively correlated with TMB and MSI in most cancer types. Patients with high MRGBP expression in these tumors may respond well to immunotherapy.

DNA and RNA methylation (m6A) are two important nucleic acid modifications that play an important role in the regulation of gene expression and participate in many biological processes (Szigeti et al., 2018). Both of them can affect tumor proliferation, differentiation, tumorigenesis, invasion, and metastasis by regulating proto-oncogenes and oncogenes. In our research, MRGBP expression was significantly associated with DNA methyltransferase genes and M6A RNA methylation-related genes, particularly DNMT3B and YTHFD1. It has been revealed that lysosomal protease mRNA could be bound by YTHDF1 to inhibit its cross-presentation ability to antigens in dendritic cells. This suggested that YTHDF1 may be a possible therapeutic target for immunotherapy (Han et al., 2019). Overall, MRGBP may influence tumor development and ultimately patient prognosis by regulating DNA methyltransferase gene and M6A RNA methylation. Also, it may be a potential target to predict the efficacy of immunotherapy in patients.

Tumor cells, which usually colonize normal tissues, can form a tumor microenvironment (TME) together with stromal cells, immune cells and their secretory factors and extracellular matrix (ECM) components. Studies have shown that immune cells and the cytokines produced by them, including immunosuppressive and inflammatory cytokines, play a dual role in promoting or preventing cancer development (Junttila and de Sauvage, 2013; Lianyuan et al., 2018; Wang et al., 2020). Nevertheless, few studies have been conducted on the role of MRGBP in the immune microenvironment. In our research, the results showed that MRGBP expression negatively correlated with stromal and immune cell content in COAD, GBM, HNSC, KIRP, LUAD, LUSC, PAAD, STAD, and UCEC according to ESTIMATE scores. Our research further clarifies that MRGBP expression was significantly related to the levels of multiple immune-related cells including NK cells, dendritic cells, M0 macrophages, M1 macrophages, M1 macrophages, regulatory T cells, $CD4^+$ T cells, and $CD8^+$ T cells. Among them, $CD8^+$ T lymphocyte plays a tumor-killing function, while regulatory T cells attenuate effector T cell activity and promote immunosuppression in TME. M1-type macrophages secrete Th1 cytokines, which play a pro-inflammatory and anti-tumor role, while the M2 macrophages in TME promote angiogenesis and tumor invasion by secreting Th2 cytokines. Also, natural killer cells are capable of killing tumor-associated immune cells by releasing granzyme and perforin or by mediating antibody-dependent cytotoxicity with their Fc segment receptors. Dendritic cells are also affected by the hypoxic and inflammatory environment in the TME, which impairs their antigen-presenting activity (Gajewski et al., 2013; Cully, 2018; da Silva et al., 2019). We selected LGG and LIHC to further verify the relationship between MRGBP expression and the level of immune cell infiltration, and the results showed that in the high MRGBP expression group, the level of M0 macrophages was significantly higher, while M1 macrophages, $CD4^+$ T and $CD8^+$ T cells were lower compared to the low MRGBP group. These suggested that MRGBP regulated macrophage infiltration and might have a potential regulatory role in macrophage polarization. Our IHC analysis results confirmed that the high MRGBP expression group had more infiltration of macrophages. Besides, MRGBP expression was considered to be significantly associated with genes encoding immune checkpoints, MHC, chemokines, and chemokine receptor proteins in our study. The results of the TIDE algorithm showed that patients with low MRGBP expression have lower TIDE scores at LGG and LIHC, implying that patients with low MRGBP expression are more likely to benefit from immunotherapy. These results suggested that MRGBP may affect patient prognosis by modulating immune infiltration of tumors, and it may be a potential target for immunotherapy.

Finally, we performed a series of enrichment analyses by screening out genes closely related to MRGBP expression, and the results showed that MRGBP contributed to tumor development by regulating ganelle fission, nuclear division, mitotic nuclear division, and histone modifications. Similarly, previous studies have found that MRGBP was associated with DNA replication,

microchromosome maintenance, and cell division in colorectal cancer (Yamaguchi et al., 2011).

In summary, our results demonstrated high MRGBP expression in a variety of tumors and revealed the correlation between MRGBP expression and tumor progression. Moreover, MRGBP expression was significantly correlated with poor prognosis, TMB, MSI, and immune cell infiltration in most types of cancer. Besides, MRGBP expression level was an independent prognostic factor for LGG and LIHC. Low MRGBP expression was associated with lower TIDE scores, which suggested a better immunotherapy efficacy. These results revealed that MRGBP can serve as a potential prognostic biomarker. It plays an important role in tumor immune infiltration in various tumors and may be a potential target for immunotherapy, especially in LGG and LIHC.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Renmin Hospital of Wuhan University Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DC, LZ, and YG came up with the design and conception, prepared the material, conducted the experiments, collected the data, and analyzed the data. DC wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript, read, and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.703415/full#supplementary-material

Supplementary Figure 1 | mRNA expression of MRGBP gene in normal human tissues and tumor cell lines. **(A)** Normal tissue expression. **(B)** mRNA expression in 38 kinds of tumor cell lines from the CCLE database.

Supplementary Figure 2 | Kaplan-Meier survival curves comparing the high and low expression of MRGBP gene in various cancer types in Kaplan-Meier Plotter.

Supplementary Figure 3 | Relationship between MRGBP expression and age in ESCA, KIRC, LGG, LIHC, LUAD, OV, SARC, STAD, and UCEC.

Supplementary Figure 4 | Relationship between MRGBP expression and tumor stage in ACC, BRCA, COAD, ESCA, HNSC, KIRC, LIHC, LUSC, TGCT, and UVW.

Supplementary Figure 5 | mRNA expression of MRGBP in different molecular subtypes of COAD, GBM, HNSC, KIRP, LGG, LUSC, OV, PCPG, PRAD, READ, SKCM, STAD ACC, BRCA, and UCEC.

Supplementary Figure 6 | Pie chart showing the percentage of the different mutation types of MRGBP in various cancers according to the COSMIC database.

Supplementary Figure 7 | Association between MRGBP expression and immune scores and stromal scores in multiple cancers.

Supplementary Figure 8 | Correlation of MRGBP expression and infiltrating immune cells in multiple cancers.

Supplementary Figure 9 | Functional state of MRGBP across 13 types of cancer. The red plots indicated that MRGBP was positively correlated and the blue plots indicated that MRGBP was negatively correlated with the functional state identified by CancerSEA.

Supplementary Figure 10 | (A) Representative images of MRGBP immunohistochemistry in COAD and READ. **(B)** Association between MRGBP expression and prognosis in CGGA datasets. **(C,D)** Hazard ratio and *P*-value of constituents involved in univariate and multivariate Cox regression and some parameters of the MRGBP genes.

Supplementary Figure 11 | (A) Immunohistochemical staining of MRGBP on LIHC tissue microarrays. **(B)** Expression of MRGBP in subgroups with different clinical parameters.

Supplementary Table 1 | Database and its website.

Supplementary Table 2 | The expression of MRGBP between immunotherapy responders and non-responders.

Supplementary Table 3 | Expression levels of MRGBP and the clinical information contained in LIHC tissue microarrays.

REFERENCES

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421. doi: 10.1038/nature12477

Baretti, M., and Le, D. T. (2018). DNA mismatch repair in cancer. *Pharmacol. Ther.* 189, 45–62. doi: 10.1016/j.pharmthera.2018.04.004

Boland, C. R., and Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology* 138, 2073–2087.e3. doi: 10.1053/j.gastro.2009.12.064

Cai, Y., Jin, J., Tomomori-Sato, C., Sato, S., Sorokina, I., Parmely, T. J., et al. (2003). Identification of new subunits of the multiprotein mammalian TRRAP/TIP60-containing histone acetyltransferase complex. *J. Biol. Chem.* 278, 42733–42736. doi: 10.1074/jbc.C300389200

Carvalho, B., Postma, C., Mongera, S., Hopmans, E., Diskin, S., van de Wiel, M. A., et al. (2009). Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut* 58, 79–89. doi: 10.1136/gut.2007.143065

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Choucair, K., Morand, S., Stanbery, L., Edelman, G., Dworkin, L., and Nemunaitis, J. (2020). TMB: a promising immune-response biomarker, and potential spearhead in advancing targeted therapy trials. *Cancer Gene Ther.* 27, 841–853. doi: 10.1038/s41417-020-0174-y

Cully, M. (2018). Cancer: re-educating tumour-associated macrophages with nanoparticles. *Nat. Rev. Drug Discov.* 17:468. doi: 10.1038/nrd.2018.102

da Silva, J. L., Dos Santos, A. L. S., Nunes, N. C. C., de Moraes Lino da Silva, F., Ferreira, C. G. M., and de Melo, A. C. (2019). Cancer immunotherapy: the art of targeting the tumor immune microenvironment. *Cancer Chemother. Pharmacol.* 84, 227–240. doi: 10.1007/s00280-019-03894-3

Dai, J., Li, Z., Amos, C. I., Hung, R. J., Tardon, A., Andrew, A. S., et al. (2019). Systematic analyses of regulatory variants in DNase I hypersensitive sites identified two novel lung cancer susceptibility loci. *Carcinogenesis* 40, 432–440. doi: 10.1093/carcin/bgy187

Devarakonda, S., Rotolo, F., Tsao, M.-S., Lanc, I., Brambilla, E., Masood, A., et al. (2018). Tumor mutation burden as a biomarker in resected non-small-cell lung cancer. *J. Clin. Oncol.* 36, 2995–3006. doi: 10.1200/JCO.2018.78.1963

Ding, F., Zhang, S., Gao, S., Shang, J., Li, Y., Cui, N., et al. (2017). MRGBP as a potential biomarker for the malignancy of pancreatic ductal adenocarcinoma. *Oncotarget* 8, 64224–64236. doi: 10.18632/oncotarget.19451

Gajewski, T. F., Schreiber, H., and Fu, Y.-X. (2013). Innate and adaptive immune cells in the tumor microenvironment. *Nat. Immunol.* 14, 1014–1022. doi: 10.1038/ni.2703

Han, D., Liu, J., Chen, C., Dong, L., Liu, Y., Chang, R., et al. (2019). Anti-tumour immunity controlled through mRNA m6A methylation and YTHDF1 in dendritic cells. *Nature* 566, 270–274. doi: 10.1038/s41586-019-0916-x

Ito, S., Kayukawa, N., Ueda, T., Taniguchi, H., Morioka, Y., Hongo, F., et al. (2018). MRGBP promotes AR-mediated transactivation of KLK3 and TMPRSS2 via acetylation of histone H2A.Z in prostate cancer cells. *Biochim. Biophys. Acta Gene Regul. Mech.* 1861, 794–802. doi: 10.1016/j.bbagrm.2018.07.014

Ito, S., Ueda, T., Ueno, A., Nakagawa, H., Taniguchi, H., Kayukawa, N., et al. (2014). A genetic screen in Drosophila for regulators of human prostate cancer progression. *Biochem. Biophys. Res. Commun.* 451, 548–555. doi: 10.1016/j.bbrc.2014.08.015

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24, 1550–1558. doi: 10.1038/s41591-018-0136-1

Junttila, M. R., and de Sauvage, F. J. (2013). Influence of tumour microenvironment heterogeneity on therapeutic response. *Nature* 501, 346–354. doi: 10.1038/nature12626

Lee, D.-W., Han, S.-W., Bae, J. M., Jang, H., Han, H., Kim, H., et al. (2019). Tumor mutation burden and prognosis in patients with colorectal cancer treated with adjuvant fluoropyrimidine and oxaliplatin. *Clin. Cancer Res.* 25, 6141–6147. doi: 10.1158/1078-0432.CCR-19-1105

Lee, V., Murphy, A., Le, D. T., and Diaz, L. A. (2016). Mismatch repair deficiency and response to immune checkpoint blockade. *Oncologist* 21, 1200–1211. doi: 10.1634/theoncologist.2016-0046

Lianyuan, T., Dianrong, X., Chunhui, Y., Zhaolai, M., and Bin, J. (2018). The predictive value and role of stromal tumor-infiltrating lymphocytes in pancreatic ductal adenocarcinoma (PDAC). *Cancer Biol. Ther.* 19, 296–305. doi: 10.1080/15384047.2017.1416932

Lingel, H., and Brunner-Weinzierl, M. C. (2019). CTLA-4 (CD152): a versatile receptor for immune-based therapy. *Semin. Immunol.* 42, 101298. doi: 10.1016/j.smim.2019.101298

Luke, J. J., Flaherty, K. T., Ribas, A., and Long, G. V. (2017). Targeted agents and immunotherapies: optimizing outcomes in melanoma. *Nat. Rev. Clin. Oncol.* 14, 463–482. doi: 10.1038/nrclinonc.2017.43

Overman, M. J., McDermott, R., Leach, J. L., Lonardi, S., Lenz, H.-J., Morse, M. A., et al. (2017). Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *Lancet Oncol.* 18, 1182–1191. doi: 10.1016/S1470-2045(17)30422-9

Rizvi, H., Sanchez-Vega, F., La, K., Chatila, W., Jonsson, P., Halpenny, D., et al. (2018). Molecular determinants of response to anti-programmed cell death (PD)-1 and anti-programmed death-ligand 1 (PD-L1) blockade in patients with non-small-cell lung cancer profiled with targeted next-generation sequencing. *J. Clin. Oncol.* 36, 633–641. doi: 10.1200/JCO.2017.75.3384

Samstein, R. M., Lee, C.-H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., et al. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* 51, 202–206. doi: 10.1038/s41588-018-0312-8

Scotto, L., Narayan, G., Nandula, S. V., Arias-Pulido, H., Subramaniyam, S., Schneider, A., et al. (2008). Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. *Genes Chromosomes Cancer* 47, 755–765. doi: 10.1002/gcc.20577

Szigeti, K. A., Galamb, O., Kalmár, A., Barták, B. K., Nagy, Z. B., Márkus, E., et al. (2018). [Role and alterations of DNA methylation during the aging and cancer]. *Orv. Hetil.* 159, 3–15. doi: 10.1556/650.2018.30927

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Tanabe, S., Quader, S., Ono, R., Cabral, H., Aoyagi, K., Hirose, A., et al. (2020). Molecular network profiling in intestinal- and diffuse-type gastric cancer. *Cancers (Basel)* 12:3833. doi: 10.3390/cancers12123833

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947. doi: 10.1093/nar/gky1015

Topalian, S. L., Hodi, F. S., Brahmer, J. R., Gettinger, S. N., Smith, D. C., McDermott, D. F., et al. (2012). Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* 366, 2443–2454. doi: 10.1056/NEJMoa1200690

van Velzen, M. J. M., Derks, S., van Grieken, N. C. T., Haj Mohammad, N., and van Laarhoven, H. W. M. (2020). MSI as a predictive factor for treatment outcome of gastroesophageal adenocarcinoma. *Cancer Treat. Rev.* 86:102024. doi: 10.1016/j.ctrv.2020.102024

Wang, Z., Song, K., Zhao, W., and Zhao, Z. (2020). Dendritic cells in tumor microenvironment promoted the neuropathic pain via paracrine inflammatory and growth factors. *Bioengineered* 11, 661–678. doi: 10.1080/21655979.2020.1771068

Watt, S. A., Pourreyron, C., Purdie, K., Hogan, C., Cole, C. L., Foster, N., et al. (2011). Integrative mRNA profiling comparing cultured primary cells with clinical samples reveals PLK1 and C20orf20 as therapeutic targets in cutaneous squamous cell carcinoma. *Oncogene* 30, 4666–4677. doi: 10.1038/onc.2011.180

Wu, T., and Dai, Y. (2017). Tumor microenvironment and therapeutic response. *Cancer Lett.* 387, 61–68. doi: 10.1016/j.canlet.2016.01.043

Xin Yu, J., Hodge, J. P., Oliva, C., Neftelinov, S. T., Hubbard-Lucey, V. M., and Tang, J. (2020). Trends in clinical development for PD-1/PD-L1 inhibitors. *Nat. Rev. Drug Discov.* 19, 163–164. doi: 10.1038/d41573-019-00182-w

Yamaguchi, K., Sakai, M., Kim, J., Tsunesumi, S., Fujii, T., Ikenoue, T., et al. (2011). MRG-binding protein contributes to colorectal cancer development. *Cancer Sci.* 102, 1486–1492. doi: 10.1111/j.1349-7006.2011.01971.x

Yamaguchi, K., Sakai, M., Shimokawa, T., Yamada, Y., Nakamura, Y., and Furukawa, Y. (2010). C20orf20 (MRG-binding protein) as a potential therapeutic target for colorectal cancer. *Br. J. Cancer* 102, 325–331. doi: 10.1038/sj.bjc.6605500

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chai, Zhang, Guan, Yuan, Li and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 05 October 2021
doi: 10.3389/fcell.2021.753221

Prediction of Ovarian Cancer-Related Metabolites Based on Graph Neural Network

Jingjing Chen[1†], Yingying Chen[1†], Kefeng Sun[1†], Yu Wang[1], Hui He[1], Lin Sun[2], Sifu Ha[2], Xiaoxiao Li[3], Yifei Ou[3], Xue Zhang[4] and Yanli Bi[5*]*

[1] Department of Obstetrics and Gynecology, First Affiliated Hospital, Heilongjiang University of Chinese Medicine, Harbin, China, [2] Department of Reproductive Medicine, Dalian Maternal and Children's Centre, Dalian, China, [3] Graduate School of Heilongjiang University of Chinese Medicine, Harbin, China, [4] Department of General Practice, Beijing Friendship Hospital, Capital Medical University, Beijing, China, [5] Department of Reproductive Medicine, The First Affiliated Hospital, Henan University of Chinese Medicine, Zhengzhou, China

Ovarian cancer is one of the three most malignant tumors of the female reproductive system. At present, researchers do not know its pathogenesis, which makes the treatment effect unsatisfactory. Metabolomics is closely related to drug efficacy, safety evaluation, mechanism of action, and rational drug use. Therefore, identifying ovarian cancer-related metabolites could greatly help researchers understand the pathogenesis and develop treatment plans. However, the measurement of metabolites is inaccurate and greatly affects the environment, and biological experiment is time-consuming and costly. Therefore, researchers tend to use computational methods to identify disease-related metabolites in large scale. Since the hypothesis that similar diseases are related to similar metabolites is widely accepted, in this paper, we built both disease similarity network and metabolite similarity network and used graph convolutional network (GCN) to encode these networks. Then, support vector machine (SVM) was used to identify whether a metabolite is related to ovarian cancer. The experiment results show that the AUC and AUPR of our method are 0.92 and 0.81, respectively. Finally, we proposed an effective method to prioritize ovarian cancer-related metabolites in large scale.

Keywords: ovarian cancer, metabolite, Graph convolutional network, support vector machine, prediction

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Fei Shen,
South China University of Technology, China
Sheng Li,
Wuhan University, China

****Correspondence:***
Xue Zhang
Vowwzx@163.com
Yanli Bi
biyanli.mary@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *04 August 2021*
Accepted: *27 August 2021*
Published: *05 October 2021*

Citation:
Chen J, Chen Y, Sun K, Wang Y, He H, Sun L, Ha S, Li X, Ou Y, Zhang X and Bi Y (2021) Prediction of Ovarian Cancer-Related Metabolites Based on Graph Neural Network. Front. Cell Dev. Biol. 9:753221. doi: 10.3389/fcell.2021.753221

INTRODUCTION

Ovarian cancer is a common gynecological malignancy and one of the deadliest female diseases. Because the underlying symptoms are not obvious, about 70% of ovarian cancer patients are already at an advanced stage when they are diagnosed (Liu et al., 2020). The survival rates of patients with ovarian cancer at different stages are very different, and the mortality rate of patients with advanced stages exceeds 75% (Hussain, 2020). Therefore, there is an urgent need to find metabolites related to ovarian cancer to improve the prognosis of ovarian cancer and improve the efficiency of individualized treatment of patients (Perrone et al., 2020). Many life activities in cells occur at the metabolite level, so metabolomics has become one of the current research hotspots in the field of omics (Blimkie et al., 2020). The research of metabolomics in the early diagnosis of malignant tumors has shown its advantages (Agakidou et al., 2020). Ovarian cancer is a disease with a very high mortality rate of gynecological malignancies. There is an urgent need for a method to diagnose the disease early. The application of metabolomics in ovarian cancer can provide ideas for the diagnosis and prevention of ovarian cancer.

The analysis of the metabolites caused by the disease will help us to more comprehensively grasp the process of disease changes and the metabolic pathway of substances in the body, so as to make the clinical diagnosis more accurate. Zhou et al. (2010) collected 44 serous papillary ovarian cancer (stage I–IV) and 50 healthy women and found that histamine, purine nucleotide, glycine, serine, and sarcosine were the differential metabolites, and alanine, serine, cysteine, threonine, and glycine were overexpressed. Fong et al. (2011) found that there are 364 kinds of biochemical substances in human ovarian metabolic tissue by gas chromatography–mass spectrometry and liquid chromatography tandem mass spectrometry. Ovarian transformation can cause changes in energy utilization, resulting in glycolysis and fatty acids (such as carnitine, acetylcarnitine, and butyrylcarnitine) β-oxidation changes. Based on the non-targeted metabonomics method of LC/MS, Chen et al. (2011) analyzed the serum samples of 27 healthy women, 28 cases of benign ovarian tumor, and 29 cases of epithelial ovarian cancer. β-Cholestane-3,7,12,24,25, pentose glucoside, phenylalanine, glycine cholic acid, and propionyl carnitine are potential biomarkers for epithelial ovarian cancer.

Garcia et al. (2011) used the 1H NMR method to analyze the concentration of alanine, valine, phospholipid choline, etc., from the serum of 170 healthy women of appropriate age and 182 ovarian cancer stage I/II patients, while β-hydroxybutyrate, acetone, and acetoacetic acid have higher concentrations. These can be qualitatively compared with the changes in the concentration distribution of serum samples of cancer patients studied by other NMR-based metabolomics. This proves that early diagnosis of ovarian cancer can significantly affect the clinical outcome of patients with ovarian cancer. Chen et al. (2011) analyzed the serum samples of 27 healthy women, 28 cases of benign ovarian tumors, and 29 cases of epithelial ovarian cancer using LC/MS combined analysis, liquid chromatography selective ion monitoring mass spectrometry technology combined with PCR, and other pattern recognition techniques. The study found that 27-nor-5β-cholestane-3,7,12,24,25 pentanol glucuronide can be used in the early diagnosis of epithelial ovarian cancer. It is elevated in the serum of early epithelial ovarian cancer (stage I). Gaul et al. (2015) used ultra-high performance liquid chromatography and high-resolution mass spectrometry from 46 early (I/II) serous epithelial ovarian cancer (EOC) patients and 49 age-matched normal healthy female controls. UPLC-MS and tandem mass spectrometry (MS/MS) methods found that 16 metabolites in lipids and fatty acids have 100% accuracy in the diagnosis of early-stage ovarian cancer patients. Woo et al. (2009) tested the metabolites in urine of 10 breast cancer patients, 9 ovarian cancer patients, 12 cervical cancer patients, and 22 normal controls. They found that 1-Methyladenosine is a powerful biomarker for diagnosing ovarian cancer. Zhang et al. (2012) found that 2-Piperidinone could be used to distinguish epithelial ovarian cancer (EOC) and benign ovarian tumor (BOT).

Although multiple metabolites have been found to be related to ovarian cancer, the time and money cost of this discovery is huge. With the development of computational method, increasing number of researchers try to use machine learning or deep learning methods to solve biological problems (Chen et al., 2019, 2020; Zhao et al., 2020b). Disease-related genes (Peng and Zhao, 2020; Zhao et al., 2021), RNAs (Gebauer et al., 2021), proteins (Zhao et al., 2020c), and drugs (Tianyi et al., 2020) have all been identified by computational methods in large scale, which significantly increases the speed of discovering knowledge. Predicting disease-related metabolites by computational methods has become a hot issue in recent years. Hu et al. (2018) used random walk to identify disease-related metabolites by similarity network in 2018. Following this research, Wang et al. (2019) fused text mining technology with random walk to further infer relationship between metabolites and diseases. Then, Peng and Zhao (2020) developed "MDBIRW," which is an improved random walk method to identify disease-related metabolites. However, these methods all traverse network by random walk, which did not fully extract the topological relationship of similarity network. In 2020, Zhao et al. (2020a) proposed "Deep-DRM," which used Graph convolutional network (GCN) to encode similarity network and achieved high accuracy. In this paper, we followed this research and focused on ovarian cancer to provide support for the treatment and diagnosis of ovarian cancer by prioritizing metabolites.

MATERIALS AND METHODS

Deep-DRM is a method that fuses GCN, principal component analysis (PCA), and deep neural network (DNN). Considering we only focus on ovarian cancer, the sample set would be much smaller. Therefore, we used support vector machine (SVM) to replace DNN to build a model with a small sample. The workflow of our method is shown in **Figure 1**.

Metabolite and Disease Similarity Network

We used "PaDEL-Descriptor" (Yap, 2011) to estimate the chemical property of metabolites by their chemical structure. The output of this tool includes 1D and 2D descriptors and fingerprints. Each metabolite could be represented as a vector of 2,325 dimension in this way.

$$m_i = [v_1, v_2, ..., v_{2325}] \tag{1}$$

The similarity of each of the two metabolites could be calculated by vector cosine.

$$sim(\hat{m}_i, \hat{m}_j) = \frac{\sum_k^{2325} \hat{m}_i^k \times \hat{m}_j^k}{\sqrt{\sum_k^{2325} (\hat{m}_i^k)^2} \times \sqrt{\sum_k^{2325} (\hat{m}_j^k)^2}} \tag{2}$$

Using the similarities of metabolites, we could build a metabolite similarity network. In the network, each node is a metabolite and each edge is the similarity between the two metabolites.

Cheng et al. (2014) proposed SemFunsim to calculate disease similarity. We used their results to build an ovarian cancer similarity network. All the nodes in this network are diseases similar to ovarian cancer. The edges are the similarities between ovarian cancer and other diseases.

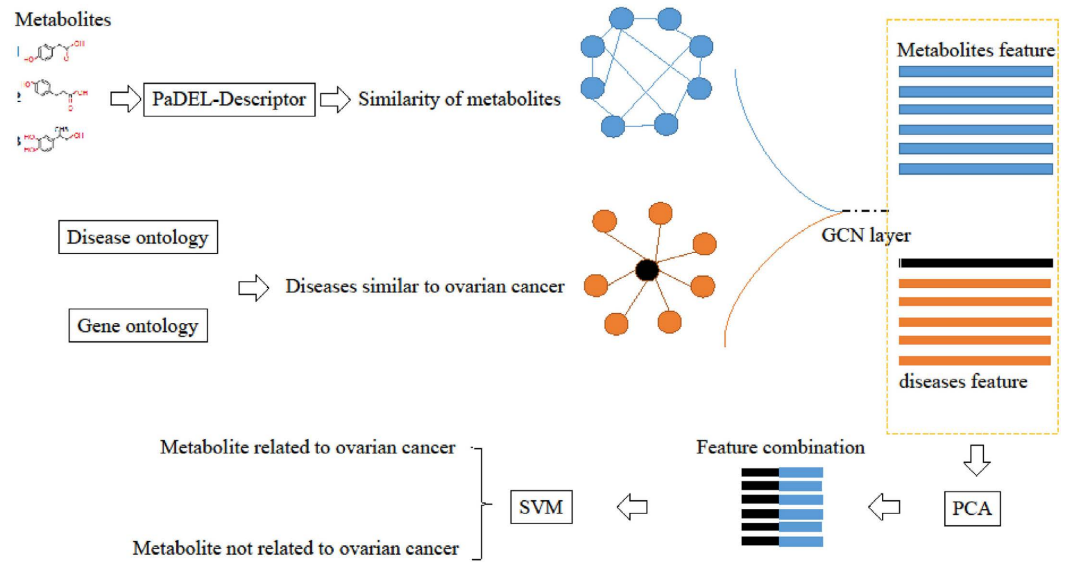


FIGURE 1 | Workflow of GPS-OCM (the fusion of GCN, PCA, and SVM to identify ovarian cancer-related metabolites).

Feature Encoding by Graph Convolutional Network

Graph convolutional network was implemented on both metabolite and disease similarity networks, respectively. The GCN-based network feature extraction method can convert the network structure into a vector output through a non-linear function:

$$H^{(l+1)} = f(H^{(l)}, A) \tag{3}$$

$H^{(0)} = X$ which is the initial feature of each node.

First, we need to perform Laplacian changes on the network, and the corresponding Laplacian matrix calculation formula is as follows:

$$L = D - A \tag{4}$$

Among them, D is the degree matrix of the graph, which can be solved by formula 5. A is the adjacency matrix.

$$\hat{D}_{ii} = \sum_{j} \hat{A}_{ij} \tag{5}$$

Since D is a diagonal matrix, only its diagonal elements need to be solved, and the remaining elements are all 0.

Then, we need to normalize the Laplacian matrix:

$$L^{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{6}$$

The final formula of GCN would be:

$$H^{(l+1)} = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(l)} W^{(l)}) \tag{7}$$

$\sigma()$ is the activation function, $W^{(l)}$ the parameter to be trained.

Finally, we obtained the encoded feature of metabolites and ovarian cancer.

Feature Dimensionality Reduction

Since the dimension of metabolites and ovarian cancer features are large, we used PCA to reduce the dimension.

Principal component analysis reduces the n-dimensional input data to r-dimensional, where r < n. PCA is essentially a basis transformation, so that the transformed data have the largest variance, that is, by rotating the coordinate axis and translating the origin of the coordinate, the variance between one of the axes (main axis) and the data point is minimized. After the coordinate conversion, the orthogonal axis with high variance is removed, and the dimensionality reduction data set is obtained.

The SVD method is used to perform PCA dimensionality reduction. Assuming that there are p × n-dimensional data samples X, there are p samples in total, and each row is n-dimensional. The p × n real matrix can be decomposed into:

$$X = U \sum V^T \tag{8}$$

Here, the dimension of the orthogonal matrix U is p × n, the dimension of the orthogonal matrix V is n × n (orthogonal matrix satisfies: $UU^T = VV^T = 1$), and Σ is a diagonal matrix of n × n. Next, divide Σ into r columns, denoted as Σr; use U and V to get the dimensionality reduction data point Yr:

$$Y_r = U \sum_{r} \tag{9}$$

After PCA, 99% of the feature information are preserved for both metabolites and ovarian cancer.

Identify Ovarian Cancer-Related Metabolites

After extracting and reducing the features of metabolites and ovarian cancer, we need to combine features of metabolites and

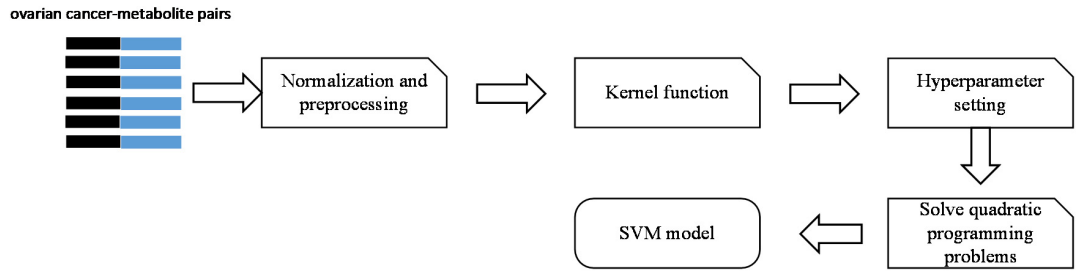


FIGURE 2 | Process of building the SVM model.

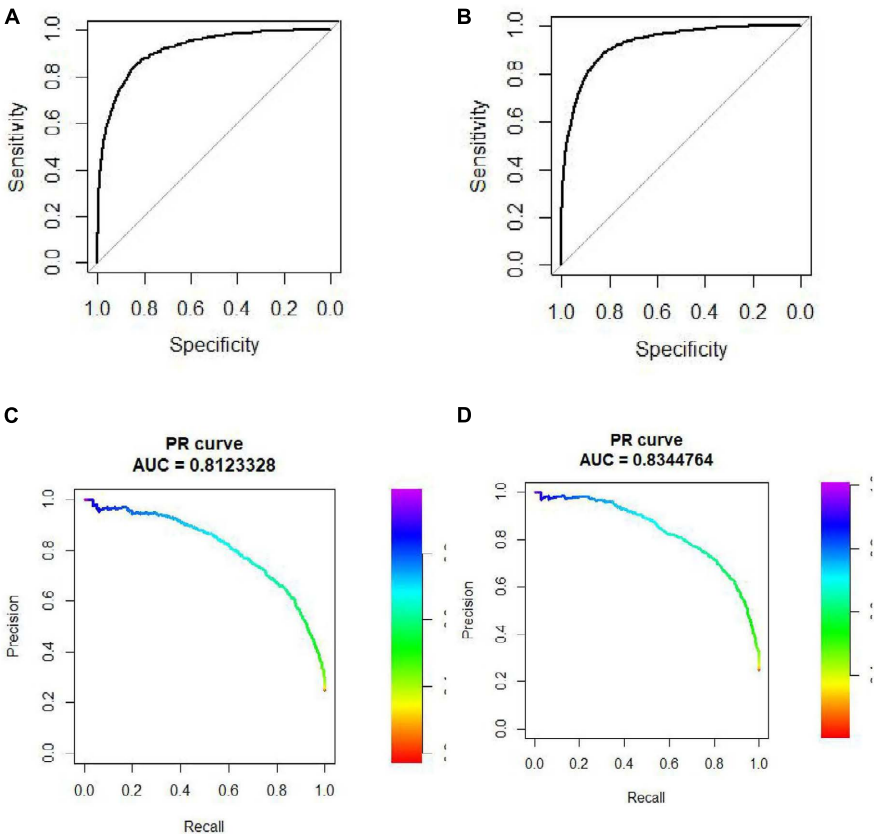


FIGURE 3 | (A) ROC curve of "SP." Experiment. **(B)** ROC curve of "SM" experiment. **(C)** PR curve of "SP" experiment. **(D)** PR curve of "SM" experiment.

ovarian cancer to make ovarian cancer–metabolite pairs. If the metabolite has relationship with ovarian cancer, the label of this pair would be 1; otherwise, the label is 0.

There are five steps to build the SVM model. The process is shown in **Figure 2**.

TABLE 1 | Comparison experiment of GPS-OCM and other methods on the "SP" test.

Method	AUC	AUPR
GPS-OCM	0.92	0.81
RWPS-OCM	0.83	0.70
GPR-OCM	0.87	0.73
GPD-OCM	0.90	0.81

EXPERIMENT RESULTS

Since we only focus on ovarian cancer, we divided our experiments into two classes. One is to identify ovarian cancer-related metabolites from known disease-related metabolites, which is named as "SP." The other one is to identify ovarian cancer-related metabolites from metabolites associated with no disease, which is names as "SM".

We did 10-cross validation on both "SP" and "SM" experiments. The AUC and AUPR of these experiments are shown in **Figure 3**.

The AUC of "SP" and "SM" experiments is 0.9168 and 0.9282, respectively. The AUPR of "SP" and "SM" experiments is 0.81 and 0.83, respectively.

To show the superiority of GPS-OCM, we compared GPS-OCM with RWPS-OCM, GPR-OCM, and GPD-OCM. We replaced GCN by Random Walk (RW) to construct RWPS-OCM. GPR-OCM is the fusion of GCN, PCR, and Random Forest (RF). GPD-OCM is to replace SVM by deep neural network (DNN). The results are shown in **Table 1**. GPS-OCM performed best among these methods.

After verifying the effectiveness of our method, we used all known ovarian cancer-related metabolites as positive samples and randomly selected equal number of other metabolites as negative samples to build a final GPS-OCM model. We totally identified 257 more metabolites that are associated with ovarian cancer. To verify whether these metabolites are associated with ovarian cancer, we chose the top five of these metabolites and did case studies.

Three of the top five metabolites have been reported to be related to ovarian cancers. Niemi et al. (2017) used morning urine samples from 23 women with benign ovarian tumors and 37 women with malignant ovarian tumors and found that N1, N12-Diacetylspermine showed significant statistical differences, and found that it can help distinguish benign and malignant ovarian tumors as well as early and advanced stage, and low malignant potential and high-grade ovarian cancers from each other, respectively. Fahrmann et al. (2021) found that 3-acetamidopropyl can significantly increase the sensitivity of ovarian cancer diagnosis by 116 ovarian cancer patients and 143 controls. Dessources et al. (2017) collected samples from 16 patients with benign ovarian pathology and 21 patients with malignant pathology and found that multiple metabolites are significantly associated with ovarian cancer including N-acetylation, acyl carnitines, and tryptophan.

CONCLUSION

Identifying ovarian cancer-related metabolites can help better understand pathogenic mechanism and disease process. In addition, metabolites in blood and urine have shown strong power in diagnosing cancer in early stage as biomarkers. However, few metabolites associated with ovarian cancer have been found at present. In order to speed up the study of metabolites related to ovarian cancer, we proposed a calculation method "GPS-OCM" based on similarity of metabolites and diseases. This method is fusion of GCN, PCA, and SVM. GCN was used to extract network topology features and PCA was implemented to reduce the dimension of disease and metabolite features. SVM was applied to do classification. The experiments show the high accuracy of our method with high AUC and AUPR. In addition, three of the top five metabolites that are identified as ovarian cancer-related metabolites by our method have been proven by previous studies, which proved the accuracy of our results.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

JC and YC wrote this manuscript. YW and HH did experiments. KS, SH, XL, and YO contributed to software analysis. LS provided important ideas. XZ and YB guided the whole work. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Scientific Research Projects of National Clinical Research Base of Chinese Medicine of Health Commission of Henan Province 2021JDZY101 and Doctoral research fund of the First Affiliated Hospital of Henan University of Chinese Medicine 2021BSJJ003 to YB.

REFERENCES

Agakidou, E., Agakidis, C., Gika, H., and Sarafidis, K. (2020). Emerging biomarkers for prediction and early diagnosis of necrotizing enterocolitis in the era of metabolomics and proteomics. *Front. Pediatr.* 8:838. doi: 10.3389/fped.2020.602255

Blimkie, T., Lee, A. H. Y., and Hancock, R. E. (2020). MetaBridge: an integrative multi-omics tool for metabolite−enzyme mapping. *Curr. Protoc. Bioinformatics* 70:e98. doi: 10.1002/cpbi.98

Chen, J., Zhang, X., Cao, R., Lu, X., Zhao, S., Fekete, A., et al. (2011). Serum 27-nor-5β-cholestane-3, 7, 12, 24, 25 pentol glucuronide discovered by metabolomics as potential diagnostic biomarker for epithelium ovarian cancer. *J. Proteome Res.* 10, 2625–2632. doi: 10.1021/pr200173q

Chen, S. C., Liu, Z. M., Li, M., Huang, Y. H., Wang, M., Wang, W., et al. (2020). Potential prognostic predictors and molecular targets for skin melanoma screened by weighted gene co-expression network analysis. *Curr. Gene Ther.* 20, 5–14. doi: 10.2174/1566523220666200516170832

Chen, X. G., Shi, W. W., and Deng, L. (2019). Prediction of disease comorbidity using hetesim scores based on multiple heterogeneous networks. *Curr. Gene Ther.* 19, 232–241. doi: 10.2174/1566523219666190917155959

Cheng, L., Li, J., Ju, P., Peng, J., and Wang, Y. (2014). SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One* 9:e99415. doi: 10.1371/journal.pone.0099415

Dessources, K., Cohen, J., Sen, K., Ramadoss, S., and Chaudhuri, G. (2017). N-acetylation and ovarian cancer: a study of the metabolomic profile of ovarian cancer compared to benign counterparts. *Gynecol. Oncol.* 147, 223–224. doi: 10.1016/j.ygyno.2017.07.089

Fahrmann, J. F., Irajizad, E., Kobayashi, M., Vykoukal, J., Dennison, J. B., Murage, E., et al. (2021). A MYC-driven plasma polyamine signature for early detection of ovarian cancer. *Cancers* 13:913. doi: 10.3390/cancers13040913

Fong, M. Y., McDunn, J., and Kakar, S. S. (2011). Identification of metabolites in the normal ovary and their transformation in primary and metastatic ovarian cancer. *PLoS One* 6:e19963. doi: 10.1371/journal.pone.0019963

Garcia, E., Andrews, C., Hua, J., Kim, H. L., Sukumaran, D. K., Szyperski, T., et al. (2011). Diagnosis of early stage ovarian cancer by 1H NMR metabonomics of serum explored by use of a microflow NMR probe. *J. Proteome Res.* 10, 1765–1771. doi: 10.1021/pr101050d

Gaul, D. A., Mezencev, R., Long, T. Q., Jones, C. M., Benigno, B. B., Gray, A., et al. (2015). Highly-accurate metabolomic detection of early-stage ovarian cancer. *Scientific reports* 5:16351. doi: 10.1038/srep16351

Gebauer, F., Schwarzl, T., Valcárcel, J., and Hentze, M. W. (2021). RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.* 22, 185–198. doi: 10.1038/s41576-020-00302-y

Hu, Y., Zhao, T., Zhang, N., Zang, T., Zhang, J., and Cheng, L. (2018). Identifying diseases-related metabolites using random walk. *BMC Bioinformatics* 19:116. doi: 10.1186/s12859-018-2098-1

Hussain, S. M. A. (2020). Molecular-based screening and therapeutics of breast and ovarian cancer in low-and middle-income countries. *Cancer Res. Stat. Treat.* 3:81. doi: 10.4103/CRST.CRST_88_20

Liu, T., Wei, Q., Jin, J., Luo, Q., Liu, Y., Yang, Y., et al. (2020). The m6A reader YTHDF1 promotes ovarian cancer progression via augmenting EIF3C translation. *Nucleic Acids Res.* 48, 3816–3831. doi: 10.1093/nar/gkaa048

Niemi, R. J., Roine, A. N., Häkkinen, M. R., Kumpulainen, P. S., Keinänen, T. A., Vepsäläinen, J. J., et al. (2017). Urinary polyamines as biomarkers for ovarian cancer. *Int. J. Gynecol. Cancer* 27, 1360–1366. doi: 10.1097/IGC.0000000000001031

Peng, J., and Zhao, T. (2020). Reduction in TOM1 expression exacerbates Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* 117, 3915–3916. doi: 10.1073/pnas.1917589117

Perrone, E., Lopez, S., Zeybek, B., Bellone, S., Bonazzoli, E., Pelligra, S., et al. (2020). Preclinical activity of sacituzumab govitecan, an antibody-drug conjugate targeting trophoblast cell-surface antigen 2 (Trop-2) linked to the active metabolite of irinotecan (SN-38), in ovarian cancer. *Front. Oncol.* 10:118. doi: 10.3389/fonc.2020.00118

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22:bbaa044. doi: 10.1093/bib/bbaa044

Wang, Y., Juan, L., Peng, J., Zang, T., and Wang, Y. (2019). Prioritizing candidate diseases-related metabolites based on literature and functional similarity. *BMC Bioinformatics* 20:574. doi: 10.1186/s12859-019-3127-4

Woo, H. M., Kim, K. M., Choi, M. H., Jung, B. H., Lee, J., Kong, G., et al. (2009). Mass spectrometry based metabolomic approaches in urinary biomarker study of women's cancers. *Clin. Chim. Acta* 400, 63–69. doi: 10.1016/j.cca.2008.10.014

Yap, C. W. (2011). PaDEL−descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32, 1466–1474. doi: 10.1002/jcc.21707

Zhang, T., Wu, X., Yin, M., Fan, L., Zhang, H., Zhao, F., et al. (2012). Discrimination between malignant and benign ovarian tumors by plasma metabolomic profiling using ultra performance liquid chromatography/mass spectrometry. *Clin. Chim. Acta* 413, 861–868. doi: 10.1016/j.cca.2012.01.026

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2020c). Identifying protein biomarkers in blood for Alzheimer's disease. *Front. Cell Dev. Biol.* 8:472. doi: 10.3389/fcell.2020.00472

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 22:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Zhou, M., Guan, W., Walker, L. D., Mezencev, R., Benigno, B. B., Gray, A., et al. (2010). Rapid mass spectrometric metabolic profiling of blood sera detects ovarian cancer with high accuracy. *Cancer Epidemiol. Biomarkers Prev.* 19, 2262–2271. doi: 10.1158/1055-9965.EPI-10-0126

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chen, Chen, Sun, Wang, He, Sun, Ha, Li, Ou, Zhang and Bi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 11 October 2021
doi: 10.3389/fcell.2021.720110

Histopathological Images and Multi-Omics Integration Predict Molecular Characteristics and Survival in Lung Adenocarcinoma

Linyan Chen[1†], Hao Zeng[1†], Yu Xiang[1], Yeqian Huang[2], Yuling Luo[2] and Xuelei Ma[1]*

[1] State Key Laboratory of Biotherapy, Department of Biotherapy, Cancer Center, West China Hospital, Sichuan University, Chengdu, China, [2] Department of Pathology, West China Hospital, Sichuan University, Chengdu, China

Histopathological images and omics profiles play important roles in prognosis of cancer patients. Here, we extracted quantitative features from histopathological images to predict molecular characteristics and prognosis, and integrated image features with mutations, transcriptomics, and proteomics data for prognosis prediction in lung adenocarcinoma (LUAD). Patients obtained from The Cancer Genome Atlas (TCGA) were divided into training set (n = 235) and test set (n = 235). We developed machine learning models in training set and estimated their predictive performance in test set. In test set, the machine learning models could predict genetic aberrations: *ALK* (AUC = 0.879), *BRAF* (AUC = 0.847), *EGFR* (AUC = 0.855), *ROS1* (AUC = 0.848), and transcriptional subtypes: proximal-inflammatory (AUC = 0.897), proximal-proliferative (AUC = 0.861), and terminal respiratory unit (AUC = 0.894) from histopathological images. Moreover, we obtained tissue microarrays from 316 LUAD patients, including four external validation sets. The prognostic model using image features was predictive of overall survival in test and four validation sets, with 5-year AUCs from 0.717 to 0.825. High-risk and low-risk groups stratified by the model showed different survival in test set (HR = 4.94, $p < 0.0001$) and three validation sets (HR = 1.64–2.20, $p < 0.05$). The combination of image features and single omics had greater prognostic power in test set, such as histopathology + transcriptomics model (5-year AUC = 0.840; HR = 7.34, $p < 0.0001$). Finally, the model integrating image features with multi-omics achieved the best performance (5-year AUC = 0.908; HR = 19.98, $p < 0.0001$). Our results indicated that the machine learning models based on histopathological image features could predict genetic aberrations, transcriptional subtypes, and survival outcomes of LUAD patients. The integration of histopathological images and multi-omics may provide better survival prediction for LUAD.

Keywords: lung cancer, histopathology, genomics, transcriptomics, proteomics

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Xiyuan Lu,
University of Texas at Austin, United States
Qi Wang,
University of Texas MD Anderson Cancer Center, United States

****Correspondence:***
Xuelei Ma
drmaxuelei@gmail.com

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *30 June 2021*
Accepted: *14 September 2021*
Published: *11 October 2021*

Citation:
Chen L, Zeng H, Xiang Y, Huang Y, Luo Y and Ma X (2021) Histopathological Images and Multi-Omics Integration Predict Molecular Characteristics and Survival in Lung Adenocarcinoma. Front. Cell Dev. Biol. 9:720110. doi: 10.3389/fcell.2021.720110

INTRODUCTION

Lung cancer is the most common cancer and the main cause of cancer death worldwide, resulting in an estimated 2.1 million new cases and 1.8 million deaths annually (Bray et al., 2018). Lung adenocarcinoma (LUAD) is the most major histological subtype, which is different from lung squamous cell carcinoma (LUSC) in clinical manifestations and therapeutic principles

(Ettinger et al., 2019). LUAD occurs more frequently in never-smokers compared to LUSC (Herbst et al., 2018). Although there were small improvements in 5-year survival rate of lung cancer patients, the survival rates of patients with lymph node invasion (29.7%) or distant metastases (4.7%) were still not optimistic (Schabath and Cote, 2019). Therefore, identifying high-risk patients with worse prognosis is critical to the treatment and management of cancer patients. In recent years, various novel biomarkers are constantly emerging to better classify LUAD patients by their probable prognosis, and promote the development of precision medicine (Vargas and Harris, 2016).

Oncogene addiction refers to the dependence of cancer cells on specific driver oncogenes that play important roles in tumorigenesis and progression (Pagliarini et al., 2015). Known driver oncogene aberrations of LUAD included mutations of *KRAS*, *EGFR*, and *BRAF* genes, and fusions of *ALK*, *ROS1*, and *RET* genes (Saito et al., 2016). Based on the concept of "oncogene addiction," cancer patients harboring such aberrations can benefit from molecular targeted therapy. Targeted agents against lung cancers with *EGFR*, *BRAF*, *ALK*, and *ROS1* aberrations (e.g., gefitinib, dabrafenib, and crizotinib) have achieved good efficacy in clinical trials, which are successful examples of clinical application of precision therapy (Saito et al., 2016; Park et al., 2019). The *EGFR* and *BRAF* mutations were found to be correlated with patient prognosis, but their prognostic implication in LUAD remains controversial (Calvayrac et al., 2017). In addition, based on unsupervised analysis of gene expression, the transcriptional subtypes of LUAD were proposed to provide clinical related classifications, and offered insights into tumors that lacked specific driver mutations (Cancer Genome Atlas Research Network, 2014). The terminal respiratory unit (TRU), proximal-inflammatory (PI), and proximal-proliferative (PP) subtypes revealed distinct genomic signatures, and the TRU subtype had a favorable prognosis (Cancer Genome Atlas Research Network, 2014). Furthermore, proteomic characterization of LUAD demonstrated the association of proteomic features with genetic aberrations and transcriptional subtypes (Gillette et al., 2020). Overall, multi-omics analysis is crucial for understanding the molecular landscape of cancer and improving the prognosis prediction and therapeutic strategies of patients.

Several clinicopathological factors are well-recognized prognostic factors for LUAD, such as TNM stage, smoking history, and performance status (Thakur and Gadgeel, 2016). Histopathological images contain numerous information about tumor morphology and its correlation with surrounding microenvironment. However, current histopathological assessment patterns (such as classification of tumor grade) are not sufficient to predict prognosis of LUAD patients (Yu et al., 2016). Computer-aided image analysis is an emerging field of artificial intelligence that converts digital pathological images into high-dimensional data, and offers a new approach to studying tumor heterogeneity and underlying pathophysiological mechanisms (Zhang et al., 2015). This digital innovation has potential to promote the modernization of pathology workflow, improve efficiency and consistency while maintaining diagnostic and prognostic accuracy, and provide decision support for clinicians (Niazi et al., 2019). Various types of quantitative image features have been defined, such as the cell size, shape, spatial distribution, and texture patterns (Tabesh et al., 2007). Based on histopathological image features, machine learning models have shown its utility in predicting tumor classification and patient outcome, such as lung (Yu et al., 2016), breast (Turkki et al., 2019), and prostate cancers (Lee et al., 2017). Furthermore, previous study indicated that the prognostic performance of histopathological image features was independent of other clinical factors in LUAD, including age, gender, tumor stage, and smoking status (Luo et al., 2017). Considering the properties of tumor and its microenvironment are closely related to molecular alterations, many researches have been conducted on the genomics and histopathological features. For example, promising result for the prediction of commonly mutated genes in lung cancer from histopathological images was reported (Coudray et al., 2018). Recent studies suggested that the integration of histopathological images and genomics data can enhance the ability to predict survival of cancer patients compared with using only one type of data (Cheng et al., 2017; Mobadersany et al., 2018; Zeng et al., 2020).

In this article, we designed an image processing pipeline to automatically extract image features from digital histopathological slides, and performed systematic analyses to correlate the features from histopathological images and omics profiles. Firstly, besides common mutations, we also built machine learning classifiers to predict the transcriptional subtypes of LUAD. In addition, we used histopathological image features alone or integrated them with genomics, transcriptomics and proteomics data to establish predictive models, and evaluated their prognostic roles for patient survival in independent datasets. We expected that the integrative models would more reliably predict survival risk and contribute to the personalized medicine of LUAD patients.

MATERIALS AND METHODS

Patient Cohorts

This study included two independent data sources. Firstly, hematoxylin and eosin (H&E)-stained histopathological images of 522 LUAD patients were obtained from The Cancer Imaging Archive (TCIA),[1] whereas the corresponding genomics, transcriptomics, and proteomics information were downloaded from The Cancer Genome Atlas (TCGA)[2] and The Cancer Proteome Atlas (TCPA) repositories.[3] The inclusion criteria were surgically resected LUAD patients with available histopathological images, genomics, and transcriptomics data, and 470 patients were finally included in this study. In addition, tissue microarrays (TMAs) of 316 LUAD patients were acquired from Shanghai Outdo Biotech Company (Shanghai, China). The TMA-LUAD datasets contained four cohorts of

[1]http://www.cancerimagingarchive.net/

[2]https://portal.gdc.cancer.gov/

[3]http://tcpaportal.org/tcpa/

patients (HLugA150Su01, n = 69; HLugA150Su02, n = 72; HLugA180Su04, n = 86; HLugA180Su05, n = 89). All tumor samples were collected by surgical resection. The pathologists first labeled the representative areas in formalin-fixed and paraffin-embedded tissues. The labeled tissues were arranged on the blank paraffin-embedded blocks by semi-automated tissue arrayer (TMArrayer), and the blocks were continuously sliced into TMAs. Digital TMA images were scanned by Aperio AT2 slide scanner. The utilization of TMAs was approved by the National Human Genetic Resources Sharing Service Platform (2005DKA21300), and informed consent was obtained from all patients.

Image Segmentation and Feature Extraction

The workflow of image processing and data integration was outlined in **Figure 1**. Since the original histopathological images (40× magnification) had extreme high resolution, we first cropped whole-slide images into millions of sub-images of 1,000 × 1,000 pixels through the Openslide-Python (Goode et al., 2013), and 60 sub-images from each whole-slide image for further analysis. The cell morphology of randomly selected sub-images still had good consistency (**Supplementary Figure 1**). Because the TMAs had much smaller size than whole-slide images, we applied the same processing method with adjusted size constraint to crop TMAs, and used all sub-images for feature extraction. We next used CellProfiler (McQuin et al., 2018) to separate the hematoxylin and eosin stains of sub-images and corrected the illumination. Then, CellProfiler automatically segmented the nuclei and cells through "Identify Primary/Secondary Objects" modules, and extracted specific image features from these cellular regions through 10 measurement modules, such as "Measure Object Size Shape," "Measure Correlation," "Measure Object Neighbors," and "Measure Texture" modules. These image features quantified the cell-level morphological characteristics, texture properties, and the relationship between neighboring objects. A total of 536 histopathological image features listed in **Supplementary Table 2** were included in this study.

Genetic Aberrations and Transcriptional Subtype Prediction

The TCGA cohort was randomly assigned into training set and test set in proportion of 1:1. In the training set, we first used gradient boosting decision tree (GBDT) (Friedman, 2001), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1997), random forest (RF) (Breiman, 2001), and extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016) to select relevant histopathological image features to reduce the risk of over-fitting. We used GBDT, RF, or XGBoost to select the top 15 features based on the ranking of importance. LASSO automatically shrunk the regression coefficients of irrelevant features to zero, and remained a variable number of features with non-zero coefficients for building classifier. Next, RF, GBDT, adaptive boosting (AdaBoost) (Collins et al., 2002), logistic regression (LR) (Collins et al., 2002), naive Bayes (NB) (Rish, 2001), support vector machine (SVM) (Cortes and Vapnik, 1995), decision tree (DT) (Safavian and Landgrebe, 1991), and K-nearest neighbor (KNN) (Weinberger and Saul, 2009) were applied to build models based on the selected features to predict genetic aberrations (*ALK*, *BRAF*, *EGFR*, *ROS1*) and transcriptional subtypes (PI, PP, TRU). We performed fivefold cross-validation on the training set to determine optimal feature combination and ensure the robustness of models. After feature selection and model construction on the training set, the predictive power of models was evaluated by the area under curve (AUC) of receiver operating characteristic (ROC) curve on the test set. True positive rate (TPR, sensitivity) refers to the proportion of predicted positive samples in all true positive samples. False positive rate (FPR, 1-specificity) is the proportion of samples that are incorrectly predicted as positive in all true negative samples.

Prognostic Model Establishment and Validation

1. Prognostic analysis of histopathological image features: By the median value of each feature, patients of training set were separated into high-value group and low-value group. Then, we conducted Cox regression analysis to estimate the prognostic effect of individual features on overall survival (OS), and calculated the hazard ratio (HR) and 95% confidence interval (CI). Kaplan-Meier survival curves and log-rank test estimated the survival outcomes of two groups. Prognostic value was significant if $p < 0.05$.

2. Data processing: Before modeling, we screened features of genomics and transcriptomics to reduce data dimension, while using all proteomics and histopathological image features. In the training set, the top 100 somatic mutations were included. To identity a limited subset of expressed genes from the whole transcriptomics profile, we selected the 100 most differently expressed genes (DEGs) between short-term (OS of 1–12 months at death) and long-term (OS $\geq$ 60 months) patients of training set in our models. DEGs between groups were identified by R package "DESeq2." Based on DEGs, we also analyzed the enriched gene function by Gene Ontology (GO) enrichment analysis on Metascape.[4]

3. Model development: We designed multiple data integration to build prognostic models, including one type of features (histopathological images, genomics, transcriptomics, proteomics), combinations of two types of features (histopathology + genomics, histopathology + transcriptomics, histopathology + proteomics), and all features (multi-omics). Input features were selected to establish models for predicting prognosis in the training set through RF (R package "randomForestSRC") with fivefold cross-validation. Survival risk for each patient was assessed by models, then patients were divided into high-risk and low-risk groups according to the median risk score. Time-dependent ROC curve, Kaplan-Meier method, and log-rank test evaluated the predictive capability, while decision curve analysis compared the clinical net benefit of each model. We further estimated the performance of these models in the test set. The model based on histopathological

[4]http://metascape.org

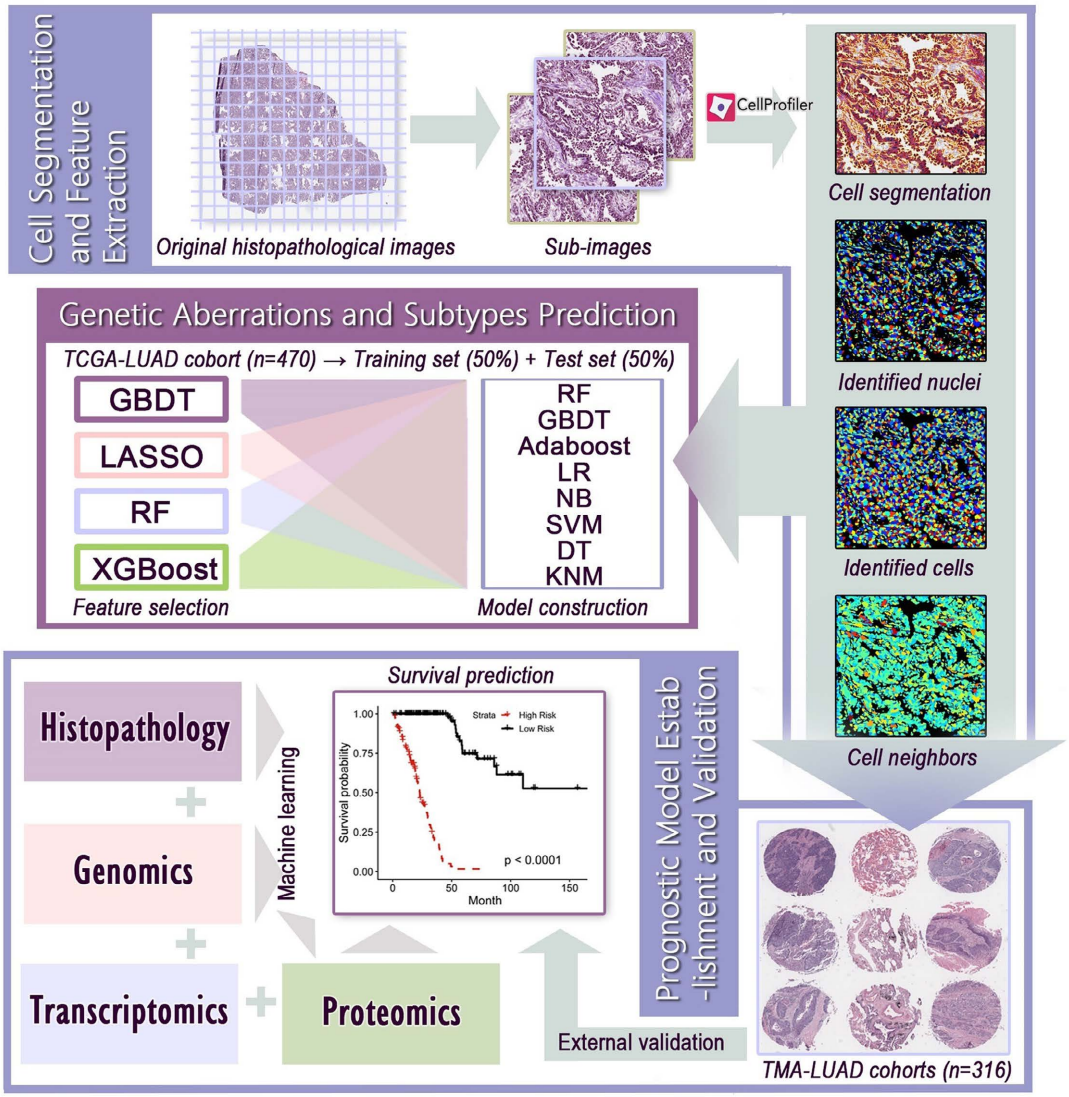


FIGURE 1 | The flowchart of histopathological image processing and data integration. (1) Whole-slide images were cropped into sub-images of 1,000 × 1,000 pixels, and 60 sub-images were randomly selected. CellProfiler was used to identify the nuclei and cells, and extract histopathological image features. (2) Image features and various machine learning algorithms were used to predict genetic aberrations and transcriptional subtypes in training and test sets. (3) Prognostic models were established using image features alone or integration of image features, genomics, transcriptomics, and proteomics, and the prognostic value was evaluated in test set or validation sets.

images was externally validated by the TMA-LUAD cohorts. Statistical analyses were performed using R 3.6.1.

RESULTS

Prediction of Genetic Aberrations and Transcriptional Subtypes

To assess whether machine learning can be trained to predict genetic aberrations and transcriptional subtypes using histopathological image features as input, we downloaded the related data from TCGA. The TCGA-LUAD cohort was then randomly divided into training set (n = 235) and test set (n = 235). The baseline clinical and molecular characteristics were not significantly different between the two sets (**Table 1**). To compare the feasibility of various machine learning methods, we applied four algorithms (GBDT, LASSO, RF, XGBoost) to select features and eight algorithms (RF, GBDT, AdaBoost, LR, NB, SVM, DT, KNN) to build models in the training set, which generated 32 combinations of two algorithms. The predictive performance of these models was estimated in the test set.

The results showed that the predictive models constructed by RF had the highest accuracy on the test set, regardless of the type of feature selection algorithm employed (**Figure 2A**). The models built only by RF were capable of predicting common gene aberrations in LUAD: *ALK* (AUC = 0.879), *BRAF* (AUC = 0.847),

TABLE 1 | Patient characteristics of the TCGA cohort.

Characteristics	TCGA-LUAD		p
	Training set (n = 235)	Test set (*n* = 235)	
Age: mean ± SD	65.1 ± 9.5	65.4 ± 10.5	0.740
Gender (%)			
Male	114 (48.5)	106 (45.1)	
Female	121 (51.5)	129 (54.9)	0.460
Tumor stage (%)			
I	132 (56.2)	122 (51.9)	
II	55 (23.4)	60 (25.5)	
III	34 (14.5)	42 (17.9)	
IV	13 (5.5)	10 (4.3)	
NA	1 (0.4)	1 (0.4)	0.605
Smoking status (%)			
Non-smoker	29 (12.3)	33 (14.0)	
Current smoker	53 (22.6)	56 (23.9)	
Former smoker	147 (63.6)	139 (59.1)	
NA	6 (2.6)	7 (3.0)	0.755
Survival status (%)			
Alive	141 (60.0)	160 (68.1)	
Deceased	94 (40.0)	75 (31.9)	0.068
***ALK* translocation (%)**			
−	106 (45.1)	91 (38.7)	
+	18 (7.7)	13 (5.5)	
NA	111 (47.2)	131 (55.7)	0.658
***BRAF* mutation (%)**			
−	219 (93.2)	225 (95.7)	
+	16 (6.8)	10 (4.3)	0.226
***EGFR* mutation (%)**			
−	210 (89.4)	200 (85.1)	
+	25 (10.6)	35 (14.9)	0.167
***ROS1* translocation (%)**			
−	228 (97.0)	219 (93.2)	
+	7 (3.0)	16 (6.8)	0.054
Transcriptional subtype (%)			
Proximal-inflammatory	32 (13.6)	39 (16.6)	
Proximal-proliferative	25 (10.6)	32 (13.6)	
Terminal respiratory unit	48 (20.4)	33 (14.0)	
NA	130 (55.3)	131 (55.7)	0.115

EGFR (AUC = 0.855), *ROS1* (AUC = 0.848), and transcriptional subtypes: PI (AUC = 0.897), PP (AUC = 0.861), and TRU (AUC = 0.894) (**Supplementary Table 1**). The combination of XGBoost and RF, XGBoost and GBDT, and GBDT and RF also performed well on the test set (**Supplementary Table 1**). Our analyses suggested that the genetic aberrations and transcriptional subtypes of LUAD could be predicted by histopathological image features with machine learning.

Evaluation and Validation of Image Features to Predict Survival

We assigned the training set into two groups by the median value of each histopathological image feature, and performed univariate Cox analysis to evaluate their prognostic value of overall survival (OS) (**Supplementary Table 2**). The results demonstrated that nine features may be able to predict OS (**Figure 2B**). With most of these features, the high-value group displayed a longer survival in Kaplan-Meier survival curve (**Supplementary Figure 2**), such as the four features with the lowest *p*-values: StDev_Cells_Texture_Correlation_3_0 (HR=0.65, 95% CI: 0.48–0.88, *p*=0.006), Granularity_16 (HR=0.66, 95% CI: 0.48–0.89, *p*=0.007), StDev_Cells_Granularity_16 (HR=0.69, 95% CI: 0.50–0.94, *p*=0.017), and Median_Cells_AreaShape_Zernike_8_2 (HR=0.70, 95% CI: 0.52–0.96, *p*=0.024).

To improve survival prediction with histopathological image features, we adopted random forest algorithm to select informative features from all image features and build prognostic model in the training set. We assessed the time-dependent

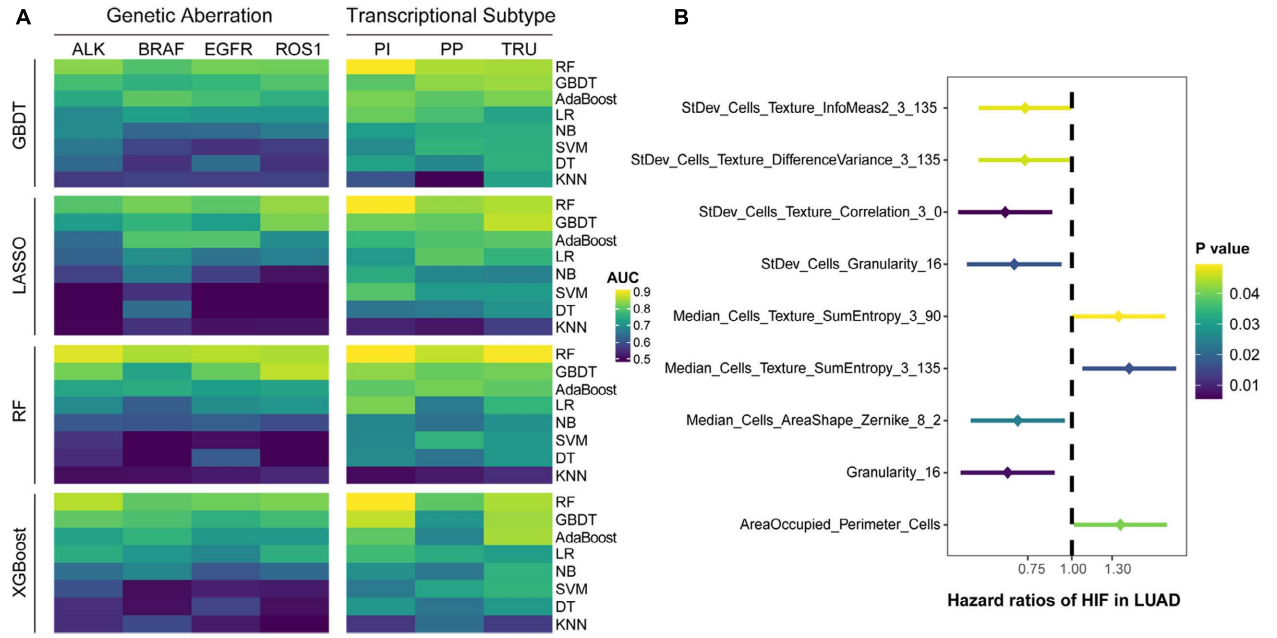


FIGURE 2 | Predictive performance of histopathological image features. **(A)** Four algorithms (GBDT, LASSO, RF, XGBoost) were utilized to select image features, and eight algorithms (RF, GBDT, AdaBoost, LR, DT, SVM, NB, KNN) were used for modeling in training set. The ability of models to predict genetic aberrations and transcriptional subtypes was assessed in test set. **(B)** Histopathological image features (HIF) with significant prognostic value ($p < 0.05$) in univariate Cox analysis.

ROC curves of the model in the test set (**Figure 3B**), which achieved good performance in predicting 1-year (AUC = 0.711), 3-year (AUC = 0.797), and 5-year OS (AUC = 0.825). Survival risk score calculated from the model divided patients into high-risk group (greater than median) and low-risk group (less than median). Compared with the individual features (**Supplementary Figure 2**), our model could better distinguish high-risk group from low-risk group in the test set (**Figure 3C**). High-risk patients had significant association with worse survival (HR = 4.94, 95% CI: 3.26–6.59, $p < 0.0001$).

In addition, the prognostic model based on image features was externally validated with the TMA-LUAD cohorts (**Table 2**). The AUC of predicting 5-year OS was 0.723 in validation set 1, 0.728 in validation set 2, 0.717 in validation set 3, and 0.743 in validation set 4 (**Figures 4A–D**). Although higher risk score was expected to be related to worse survival, this association was not strong in validation set 1 ($p = 0.077$; **Figure 4E**). However, there were significant survival differences in validation set 2 (HR = 1.64, 95% CI: 1.01–2.66, p = 0.044), validation set 3 (HR = 1.97, 95% CI: 1.11–3.52, $p = 0.021$), and validation set 4 (HR = 2.20, 95% CI: 1.17–3.08, $p = 0.010$; **Figures 4F–H**). **Supplementary Figure 3** presented several examples of histopathological images of high-risk and low-risk groups in the TCGA and TMA cohorts. In the TMA validation sets, patients with the same histological grade had different survival results. The subtle morphological differences between tumor cells of the same grade were not easily distinguished by visual evaluation, but could be quantified through cell segmentation and feature extraction on images. These results indicated the feasibility of histopathological image features for survival prediction in LUAD patients.

Integration of Image Features and Genomics for Survival Prediction

Genetic aberrations especially driver oncogenes were closely related to the genesis and development of tumors. Here, we estimated the prognostic role of genomics data, and attempted to integrate image features with genomics to enhance survival prediction. We selected 100 genes with the highest mutation frequency in the training set for modeling (**Supplementary Table 3** and **Figure 3A**). The process of data analysis and random forest method were the same as above. As shown in **Figure 3B**, the prognostic model of histopathological image features (HIF) performed better than the genomics model (G) on the test set. Integration of image features and genomics (HIF + G) greatly improved the AUCs of 1-year (0.795 vs. 0.637), 3-year (0.812 vs. 0.715), and 5-year OS (0.832 vs. 0.745) compared to genomics alone, indicating the complementary value of image features in prognostic models. According to the risk score of each model, we further stratified the test set into high-risk and low-risk sub-sets, and found that the integrative model had significant predictive ability for OS (HR = 6.14, 95% CI: 3.59–7.36, $p < 0.0001$; **Figure 3C**).

Integration of Image Features and Transcriptomics for Survival Prediction

Besides genomics analysis, the transcriptomics profile has become increasingly important for understanding the tumor molecular characteristics. Therefore, we combined histopathological image features and mRNA transcription

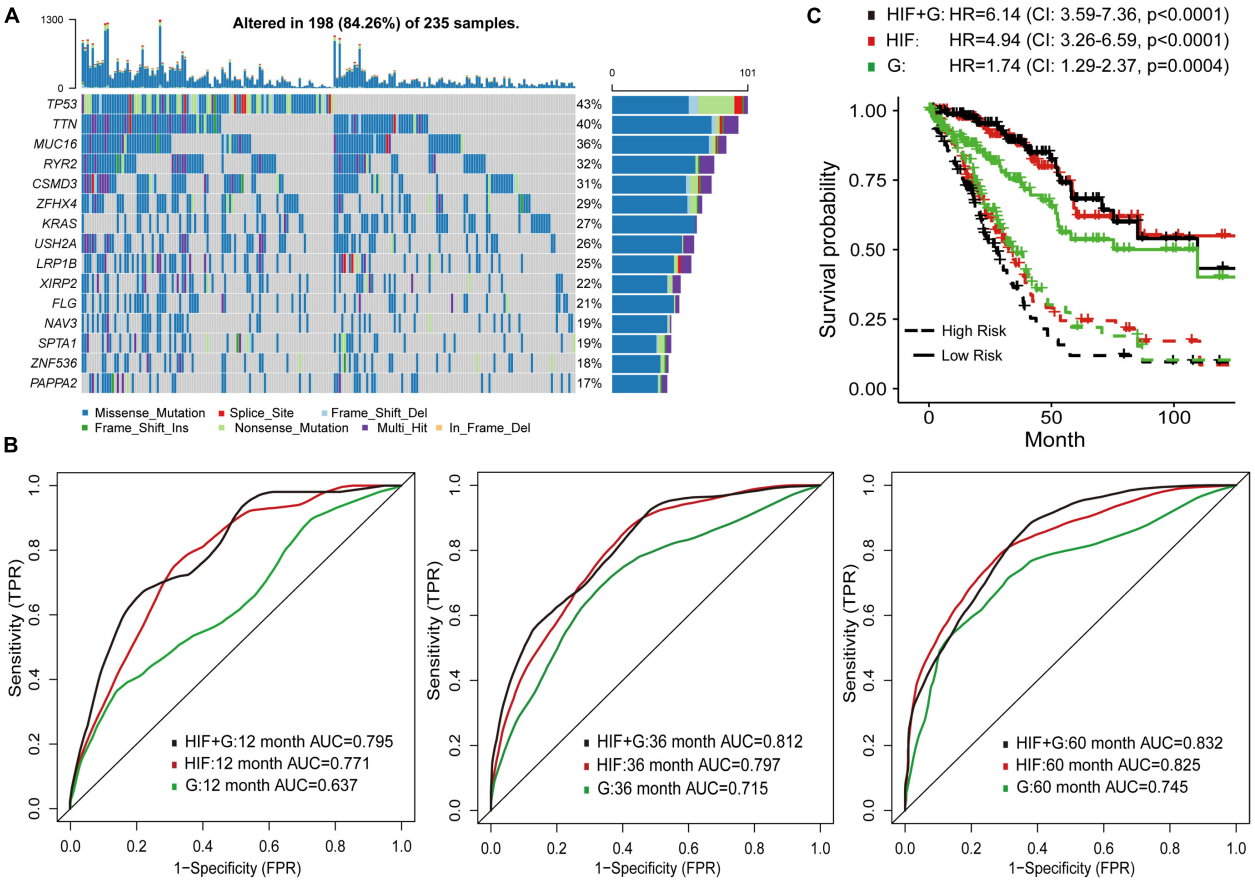


FIGURE 3 | Prognostic models based on histopathological image features (HIF) and genomics (G). **(A)** The 15 most common somatic mutations in training set. **(B)** Predictive power of HIF model, genomics model, and HIF+G model for 1-, 3-, and 5-year survival in test set. **(C)** Kaplan-Meier survival curves of high-risk and low-risk groups of test set predicted by the three models.

data to more accurately predict prognosis of LUAD patients. We analyzed the differently expressed genes (DEGs) between short-term (OS of 1–12 months at death, $n = 24$) and long-term (OS $\geq$ 60 months, $n = 29$) patients of training set. GO enrichment analysis was next conducted to show the regulatory functions of DEGs. The GO enrichment terms were most abundant in cell differentiation, such as the differentiation of epithelial cells, cardiac progenitors, and neurons. These genes were also associated with functions of thrombosis, anticoagulation, and blood circulation (**Figure 5A**).

Next, the top 100 DEGs were involved in the modeling process (**Supplementary Table 4**). In the test set, the transcriptomics model (RNA) reached higher prognostic accuracy of 1-year OS (AUC = 0.801), but lower accuracy of 3-year (AUC = 0.792), and 5-year OS (AUC = 0.786) than the histopathological image feature (HIF) model (**Figure 5B**). The integrative model of image features and transcriptomics (HIF + RNA) achieved superior prediction results (1-year AUC = 0.819, 3-year AUC = 0.837, 5-year AUC = 0.840) than the other two models. Based on the HIF + RNA model, there was remarkable difference in OS between high-risk and low-risk patients of the test set (HR = 7.34, 95% CI: 4.34–9.24, $p < 0.0001$; **Figure 5C**).

Integration of Image Features and Proteomics for Survival Prediction

The proteomics characteristics can reflect the properties and mechanisms of tumor, and offer an opportunity for better prognosis prediction. Therefore, the expression levels of 218 proteins were used to establish the prognostic models (**Supplementary Table 5**). The proteomics model (P) performed comparably to the histopathological image feature model (HIF) in predicting survival of test set patients (**Figures 6A–C**). In addition, improved predictive performance was observed when the model was built by combining image features and proteomics data (HIF+P), which obtained AUC of 0.825, 0.845, and 0.850 for 1-, 3-, and 5-year OS, respectively. In the test set, the high-risk and low-risk patients predicted by these models displayed different survival outcomes (**Figure 6D**), especially the HIF + P model (HR = 10.99, 95% CI: 7.75–18.57, $p < 0.0001$).

Integration of Image Features and Multiple Omics for Survival Prediction

The above results demonstrated that the combination of histopathological image features and single omics (genomics,

TABLE 2 | Patient characteristics of the TMA cohorts.

Characteristics	TMA-LUAD			
	Validation set 1	**Validation set 2**	**Validation set 3**	**Validation set 4**
No. of patients	69	72	86	89
Microarray number	HLugA150Su01	HLugA150Su02	HLugA180Su04	HLugA180Su05
Surgery time	2004.7–2007.8	2007.9–2009.6	2008.1–2013.7	2004.7–2009.6
Last follow-up	2012.7	2012.7	2016.6	2014.8
Age: mean ± SD	58.6 ± 12.2	59.7 ± 10.7	62.8 ± 10.0	62.5 ± 9.9
Gender (%)				
Male	35 (50.7)	39 (54.2)	47 (54.7)	48 (53.9)
Female	34 (49.3)	33 (45.8)	39 (45.3)	41 (46.1)
Tumor stage (%)				
I	30 (43.5)	23 (31.9)	23 (26.7)	27 (30.3)
II	10 (14.5)	14 (19.4)	14 (16.3)	16 (18.0)
III	16 (23.2)	18 (25.0)	18 (20.9)	28 (31.5)
IV	3 (4.3)	0 (0.0)	2 (2.3)	1 (1.1)
NA	10 (14.5)	17 (23.6)	29 (33.7)	17 (19.1)
Histological grade (%)				
G1	10 (14.5)	6 (8.3)	0 (0.0)	3 (3.4)
G1-2	3 (4.3)	5 (6.9)	9 (10.5)	4 (4.5)
G2	37 (53.6)	45 (62.5)	48 (55.8)	55 (61.8)
G2-3	9 (13.0)	3 (4.2)	23 (26.7)	16 (18.0)
G3	10 (14.5)	13 (18.1)	6 (7.0)	11 (12.4)
Survival status (%)				
Alive	34 (49.3)	49 (68.1)	38 (44.2)	21 (23.6)
Deceased	35 (50.7)	23 (31.9)	48 (55.8)	68 (76.4)

transcriptomics, proteomics) provided a greater power to predict survival of LUAD patients. Accordingly, we investigated whether incorporating all omics and histopathological image features would cause further optimization. In the test set, the multi-omics model further enhanced 1-year AUC to 0.899, 3-year AUC to 0.938, and 5-year AUC to 0.908 for predicting OS (**Figure 6E**). Kaplan-Meier curves and Cox regression analysis showed that the predicted high-risk group of test set was remarkably related to worse prognosis (HR = 19.98, 95% CI: 13.96–26.81, $p < 0.0001$, **Figure 6F**). We also performed decision curve analysis of models on the test set, and the multi-omics model obtained higher net benefit than other models in clinical decision-making (**Figure 6G**).

DISCUSSION

Considering the importance of histopathological images in cancer prognosis, we proposed an image processing and data analysis workflow to extract quantitative features from histopathological images, and to develop predictive models for molecular characteristics and survival outcomes of LUAD patients using histopathological image features and machine learning methods. Moreover, we integrated image features and multiple omics information (i.e., somatic mutation, gene transcription, and protein expression) for survival prediction. From test analyses, the results showed that image features had significant value in predicting the genetic aberrations and transcriptional subtypes of LUAD. Survival prediction based on the integrative models became more precise than using image features or omics data alone, which indicated that image features and omics data may be complementary in prognosis prediction. The integrative models might better stratify high-risk and low-risk patients, which was beneficial for the treatment of LUAD patients.

The potential differences of molecular expression in tumors tend to present as the alterations of tissue structure and nuclear morphology (Madabhushi and Lee, 2016; Ni et al., 2021). Previously, it has been found that *STK11*, *EGFR*, *FAT1*, *SETBP1*, *KRAS*, and *TP53* mutations of LUAD could be predicted from histopathological images using deep learning (AUCs ranged from 0.733 to 0.856) (Coudray et al., 2018). Afterward, several studies also suggested that histopathological image analysis could detect the gene mutation status in liver, colorectal, and ovarian cancers (Chen et al., 2020; Jang et al., 2020; Zeng et al., 2021). In this study, although we also focused on the mutation prediction, we conducted multiple combinations of machine learning methods, and the random forest models had high prediction accuracy for *ALK* (AUC = 0.879), *BRAF* (AUC = 0.847), *EGFR* (AUC = 0.855), and *ROS1* (AUC = 0.848) aberrations. In our previous radiomics study of lung cancer, random forest also had good classification accuracy for primary and metastatic lung lesions (Zhou et al., 2021). Meanwhile, the models were trained to predict the transcriptional subtypes of LUAD for the first time, with AUCs from 0.861 to 0.897. The PP subtype had more *KRAS*-mutated and *STK11*-inactivated tumors. The PI subtype was enriched

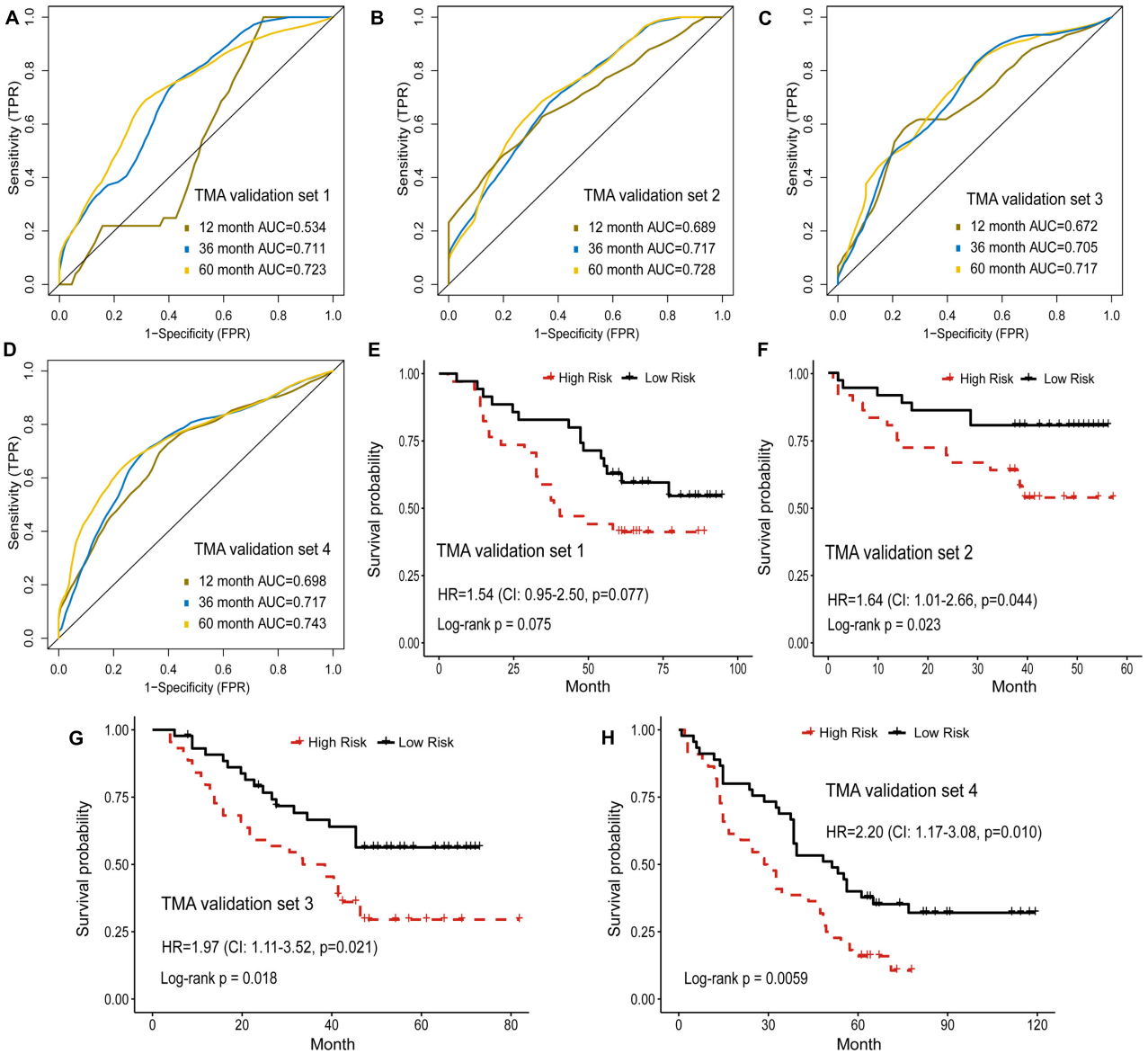


FIGURE 4 | Validation of histopathological image features for survival prediction. **(A–D)** The 1-, 3-, and 5-year time-dependent ROC curves of the model built by histopathological image features in four validation sets. **(E–H)** Kaplan-Meier survival curves of validation sets analyzed with the image feature model.

for solid histopathology, and co-mutation of *NF1* and *TP53*. Finally, the TRU subtype was characterized by *EGFR* mutation, kinase fusion expression, and better prognosis (Cancer Genome Atlas Research Network, 2014). Our finding indicated that gene mutation and expression may affect the patterns of tumor cells on histopathological images. Histopathological image features could be applied as a convenient and cost-effective approach to predict these important genetic aberrations and transcriptional subtypes in LUAD.

We evaluated the prognostic value of histopathological image features identified by computational recognition modules, including anatomical characteristics (e.g., area and shape) and patterns (e.g., correlation and neighborship) of tumor cells. The Zernike shape features might predict survival outcomes in LUAD, which was consistent with previous studies (Yu et al., 2016; Luo et al., 2017). Texture features may also be related to prognosis, such as correlation, granularity, and sum entropy. However, the prognostic power of individual image features was limited and controversial; thus, we combined image features by random forest method to provide better survival prediction. Compared to the previous researches using publicly available datasets in lung cancer (Yu et al., 2016; Luo et al., 2017; Wang et al., 2018), we extended the verification of prognostic model with four TMA validation sets, besides the internal test set. Our model successfully predicted survival results in three validation sets, which proved that the random forest model had

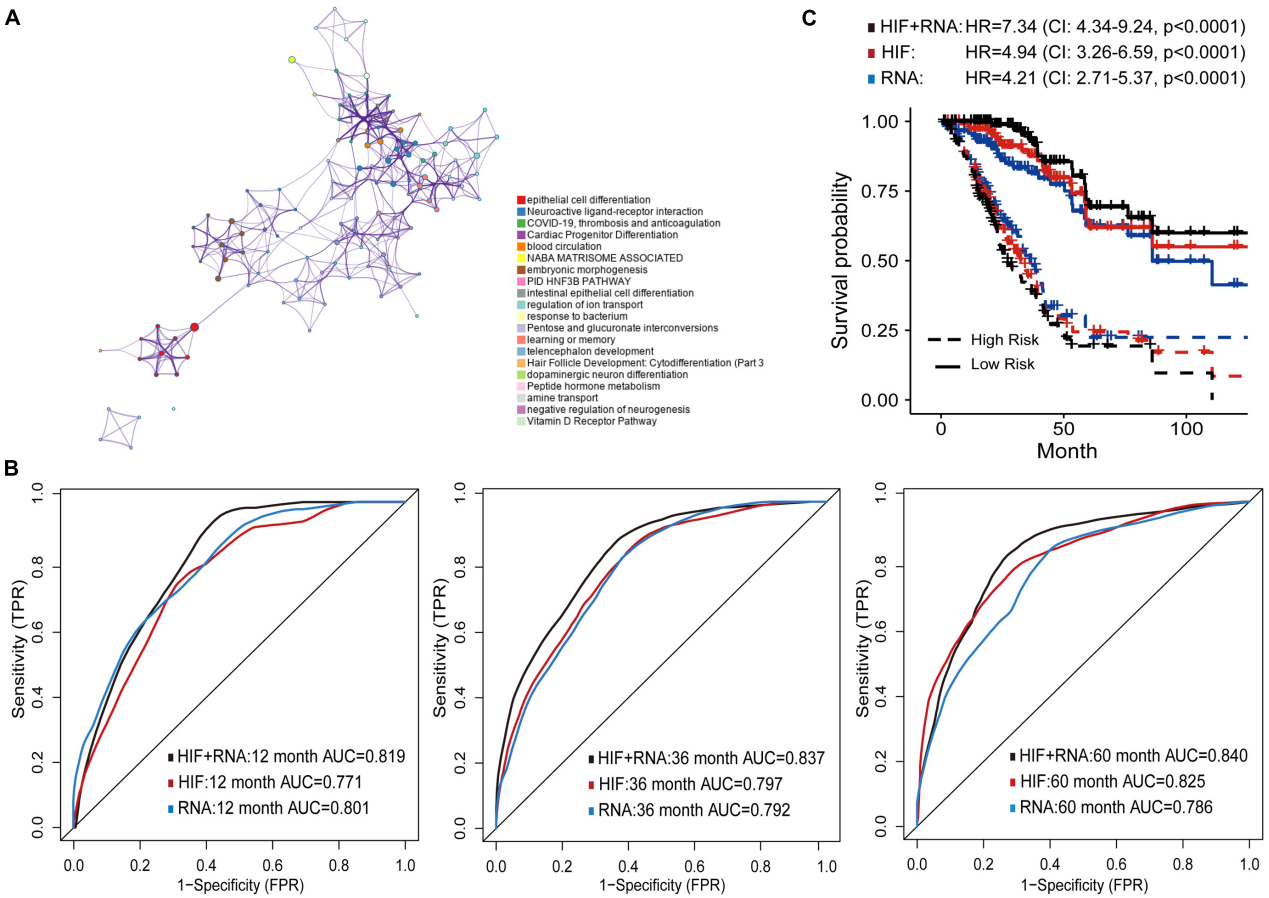


FIGURE 5 | Prognostic models based on histopathological image features (HIF) and transcriptomics (RNA). **(A)** Gene Ontology enrichment network of differently expressed genes. The most enriched term was used to describe each cluster (see legends). **(B)** Area under the time-dependent ROC curves of HIF model, transcriptomics model, and HIF + RNA model in test set. **(C)** Kaplan-Meier survival curves of the three models in test set.

certain robustness and generalizability across external cohorts. The prognostic ability of model was weak in validation set 1, possibly because of the heterogeneity of patients collected from different institutions. Moreover, some confounding factors that potentially affect prognosis were unknown in TMA datasets, such as smoking history, complication, and adjuvant therapy. It also suggested that our model still needed to be improved. Further training on large samples was required to adapt to the heterogeneous histopathological images of different populations.

Since the tumor progression is the result of complex biological processes, multiple molecular levels of data can describe more characteristics of tumors, which may contribute to the prognosis evaluation and therapeutic intervention (Gallo Cantafio et al., 2018). Several studies have combined histopathological image features with gene transcription data to improve survival prediction in LUAD (Zhu et al., 2016; Yu et al., 2017). The difference was that our study also involved the analysis of somatic mutation and protein expression of LUAD. Moreover, the time-dependent ROC curve was used to calculate dynamic AUC value of whole survival time (Kamarudin et al., 2017), rather than simply dividing the patients into short-term or long-term survivors. We founded that the proteomics model had higher accuracy than the models based on gene mutation or transcription, possibly because proteins were the functional executors of the cells. Proteomic analysis also showed the prognostic significance of expressed proteins and proteomics clustering in LUAD (Xu et al., 2020). Afterward, we developed and verified the prognostic models based on diverse feature integration of genomics, transcriptomics, proteomics, and histopathological images. The prognostic performance of the models using image features and single omics outperformed than the models using only one type of data. The model integrating multi-omics with image features achieved the best performance, which may contribute significantly to personalized risk stratification for LUAD patients.

There were some limitations in this study. Firstly, for genetic aberrations and subtype prediction, the small number of positive cases limited the accuracy; thus, it was necessary to expand the study sample in the future. Secondly, although we validated the prognostic model by four TMA validation sets, these validation sets were less diverse due to the TMAs processed by the same

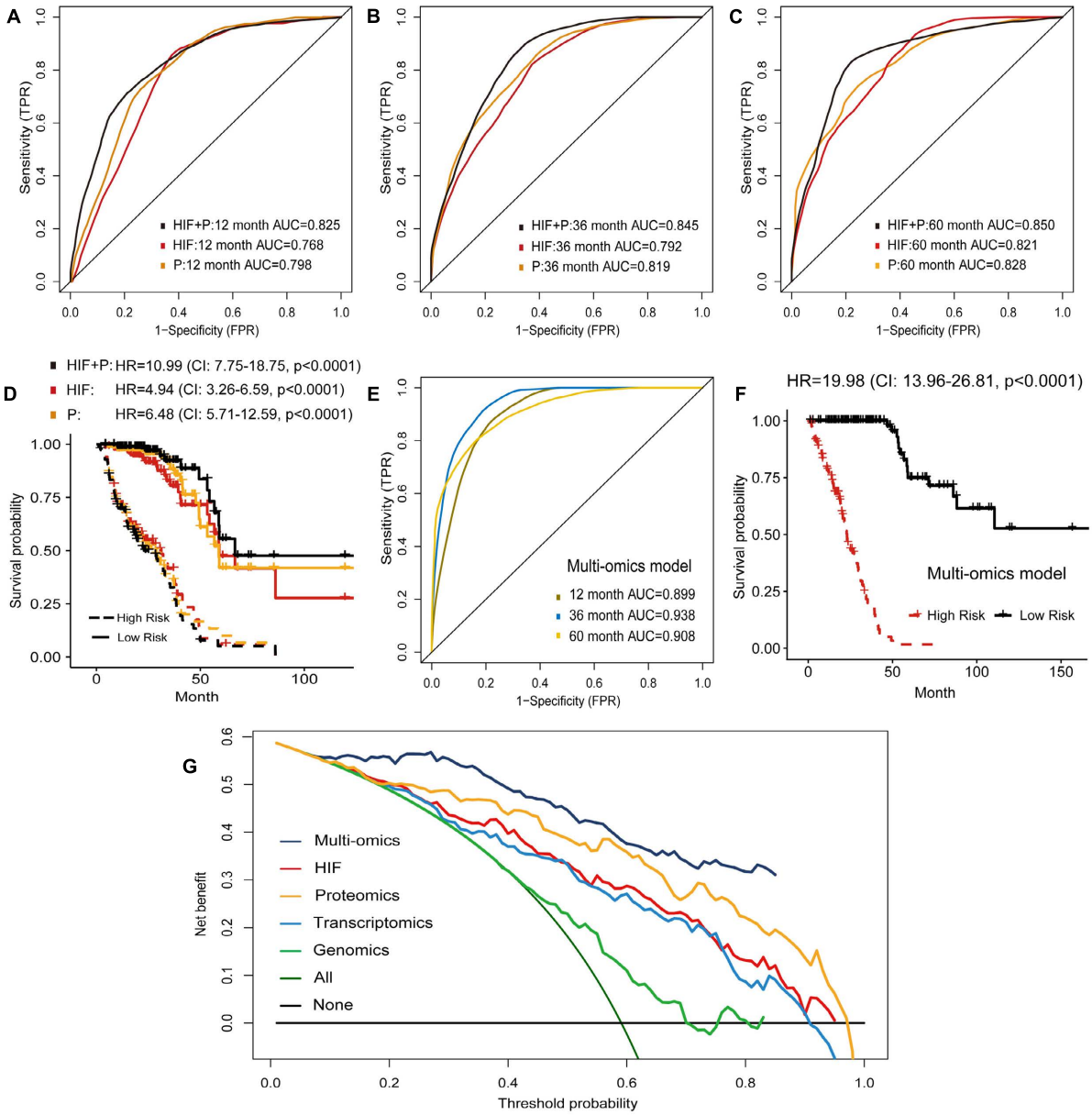


FIGURE 6 | Prognostic models based on histopathological image features (HIF), proteomics (P), and multi-omics. **(A–C,E)** Predictive performance of HIF model, proteomics model, HIF + P model, and multi-omics model in test set. **(D,F)** Kaplan-Meier survival curves of high-risk and low-risk patients in test set based on these models. **(G)** Decision curve analysis. The horizontal black line was net benefit of treating no patient, and oblique green line was net benefit of treating all patients. The net benefit of treating patients according to the multi-omics model was highest when threshold probability was greater than 10%.

institution. The cases from TCGA and TMA datasets may exist potential biases, because representative images were more likely to be chosen. Therefore, the application of the prognostic model in actual clinical practice needed further investigation. In addition, the retrospective TMA datasets lacked genetic data; thus, the efficacy and generalizability of integrative prognostic models remained to be validated. Furthermore, the threshold of risk score was simply based on the median value, and the large-scale research would more rigorously determine the threshold to optimize patient stratification.

CONCLUSION

In conclusion, our findings demonstrated that the machine learning models based on histopathological image features had great potential to predict genetic aberrations, transcriptional subtypes, and survival outcomes in patients with LUAD. In addition, the workflow of integrating histopathological image features, genomics, transcriptomics, and proteomics to develop models may improve survival prediction and benefit the precision medicine of LUAD patients.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the National Human Genetic Resources Sharing Service Platform (2005DKA21300). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XM, LC, and HZ contributed to the conception and design of the study. LC and HZ performed the data analysis, interpretation, and manuscript drafting. YX, YH, and YL provided and assessed the clinicopathological data. All authors contributed to manuscript revision and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.720110/full#supplementary-material

Supplementary Figure 1 | Examples of whole-slide histopathological images and part of selected sub-images.

Supplementary Figure 2 | Kaplan-Meier analysis of groups with high-value and low-value image features.

Supplementary Figure 3 | Histopathological image examples of high-risk and low-risk patients. Survival risk score was estimated by the model of histopathological image features, and then patients in the TCGA and TMA cohorts were divided into high-risk and low-risk groups by median risk score. The sub-images were processed with cell segmentation.

Supplementary Table 1 | Prediction of genetic aberrations and transcriptional subtypes by machine learning algorithms in test set.

Supplementary Table 2 | Hazard ratio of histopathological image features in univariate Cox analysis.

Supplementary Table 3 | Gene mutation frequency in training set.

Supplementary Table 4 | Differently expressed genes in groups with short or long survival.

Supplementary Table 5 | Included proteins.

REFERENCES

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Calvayrac, O., Pradines, A., Pons, E., Mazières, J., and Guibert, N. (2017). Molecular biomarkers for lung adenocarcinoma. *Eur. Respir. J.* 49:1601734. doi: 10.1183/13993003.01734-2016

Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550. doi: 10.1038/nature13385

Chen, M., Zhang, B., Topatana, W., Cao, J., Zhu, H., Juengpanich, S., et al. (2020). Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis. Oncol.* 4:14. doi: 10.1038/s41698-020-0120-3

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY: Association for Computing Machinery), 785–794. doi: 10.1145/2939672.2939785

Cheng, J., Zhang, J., Han, Y., Wang, X., Ye, X., Meng, Y., et al. (2017). Integrative Analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer Res.* 77, e91–e100. doi: 10.1158/0008-5472.CAN-17-0313

Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic regression, adaboost and Bregman distances. *Mach. Learn.* 48, 253–285. doi: 10.1023/A:1013912006537

Cortes, C., and Vapnik, V. (1995). Support vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24, 1559–1567. doi: 10.1038/s41591-018-0177-5

Ettinger, D. S., Wood, D. E., Aggarwal, C., Aisner, D. L., Akerley, W., Bauman, J. R., et al. (2019). NCCN guidelines insights: non-small cell lung cancer, version 1.2020. *J. Natl. Compr. Canc. Netw.* 17, 1464–1472. doi: 10.6004/jnccn.2019.0059

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Gallo Cantafio, M. E., Grillone, K., Caracciolo, D., Scionti, F., Arbitrio, M., Barbieri, V., et al. (2018). From single level analysis to multi-omics integrative approaches: a powerful strategy towards the precision oncology. *High Throughput* 7:33. doi: 10.3390/ht7040033

Gillette, M. A., Satpathy, S., Cao, S., Dhanasekaran, S. M., Vasaikar, S. V., Krug, K., et al. (2020). Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* 182, 200–225.e35. doi: 10.1016/j.cell.2020.06.013

Goode, A., Gilbert, B., Harkes, J., Jukic, D., and Satyanarayanan, M. (2013). OpenSlide: a vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* 4:27. doi: 10.4103/2153-3539.119005

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature* 553, 446–454. doi: 10.1038/nature25183

Jang, H. J., Lee, A., Kang, J., Song, I. H., and Lee, S. H. (2020). Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World J. Gastroenterol.* 26, 6207–6223. doi: 10.3748/wjg.v26.i40.6207

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* 17:53. doi: 10.1186/s12874-017-0332-6

Lee, G., Veltri, R. W., Zhu, G., Ali, S., Epstein, J. I., and Madabhushi, A. (2017). Nuclear shape and architecture in benign fields predict biochemical recurrence in prostate cancer patients following radical prostatectomy: preliminary findings. *Eur. Urol. Focus* 3, 457–466. doi: 10.1016/j.euf.2016.05.009

Luo, X., Zang, X., Yang, L., Huang, J., Liang, F., Rodriguez-Canales, J., et al. (2017). Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J. Thorac. Oncol.* 12, 501–509. doi: 10.1016/j.jtho.2016.10.017

Madabhushi, A., and Lee, G. (2016). Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. Image Anal.* 33, 170–175. doi: 10.1016/j.media.2016.06.037

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., et al. (2018). CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol.* 16:e2005970. doi: 10.1371/journal.pbio.2005970

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., et al. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. U.S.A.* 115, E2970–E2979. doi: 10.1073/pnas.1717139115

Ni, Y., Zhou, X., Yang, J., Shi, H., Li, H., Zhao, X., et al. (2021). The role of tumor-stroma interactions in drug resistance within tumor microenvironment. *Front. Cell Dev. Biol.* 9:637675. doi: 10.3389/fcell.2021.637675

Niazi, M. K. K., Parwani, A. V., and Gurcan, M. N. (2019). Digital pathology and artificial intelligence. *Lancet Oncol.* 20, e253–e261. doi: 10.1016/S1470-2045(19)30154-8

Pagliarini, R., Shao, W., and Sellers, W. R. (2015). Oncogene addiction: pathways of therapeutic response, resistance, and road maps toward a cure. *EMBO Rep.* 16, 280–296. doi: 10.15252/embr.201439949

Park, H., Sholl, L. M., Hatabu, H., Awad, M. M., and Nishino, M. (2019). Imaging of precision therapy for lung cancer: current state of the art. *Radiology* 293, 15–29. doi: 10.1148/radiol.2019190173

Rish, I. (2001). "An empirical study of the Naive Bayes classifier," in *Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, (New York, NY: IBM), 41–46.

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man. Cybern.* 21, 660–674. doi: 10.1109/21.97458

Saito, M., Shiraishi, K., Kunitoh, H., Takenoshita, S., Yokota, J., and Kohno, T. (2016). Gene aberrations for precision medicine against lung adenocarcinoma. *Cancer Sci.* 107, 713–720. doi: 10.1111/cas.12941

Schabath, M. B., and Cote, M. L. (2019). Cancer progress and priorities: lung cancer. *Cancer Epidemiol. Biomarkers Prev.* 28, 1563–1579. doi: 10.1158/1055-9965.EPI-19-0221

Tabesh, A., Teverovskiy, M., Pang, H. Y., Kumar, V. P., Verbel, D., Kotsianti, A., et al. (2007). Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans. Med. Imaging* 26, 1366–1378. doi: 10.1109/TMI.2007.898536

Thakur, M. K., and Gadgeel, S. M. (2016). Predictive and prognostic biomarkers in non-small cell lung cancer. *Semin. Respir. Crit. Care Med.* 37, 760–770. doi: 10.1055/s-0036-1592337

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Turkki, R., Byckhov, D., Lundin, M., Isola, J., Nordling, S., Kovanen, P. E., et al. (2019). Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast Cancer Res. Treat.* 177, 41–52. doi: 10.1007/s10549-019-05281-1

Vargas, A. J., and Harris, C. C. (2016). Biomarker development in the precision medicine era: lung cancer as a case study. *Nat. Rev. Cancer* 16, 525–537. doi: 10.1038/nrc.2016.56

Wang, S., Chen, A., Yang, L., Cai, L., Xie, Y., Fujimoto, J., et al. (2018). Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Sci. Rep.* 8:10393. doi: 10.1038/s41598-018-27707-4

Weinberger, K. Q., and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 10, 207–244.

Xu, J. Y., Zhang, C., Wang, X., Zhai, L., Ma, Y., Mao, Y., et al. (2020). Integrative proteomic characterization of human lung adenocarcinoma. *Cell* 182, 245–261.e17. doi: 10.1016/j.cell.2020.05.043

Yu, K. H., Berry, G. J., Rubin, D. L., Ré, C., Altman, R. B., and Snyder, M. (2017). Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell Syst.* 5, 620–627.e3. doi: 10.1016/j.cels.2017.10.014

Yu, K. H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., et al. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* 7:12474. doi: 10.1038/ncomms12474

Zeng, H., Chen, L., Huang, Y., Luo, Y., and Ma, X. (2020). Integrative models of histopathological image features and omics data predict survival in head and neck squamous cell carcinoma. *Front. Cell Dev. Biol.* 8:553099. doi: 10.3389/fcell.2020.553099

Zeng, H., Chen, L., Zhang, M., Luo, Y., and Ma, X. (2021). Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer. *Gynecol. Oncol.* doi: 10.1016/j.ygyno.2021.07.015 [Epub ahead of print].

Zhang, X., Xing, F., Su, H., Yang, L., and Zhang, S. (2015). High-throughput histopathological image analysis via robust cell segmentation and hashing. *Med. Image Anal.* 26, 306–315. doi: 10.1016/j.media.2015.10.005

Zhou, Y., Ma, X. L., Zhang, T., Wang, J., Zhang, T., and Tian, R. (2021). Use of radiomics based on 18F-FDG PET/CT and machine learning methods to aid clinical decision-making in the classification of solitary pulmonary lesions: an innovative approach. *Eur. J. Nucl. Med. Mol. Imaging* 48, 2904–2913. doi: 10.1007/s00259-021-05220-7

Zhu, X., Yao, J., Luo, X., Xiao, G., Xie, Y., Gazdar, A., et al. (2016). "Lung cancer survival prediction from pathological images and genetic data — an integration study," in *Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, (Prague), 1173–1176. doi: 10.1109/ISBI.2016.7493475

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chen, Zeng, Xiang, Huang, Luo and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 11 October 2021
doi: 10.3389/fcell.2021.728242

Identification of Key eRNAs for Spinal Cord Injury by Integrated Multinomial Bioinformatics Analysis

Runzhi Huang[1,2,3†], Siqiao Wang[3†], Rui Zhu[1,2,3†], Shuyuan Xian[3], Zongqiang Huang[4], Liming Cheng[1,2*] and Jie Zhang[1,2*]*

[1] Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration (Tongji University), Ministry of Education, Shanghai, China, [2] Division of Spine Surgery, Department of Orthopedics, Tongji Hospital, Tongji University School of Medicine, Shanghai, China, [3] Tongji University School of Medicine, Shanghai, China, [4] Department of Orthopedics, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

OPEN ACCESS

Edited by:
Chuan-Xing Li,
Karolinska Institutet (KI), Sweden

Reviewed by:
Zhi-an Cheng,
Guangdong Provincial Hospital of Chinese Medicine, China
Lixin Cheng,
Jinan University, China

***Correspondence:**
Zongqiang Huang
gzhuangzq@163.com
Liming Cheng
limingcheng@tongji.edu.cn
Jie Zhang
jiezhang@tongji.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 21 June 2021
Accepted: 25 August 2021
Published: 11 October 2021

Citation:
Huang R, Wang S, Zhu R, Xian S, Huang Z, Cheng L and Zhang J (2021) Identification of Key eRNAs for Spinal Cord Injury by Integrated Multinomial Bioinformatics Analysis. Front. Cell Dev. Biol. 9:728242. doi: 10.3389/fcell.2021.728242

Background: Spinal cord injury (SCI) is a severe neurological deficit affecting both young and older people worldwide. The potential role of key enhancer RNAs (eRNAs) in SCI remains elusive, which is a prominent challenge in the trauma repair process. This study aims to investigate the roles of key eRNAs, transcription factors (TFs), signaling pathways, and small-molecule inhibitors in SCI using multi-omics bioinformatics analysis.

Methods: Microarray data of peripheral blood mononuclear cell (PBMC) samples from 27 healthy volunteers and 25 chronic-phase SCI patients were retrieved from the Gene Expression Omnibus database. Differentially expressed transcription factors (DETFs), differentially expressed enhancer RNAs (DEeRNAs), and differentially expressed target genes (DETGs) were identified using the Linear Models for Microarray Data (limma) package. Fraction of immune cells was estimated using CIBERSORT algorithm. Gene Set Variation Analysis (GSVA) was applied to identify the downstream signaling pathways. The eRNA regulatory network was constructed based on the correlation results. Connectivity Map (CMap) database was used to find potential drugs for SCI patients. The cellular communication analysis was performed to explore the molecular regulation mechanism of SCI based on single-cell RNA sequencing (scRNA-seq) data. Chromatin immunoprecipitation sequencing (ChIP-seq) and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) data were used to validate the key regulatory mechanisms. scRNA-seq dataset was used to validate the cell subtype localization of the key eRNAs.

Results: In total, 21 DETFs, 24 DEeRNAs, and 829 DETGs were identified. A regulatory network of 13 DETFs, six DEeRNAs, seven DETGs, two hallmark pathways, two immune cells, and six immune pathways was constructed. The link of Splicing factor proline and glutamine rich (SFPQ) (TF) and vesicular overexpressed in cancer prosurvival protein 1 (VOPP1) (eRNA) ($R = 0.990$, $p < 0.001$, positive), VOPP1 (eRNA) and epidermal growth factor receptor (EGFR) (target gene) ($R = 0.974$, $p < 0.001$, positive), VOPP1, and T helper (Th) cells ($R = -0.987$, $p < 0.001$, negative), and VOPP1 and hallmark

coagulation ($R = 0.937$, $p < 0.001$, positive) was selected. Trichostatin A was considered the best compound target to SCI-related eRNAs (specificity = 0.471, $p < 0.001$).

Conclusion: VOPP1, upregulated by SFPQ, strengthened the transient expression of EGFR. Th cells and coagulation were the potential downstream pathways of VOPP1. This regulatory network and potential inhibitors provide novel diagnostic biomarkers and therapeutic targets for SCI.

Keywords: spinal cord injury, multi-omics bioinformatics analysis, biomarkers, eRNA regulatory network, therapeutic targets

INTRODUCTION

Spinal cord injury (SCI) refers to functional or structural injury of the spinal cord causing total or partial loss of motor, sensory, and sphincter function below the injured segment (Nakae et al., 2011; Anderson et al., 2018). Whether the pathogenesis is disease or trauma, SCI exhibits high disability rates. It affects approximately 347,000 individuals in the United States, with about 17,500 new cases diagnosed every year (Badhiwala et al., 2019; GBD 2016 Neurology Collaborators, 2019). According to the illness course, SCI can be divided in three phases: the acute phase (0–15 days), the subacute phase (3–5 months), and the chronic phase (6–12 months) (Pouw et al., 2011). The most common pathological characteristics of SCI are a cascade of molecular and cellular events triggered by inflammation, and excitotoxicity impairs endogenous regeneration, namely, axonal outgrowth and remyelination.

It is arduous to repair injured neurons and restore conducting function of axons, so treatments of SCI have become worldwide problems (Borton et al., 2014). Importantly, there are no efficacious drugs or therapeutic approaches for SCI. Thus, many patients suffer substantial physical and psychological consequences. To date, the molecular mechanisms of SCI remain unclear, so it is difficult to develop novel drugs or treatments. Thus, it is urgent to determine specific molecular mechanisms that underlie the pathogenesis of SCI.

The gene mutation and aberrant expression are implicated in the development of numerous diseases and may involve various types of regulatory molecules. Enhancer refers to a kind of distal regulatory DNA element that is able to enhance the transcription of corresponding target genes *via* coordinating with target gene promoters (Blackwood and Kadonaga, 1998). Widely acknowledged as DNA elements that nucleate transcription factor (TF) binding, enhancers were identified to also transcribe non-coding RNAs recently, which are referred to as enhancer RNAs (eRNAs) (Kim et al., 2010). Numerous eRNAs have been identified in *Homo sapiens* so far, and many of them were suggested to play crucial parts in the transcriptional circuitry (Li et al., 2016).

Inflammation after SCI is a fundamental basis of secondary damage. Inflammatory response is coordinated by various signaling modalities that include the epigenetic modification of promoters and eRNAs (Rudman et al., 2018). Large amounts of cytokines are quickly generated and released from various cells in the damaged structures after SCI (Hayashi et al., 2000; Pineau and Lacroix, 2007). Cytokines activate apoptosis of pathological nerve cells and recruitment of active leukocytes in myelopathic lesions (Beck et al., 2010; Chen et al., 2011). Leukocyte infiltration contributes to cascaded amplification of inflammation cytokine signaling, increasing neurotoxicity, and thus promoting fibrotic scar formation *via* recruiting fibroblasts in damaged regions (Zhu et al., 2015). In addition, damaged areas chronically maintain pro-inflammatory phenotype after injury, hindering regeneration and healing of the injured spinal cord (Beck et al., 2010). Because of the detrimental inflammatory effects after SCI, reducing inflammatory response is a major demand for improving clinical outcomes of SCI patients. Nevertheless, eRNA-targeted anti-inflammation agents that effectively inhibit inflammation and improve the prognosis of SCI patients are still insufficient, and novel anti-inflammation interventions are urgently needed.

In this study, transcription factors (TFs), differentially expressed enhancer RNAs (DEeRNAs), and target genes in patients with SCI were identified. Significant immune cells and immune-related pathways were identified by cell type identification by estimating relative subsets of RNA transcripts (CIBERSORT) and single-sample Gene Set Enrichment Analysis (ssGSEA) algorithms, respectively. In addition, TFs, eRNAs, target genes, immune cells, immune-related gene sets, hallmark gene sets (signaling pathways) acquired by the Gene Set Variation Analysis (GSVA) were merged in correlation analysis. Only remarkable interactions were extracted for construction of eRNA regulatory network. Moreover, we performed Connectivity Map (CMap) analysis to explore the potential inhibitors specific to SCI-related eRNAs. Furthermore, chromatin immunoprecipitation sequencing (ChIP-seq) and Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq) data were utilized for validation of the key regulation mechanisms in this network.

MATERIALS AND METHODS

Data Acquisition

The present study was endorsed by the Ethics Committee of the Shanghai Tongji Hospital affiliated to Tongji University. Microarray data and clinical information of 27 peripheral blood mononuclear cell (PBMC) samples from healthy volunteers and 25 PBMC samples from SCI patients were obtained from the Gene Expression Omnibus (GEO) database (accession

number: GSE82152)[1] (Barrett et al., 2011) and E-GEOD-69901 (ArrayExpress)[2] (Athar et al., 2019), respectively. In addition, these two batches of microarray data were both retrieved from a platform called GPL21975 [PrimeView] Affymetrix Human Gene Expression Array [Brainarray ENTREZG Version 17],[3] compared with data merging based on multiple different platforms, and resulted in less error. TFs were obtained from the Cistrome Cancer database[4] (Zheng et al., 2019). Immunologically related genomic expression profiles were downloaded from the ImmPort database[5] (Bhattacharya et al., 2014). Hallmark signaling pathways were collected from the Molecular Signatures Database (MSigDB, v7.2)[6] (Liberzon et al., 2015). ChIP-seq data of H3K27ac (accession number: GSE134744) and ATAC-seq data of key eRNAs (accession number: GSE139099) were obtained from the GEO database (see text foot note 1) (Barrett et al., 2011).

Data Preprocessing and Differential Expression Analysis

Samples and patients with incomplete clinical information were excluded. All original microarray data were read using affy package, followed by robust multi-array average (RMA) background correction, standardization, probe-specific background correction, and summarizing probe set values in one expression measure (Irizarry et al., 2003; Gautier et al., 2004). Then, two batches of microarray data aforementioned were corrected using normalizeBetweenArrays function in Linear Models for Microarray Analysis (limma) package, which were then merged for differential expression analysis. Differential expression analysis of eRNAs (DEeRNAs), target genes (DETGs) and TFs (DETFs) between SCI and normal blood was conducted using the Linear Models for Microarray Data (limma) package (Smyth, 2004). For p-value, false discovery rate (FDR) was utilized for multiple testing correction. Absolute $\log_2$[Fold Change (FC)] ≥ 1.0 and the FDR < 0.05 were cutoff criteria.

Functional Enrichment Analysis

We conducted Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses to identify biological processes and pathways that were most related to DEeRNAs with FDR p-value < 0.05 as cutoff value (Zhang T. et al., 2013).

Identification of Potential Immune Cells, Immune Pathways, and Hallmark Pathways

Immune cell proportions in the SCI and normal samples were analyzed utilizing the cell type identification by estimating relative subsets of RNA transcripts (CIBERSORT) algorithm (Newman et al., 2015). To identify the association between eRNA signature and immune cell infiltration in SCI tissues, we uploaded gene expression matrix data to CIBERSORT database[7] to purify cellular subtype-specific gene expression. Infiltrating immune cells were all extracted for further analysis. Furthermore, non-parametric tests were used to identify correlations between these immune cells/pathways and various clinical phenotypes.

In addition, ssGSEA algorithm was conducted to evaluate and quantify the enrichment level of 10 immune-related pathways in SCI samples (Xiao et al., 2020). Furthermore, Pearson correlation analysis was carried out to determine the correlations between key eRNAs and immune-related signaling pathways or immune genes, in which a p-value < 0.05 was considered statistically significant.

Identification of Potential Downstream Hallmark Pathways

GSVA (Ferreira et al., 2021) and GSEA (Subramanian et al., 2007) were both utilized for exploring potential downstream hallmark pathways of DEeRNAs. Absolute quantification of 50 hallmark signaling pathways was evaluated to extract differentially expressed hallmark pathways between SCI and normal blood using ClusterProfiler package and GSVA package (Yu et al., 2012; Hanzelmann et al., 2013). GSEA was carried out to reveal the significant enrichment of upregulated and downregulated hallmark pathways in SCI and normal blood. Furthermore, correlations of hallmark pathways of GSVA and GSEA were extracted, and the interactional pathways were suggested as key pathways.

Construction of Regulation Network for Transcription Factors, Enhancer RNAs, Immune Cells, Immune-Related Pathways, Target Genes, and Hallmark Gene Sets

Pearson correlation analysis was conducted based on key TFs, eRNAs, immune genes, immune cells, target genes, and hallmark pathways aforementioned. An eRNA regulatory network was then reconstructed by Cytoscape (3.7.1) (Kohl et al., 2011).

Molecular complex detection (MCODE) plugin of Cytoscape software was utilized to identify significant molecules within this network in a recognition standard MCODE score ≥ 4 to extract the modules of hub genes. Interaction relationships between key eRNAs and other components were controlled based on p-value < 0.05 and $|$ correlation coefficient$| > 0.40$.

Connectivity Map Analysis

Here, we used CMap (build 02) to find potential inhibitors that may target SCI-related eRNAs. In total, 6,100 gene expression cases covering 1,309 drugs were obtained from the CMap database[8] (Lamb et al., 2006). That is, a candidate drug might correspond to various gene expression cases. Genes in each case were ranked *via* taking differential expression values between drug-untreated and drug-treated cell lines, and 6,100 gene lists

[1] https://www.ncbi.nlm.nih.gov/geo/

[2] https://www.ebi.ac.uk/arrayexpress/

[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL21975

[4] http://cistrome.org/

[5] http://www.immport.org/

[6] http://software.broadinstitute.org/gsea/msigdb/

[7] https://cibersort.stanford.edu/

[8] https://portals.broadinstitute.org/cmap/

related to drugs were then generated. Based on identified key DEeRNAs involved in SCI and 6,100 drug-related cases, we conducted a non-parametric test to explore the relationship between drugs and SCI.

Information on targeting compounds is available in the mechanism of actions (MoA)[9] that includes transcriptional responses of various human cell lines to perturbagens, structural formulas, and protein targets. On the basis of MoA, compounds that may target SCI-related eRNAs/enhancers in this study were extracted.

Single-Cell RNA Sequencing Data Processing

Following the procedure of 10x Genomics Chromium[10] (Chen et al., 2020), the preprocessing of samples and scRNA-seq data was conducted. After demultiplexing, the sequencing results were divided into two pair-ended reads fastq files that were then trimmed to eliminate template switch oligo (TSO) sequence and polyA tail sequence. Additionally, clean reads were aligned with the GRCh38 (Version: 100) genome assembly, which were quantified using the Cell Ranger Software (Version 1.0.0).[11]

The quantitative gene expression matrices (The row names of matrices were genes and column names were barcodes) acquired from seven libraries that included seven vertical section samples and 14 cross-section samples were analyzed with Seurat pipeline (Version: 3.2.2) for further analysis (Butler et al., 2018). Only cells with less than 10% mitochondrial gene mapped and more than 100,000 transcripts expressing were extracted for subsequent analysis. In addition, genes that expressed in over three single cells were included in follow-up analyses. After the completion of quality control (QC), all samples were merged in one Seurat object with the function of "IntegrateData" that were then scaled and standardized with the function of "ScaleData." Top 1,500 variable genes were identified using "vst" method. To reduce model dimensionality, principal component (PC) analysis (PCA) was initially carried out, and the top 20 PCs were incorporated as input file for Uniform Manifold Approximation and Projection (UMAP) for dimension reduction analysis. The UMAP plots that illustrated cell subclusters were constructed using the "DimPlot" and "RunUMAP" function.

Differential Expression Analysis of Single-Cell RNA Sequencing

Genes with remarkable differential expression from the top 1,500 variable genes were defined as differentially expressed genes (DEGs) with "wilcox" method using "FindAllMarkers" function.

Cell Type Annotation

To identify the cell type of each unsupervised cluster, DEGs of all subclusters were utilized as potential references that were combined with known specific cell surface biomarkers obtained from CellMarker[12] (Zhang et al., 2019) for a comprehensive annotation of cell type. Given the variable gene expression patterns in SCI, a specific cellular annotation method was utilized in the present study. Firstly, known cell surface biomarkers of neuron (Rph3a, Tubb3, Gnal), neuron precursor cell (NPC) (Sox2, Prom1, Sox9), astrocyte (Gfap, S100b, Vim), fibroblast (Col1a1, Col1a2, Col4a1), oligodendrocyte (Mbp, Olig2, Gjb1), oligodendrocyte precursor cell (OPC) (Olig1, Apoc4, Epn2), microglia (Cx3cr1, Aif1, Itgam), and macrophage (Cd68, Itgam, Fcgr3) were used to annotate the cell type of the eight cells aforementioned. Furthermore, because several biomarkers of macrophage and microglia were overlapping, cells with transcripts per million (TPM) of Cd68 < 2 and Aif1 > 1.38 were identified as microglia, while cells with TPM of Cd68 > 2 were identified as macrophages. In addition, the cellular feature plots, dot plots, violin plots, and heat maps were constructed to show the marker genes of each cell type using SCANPY (Version: 1.7.1) and Seurat R package (Version: 3.2.2) under the environment of Python 3.6 (Butler et al., 2018; Wolf et al., 2018).

Cellular Communication Analysis

In order to elucidate the significant cellular communication patterns and ligand–receptor pairs among various different cell types in the spinal cord, cellular communication analysis was carried out using iTALK R package (Version: 0.1.0)[13] (Wang et al., 2019). Firstly, because of the scarcity of musculus resources in present mainstream cellular communication algorithms, the top 1,500 variable genes were transformed to human genes using biomaRt package (Version: 2.46.0) in a certain homologous degree with "getLDS" function (Durinck et al., 2009). Secondly, the normalized expression matrix of these genes was incorporated into the iTALK object with the "rawParse" function. Finally, the top 200 ligand–receptor pairs were shown by ligand–receptor plots and iTALK networks.

Chromatin Immunoprecipitation Sequencing Validation

Histone H3K27ac was implicated in enhancer-specific modifications, which were essential for enhancers to activate the transcription of relevant target genes (Kang et al., 2021). Role of H3K27ac in eRNA transcription was evaluated by analyzing ChIP-seq data (accession number: GSE134744) in peripheral blood (Grubert et al., 2020).

The ChIP-seq data of Splicing factor proline and glutamine rich (SFPQ) (accession number: GSM2827312, GSM1411215, GSM2825596, GSM1097497, GSM2827311, and GSM1097496) (Dunham et al., 2012; ENCODE Project Consortium, 2012; West et al., 2014) and H3K27 (accession number: GSM732912, GSM575294, and GSM663427) (Lister et al., 2009; Guenther et al., 2010; Wang et al., 2011) were obtained from the Cistrome database for validating binding relationships between key eRNAs and other significant biomarkers in this study. Furthermore, we determined the eRNA binding relationships using the University of California Santa Cruz (UCSC) Genome Browser[14] based on the original ChIP-seq data (Robinson et al., 2011; Li et al., 2015).

[9] http://clue.io/

[10] https://www.10xgenomics.com/instruments/chromium-x-series

[11] http://10xgenomics.com/

[12] http://biocc.hrbmu.edu.cn/CellMarker/

[13] https://github.com/Coolgenome/iTALK

[14] http://genome.ucsc.edu

Assay for Transposase-Accessible Chromatin Using Sequencing Validation

The ATAC-seq (Buenrostro et al., 2013) refers to an impressively flexible, simple, and powerful technique to profile chromatin regions genome-wide compared to traditional methods like functional assays or sequence conservation analyses.

We downloaded peak data of key TFs and eRNAs (accession number: GSE139099) (Grubert et al., 2020) and performed the correlation analysis using ggpubr.[15] In order to explore the potential super-enhancers, we performed the peak calling with macs14 (Feng et al., 2012).[16] Furthermore, we ranked enhancers and determined super-enhancers using ROSE (Whyte et al., 2013).[17] Multiple SCI-enriched ATAC-seq peaks that represented candidate regulation elements were located near established eRNAs and were enriched in distinct sets of TF binding sites.

Online Single-Cell RNA Sequencing Validation

To identify the expression of the key biomarkers in single-cell level, scRNA-seq data of developing mouse spinal cord (accession number: GO0006836 and GO0098609) were downloaded from Single Cell Expression Atlas (Delile et al., 2019).[18] Then, we validated the cellular location of these biomarkers' expression.

Statistics Analysis

All statistical analyses were put into effect using R version 3.6.1 (Institute for Statistics and Mathematics, Vienna, Austria).[19] In descriptive statistics, mean ± standard deviation was utilized for continuous variables in normal distribution.

In addition, when encountering continuous variables in abnormal distribution, the median was utilized. Two-sided p-value < 0.05 was suggested to be necessary for statistics.

RESULTS

Identification of Differentially Expressed Enhancer RNAs

Analysis process of the present study was illustrated in **Figure 1**. Samples and patients with incomplete clinical information were excluded, and conformers were shown in **Supplementary Table 1**. Based on the threshold, a total of 3,979 eRNAs were identified as DEeRNAs between 27 healthy volunteers and 25 SCI patients from 5,100 eRNAs, which were shown in the Heat map (**Figure 2A**) and volcano plot (**Figure 2B**).

Functional Enrichment Analysis

GO and KEGG enrichment analyses were both conducted to explore the potential mechanism of identified DEeRNAs. In GO analysis, the most important terms in biological process (BP), cellular component (CC), and molecular function (MF) were neutrophil activation, endosomal part, and cell adhesion molecule binding (**Figure 2C**). The most significant KEGG pathway was mitogen-activated protein kinase (MAPK) signaling pathway (**Figure 2D**). In the heat map (**Figure 3A**) and volcano plot (**Figure 3B**), 24 differentially expressed SCI-related eRNAs were identified, which were considered to be key eRNAs in patients with SCI.

Correlation Analysis of Key Differentially Expressed Enhancer RNAs and Differentially Expressed Transcription Factors

Twenty-one DETFs (log2 FC > 1 or < -1 and FDR < 0.05) were identified using limma package that were shown in the heat map (**Figure 3C**) and volcano plot (**Figure 3D**). In addition, to explore the relationship between DETFs and DEeRNAs, Pearson correlation analysis was carried out, and 27 regulatory relationships were identified (correlation coefficient < -0.300 or > 0.300, and p-value < 0.001). Based on the correlation analysis results for regulation relationships of key DETFs and DEeRNAs, link of SFPQ (TF) and vesicular overexpressed in cancer prosurvival protein 1 (VOPP1, eRNA) was extracted ($R = 0.990$, $p < 0.001$, positive).

Correlation Analysis of Key Differentially Expressed Enhancer RNAs and Differentially Expressed Target Genes

In total, 829 DETGs (log2 FC > 1 or < -1 and FDR < 0.05) were identified using limma package and illustrated in the heat map (**Figure 4A**) and volcano plot (**Figure 4B**). Furthermore, to determine the relationships between the identified DETGs and DEeRNAs, eight regulation relationships were identified using Pearson correlation analysis (correlation coefficient < -0.300 or > 0.300 and p-value < 0.001). Additionally, based on significant regulation relationships of key DEeRNAs and DETGs acquired by Pearson analysis, the link of VOPP1 (eRNA) and epidermal growth factor receptor (EGFR, target gene) was extracted ($R = 0.974$, $p < 0.001$, positive).

Correlation Analysis of Key Differentially Expressed Enhancer RNAs, Immune Cells/Pathways

Twenty-two types of immune cells/pathways were identified using CIBERSORT. The bar chart showed the percentage of 22 kinds of immune cells in 27 healthy volunteers and 25 SCI patients (**Figure 4C**). The heat map showed the expression level of 22 immune cells by CIBERSORT (**Figure 4D**). Results of non-parametric tests demonstrated significant correlation between immune cells/pathways and SCI (**Figure 4E**). Furthermore, ssGSEA was conducted to identify 28 immune cells/pathways that were significantly correlated with differentially expressed SCI-related eRNAs. Results of ssGSEA were shown in the heat map (**Figure 4F**).

[15]https://github.com/kassambara/ggpubr

[16]https://liulab-dfci.github.io/

[17]http://younglab.wi.mit.edu/super_enhancer_code.html

[18]https://www.ebi.ac.uk/gxa/sc/experiments/E-MTAB-7320/results/tsne

[19]www.r-project.org

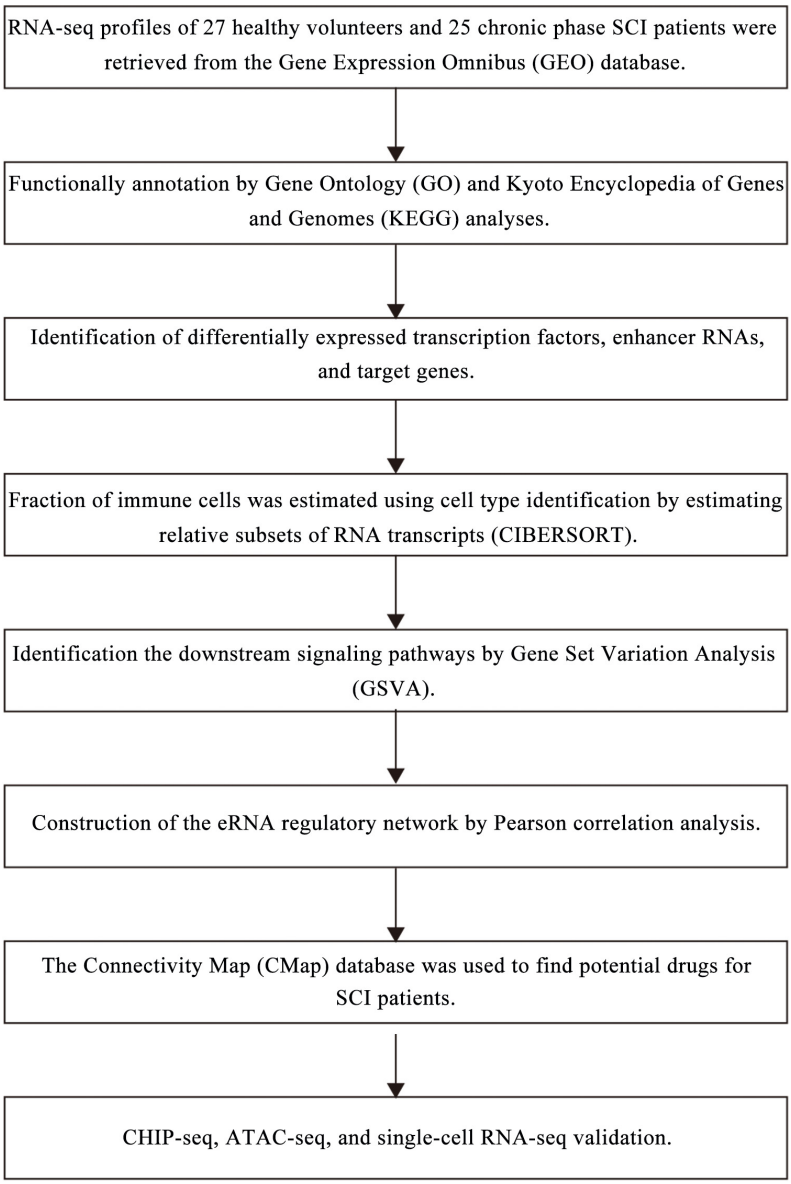


FIGURE 1 | The analysis flowchart of all analysis processes.

Then, by co-analyzing VOPP1 and 22 types of immune cells or pathways, the top 3 immune cells or immune-related pathways were extracted for further analysis, which included Antigen-presenting cells (APC) co-inhibition ($R = 0.936$, $p < 0.001$, positive), MHC class I ($R = -0.973$, $p < 0.001$, negative), and T helper (Th) cells ($R = -0.987$, $p < 0.001$, negative). Th cells were finally extracted in further analyses.

Identification of Potential Downstream Hallmark Pathways of Key Differentially Expressed Enhancer RNAs

Results of GSVA ($|\log2\ FC| > 0.1$, $p < 0.05$) were shown in the heat map (**Figure 5A**). The volcano plot showed 28 differentially expressed hallmark pathways between 27 healthy volunteers and 25 SCI patients (**Figure 5B**). The *t*-value of GSVA score of these pathways was shown in the *t*-test bar chart (**Figure 5C**). Furthermore, 38 hallmark pathways were extracted from the 50 hallmark pathways for further analysis, controlled by cutoff values of $|\log2\ FC| > 0.1$ and $p < 0.05$. Importantly, VOPP1 and hallmark coagulation ($R = 0.937$, $p < 0.001$, positive) showed the strongest interaction.

Construction of the Enhancer RNA Regulation Network

A co-expression regulation network of key DETFs, DEeRNAs, DETGs, immune cells/pathways, and hallmark pathways was

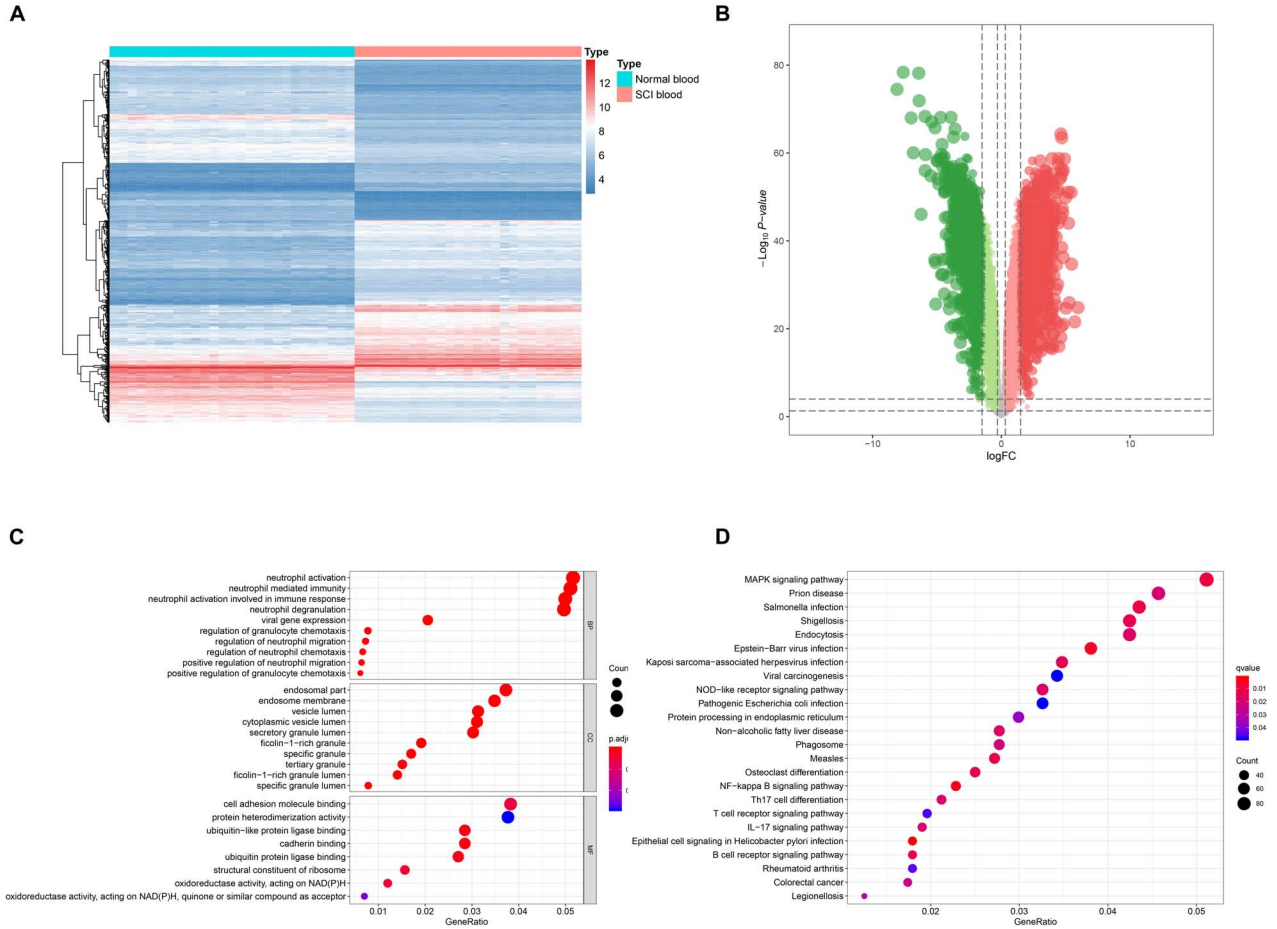


FIGURE 2 | Identification of differentially expressed enhancer RNAs (DEeRNAs) and functional annotation. **(A)** Heat map showing the expression level of DEeRNAs. **(B)** Volcano plot showing the *p*-value and log| FC| value of DEeRNAs by the differential expression analysis above. **(C)** The output of Gene Ontology (GO) analysis. **(D)** The output of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis.

constructed, which elaborated on the regulatory relationships among the aforementioned components (**Figure 5D**). Furthermore, to quantify the interaction coefficients among them, co-expression analysis was performed at the transcriptional level (**Figure 6A**).

Connectivity Map Analysis

Because eRNAs were implicated in the pathological processes of SCI and traditional long-term treatment with drugs may result in severe side effects and/or insufficient inflammation and pain relief, it is urgent to find potential compounds that target SCI-related eRNAs. Through exploring potential compounds in CMap database, we identified 19 compounds that have been validated to target SCI-related eRNAs in multiple clinical trials, including 15-delta prostaglandin J2, anisomycin, azaperone, bumetanide, dosulepin, felodipine, lycorine, metrizamide, MG-262, mitoxantrone, naftifine, pirinixic acid, Prestwick-864, probenecid, puromycin, rifampicin, trichostatin A (TSA), troglitazone, and valinomycin (**Figure 6B**). TSA (specificity = 0.471, $p < 0.001$) and 15-delta prostaglandin J2 (specificity = 0.128, $p < 0.001$) with the highest specificity were considered the best compounds to target SCI-related eRNAs (**Figures 6C,D**), and TSA was extracted for further analysis.

Integrated Analysis of Single-Cell RNA Sequencing

The UMAP scatter plots and cellular feature plots showed the cell types and marker genes of different cell types (**Figure 7A**). Furthermore, the top 1,500 variable genes (| log2(FC)| > 0.5 and FDR < 0.05) were extracted by DEG analysis. Gene expression levels of the most significant DEGs of multiple inflammation cells (astrocyte, dendritic, Div-myeloid, endothelial, ependymal, fibroblast, lymphocyte, macrophage, microglia, monocyte, neutrophil, oligodendrocyte, OPC, and pericyte) and neuron were illustrated in **Figure 7B**. In addition, the proportion of these cells in different SCI samples was also shown in the bar plot (**Figure 7B**). Specifically, average number and cell proportion of these cells were illustrated in **Figure 7C**. The feature plots of VOPP1, EGFR, and SFPQ were shown in **Figure 7D**. Importantly, VOPP1 was significantly highly

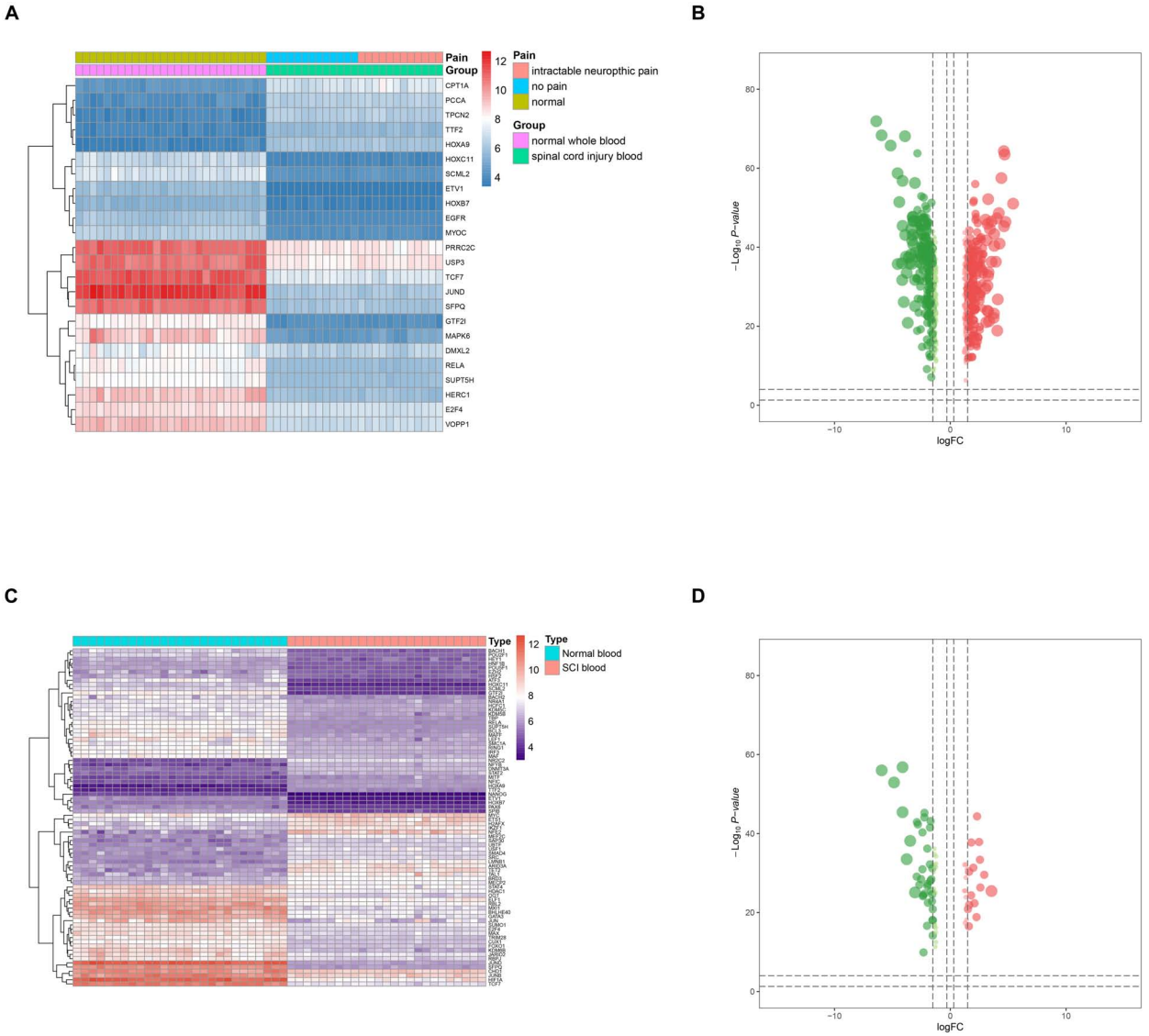


FIGURE 3 | Identification of differentially expressed spinal cord injury (SCI)-related enhancer RNAs (eRNAs) and transcription factors (TFs). **(A)** Heat map showing the expression level of differentially expressed SCI-related eRNAs. **(B)** Volcano plot showing the *p*-value and log| FC| value of differentially expressed SCI-related eRNAs. **(C)** Heat map showing the expression level of differentially expressed TFs. **(D)** Volcano plot showing the *p*-value and log| FC| value of differentially expressed TFs.

expressed in lymphocytes, which was consistent with the previous analysis. In addition, EGFR and SFPQ were highly expressed in fibroblast and microglia, respectively. Finally, intersected cellular communication network and ligand–receptor plot showed the mechanisms of intercellular signal transduction, and fibroblast was the core cellular component among these intersected ligand–receptor pairs (**Figures 7E,F**). It showed that lymphocytes received signals from myeloid cells and various neurocytes including microglia and then played a role in the downstream pathological process of SCI.

The UMAP scatter plots and cellular feature plots showed various cell types and marker genes for corresponding cell types, respectively (**Figure 8A**). Cellular feature plots of marker genes in different cell types were shown in **Figure 8B**. The most significant DEGs of neuron and inflammation cells were annotated on the DEG heat map (**Figure 8C**).

Chromatin Immunoprecipitation Sequencing and Assay for Transposase-Accessible Chromatin Using Sequencing Validation

In order to explore the role of enhancer-specific histone in modifications of eRNA transcription, ChIP-seq data of H3727ac were obtained and analyzed. The UCSC Genome Browser tracks illustrated enrichment of H3K27ac on multiple loci in various key eRNAs identified in this study (CPT1A, MAPK6, PCCA, PRRC2C, USP3, and VOPP1) (**Figure 9**).

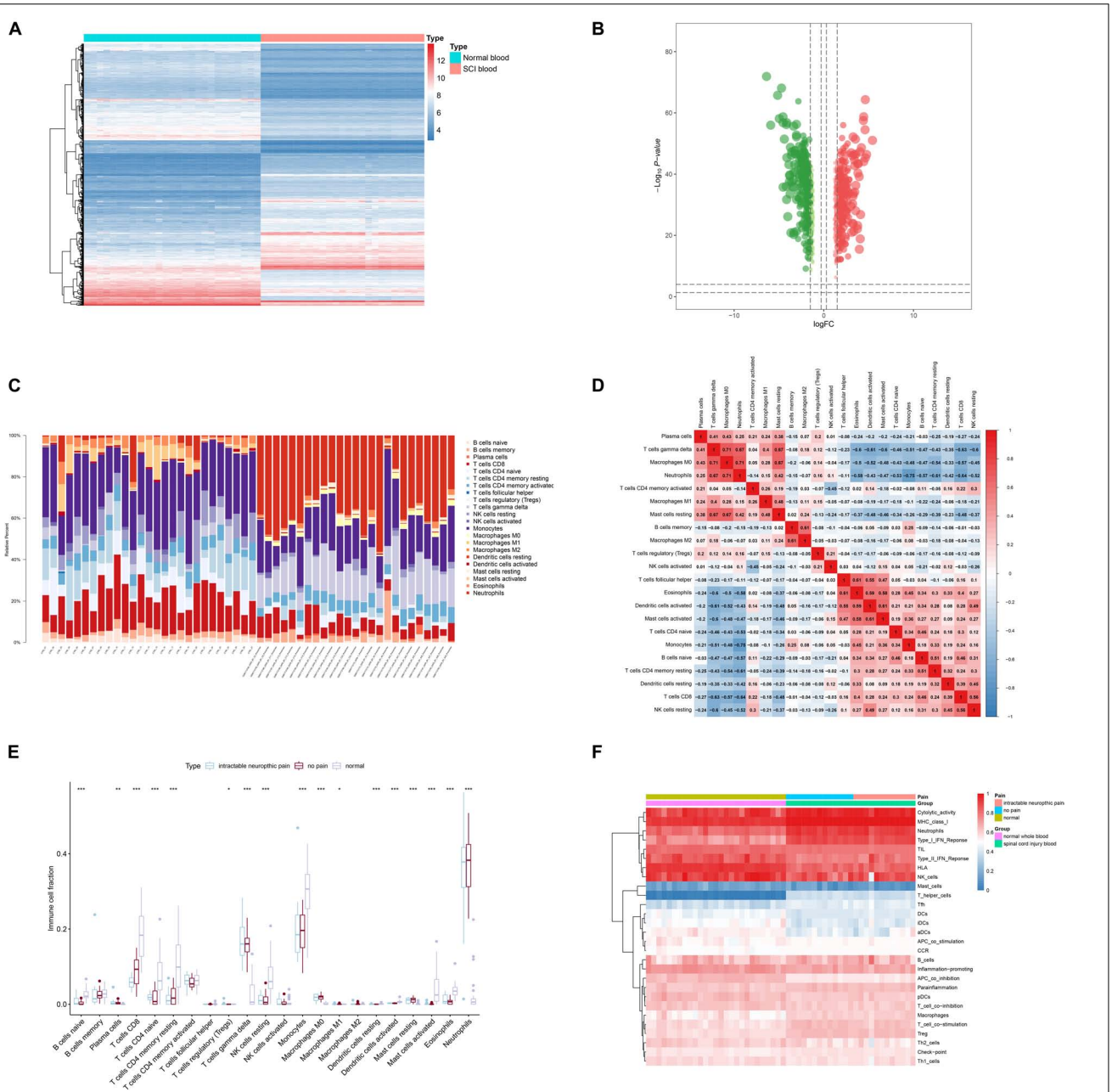


FIGURE 4 | Identification of differentially expressed target genes (DETGs) and evaluation of immune cells/gene sets. **(A)** Heat map showing the expression level of DETGs. **(B)** Volcano plot showing the *p*-value and log| FC| value of DETGs. **(C)** Bar chart showing the percentage of 22 kinds of immune cells from healthy volunteers and spinal cord injury (SCI) samples by cell type identification by estimating relative subsets of RNA transcripts (CIBERSORT). **(D)** Heat map showing the expression level of 22 immune cells by CIBERSORT. **(E)** Box plot showing the fraction of 22 immune cells from healthy volunteers and SCI samples by non-parametric tests. **(F)** Heat map showing the expression level of 28 immune cells/pathways *via* single-sample Gene Set Enrichment Analysis (ssGSEA).

Regulatory relationships between VOPP1 and other key markers (SFPQ and H3K27) were further explored using ChIP-seq and ATAC-seq methods. We retrieved relative data from the Cistrome database to detect DNA fragments binding with SFPQ and H3K27ac in various SCI-related and inflammation-related cases. The results showed that in ChIP-seq data of SFPQ (**Supplementary Figure 1**) and H3K27 (**Supplementary Figure 2**), there were strong binding peaks in the position of VOPP1, which basically coincided with the open chromosome region shown in the ATAC-seq results (**Supplementary Figure 3**). Furthermore, we downloaded the original ChIP-seq data of SFPQ and H3K2 to analyze the results with UCSC Genome Browser. Binding peaks in SFPQ and H3K27 groups were obviously stronger than those in control groups

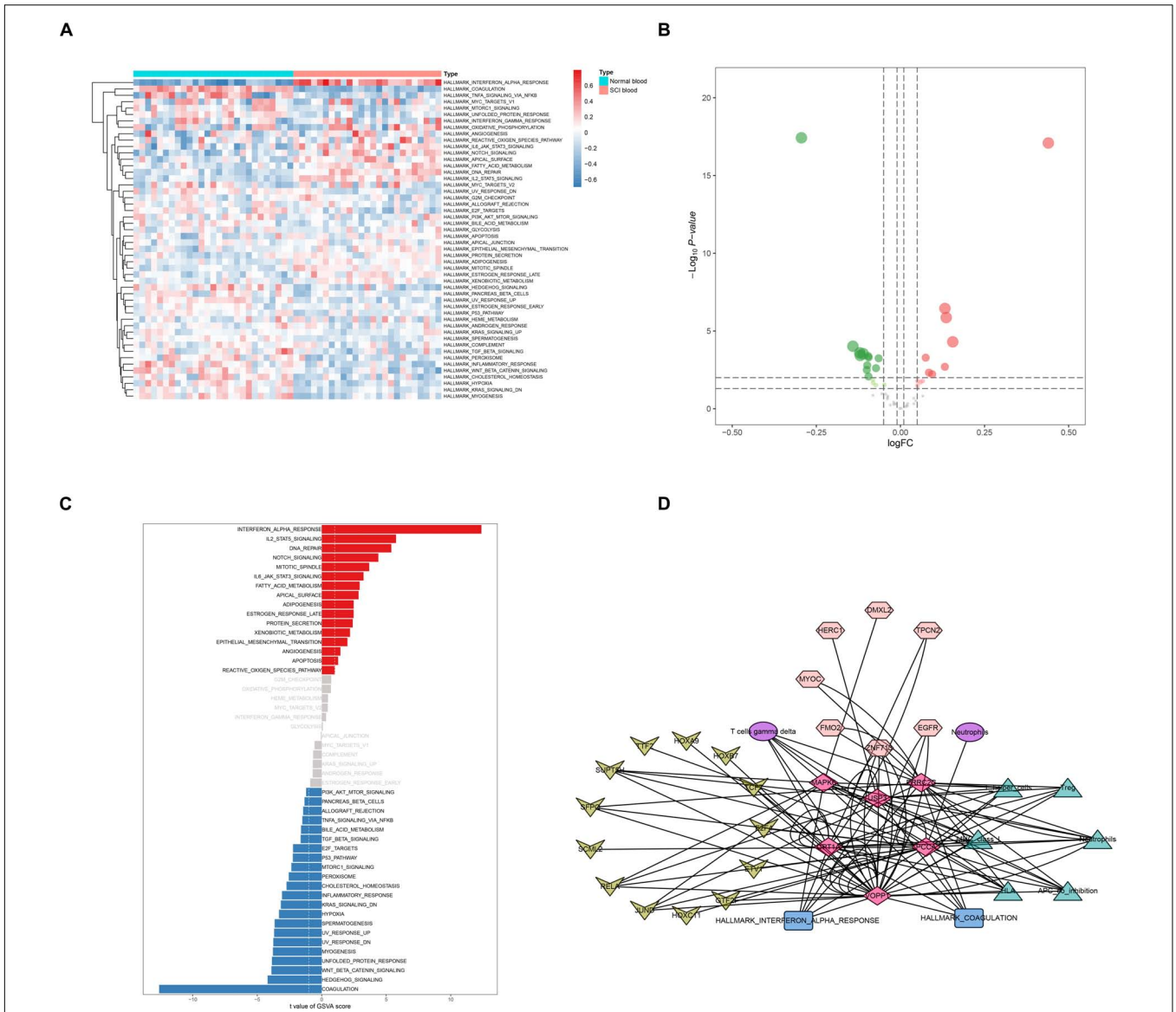


FIGURE 5 | Identification of key hallmark pathways. **(A)** Heat map showing the co-expression level of hallmark sets in spinal cord injury (SCI) samples by Gene Set Variation Analysis (GSVA). **(B)** Volcano plot showing the co-expression level of hallmark sets in SCI samples by GSVA. **(C)** The bar plot revealing the *t*-value of GSVA score. **(D)** Overview of the protein–protein interaction network of OA-related enhancer RNAs (eRNAs), translational factors (TFs), immune cells/pathways, and hallmark gene sets. The diamonds represented key eRNAs. The arrows represented TFs. The ellipses represented immune cells. The triangles represented immune pathways. The hexagons represented target genes. The rectangle represented hallmark gene sets.

in the chromosomal position of VOPP1, further validating the binding relationship.

Online Single-Cell RNA Sequencing Validation

To explore the cell subtype localization of the key eRNAs in the speculative regulation mechanisms, transcriptome combined with scRNA-seq data of SCI samples was analyzed. Sixteen clusters were identified using t-distributed stochastic neighbor embedding (t-SNE) (**Supplementary Figure 4A**). The heat map plot showed the top 5 marker genes of each cluster in SCI samples (**Supplementary Figure 4B**). The feature plots illustrated the distribution and expression level of SFPQ (**Supplementary Figure 4C**), VOPP1 (**Supplementary Figure 4D**), and EGFR (**Supplementary Figure 4E**) in SCI tissues.

DISCUSSION

Multiple pathogenic factors contribute to the damage of the spinal cord and lead to SCI, which is a common orthopedic disease (Phillips et al., 2015). SCI is a pathological process with multi-gene mutation and multi-pathway dysfunction that causes

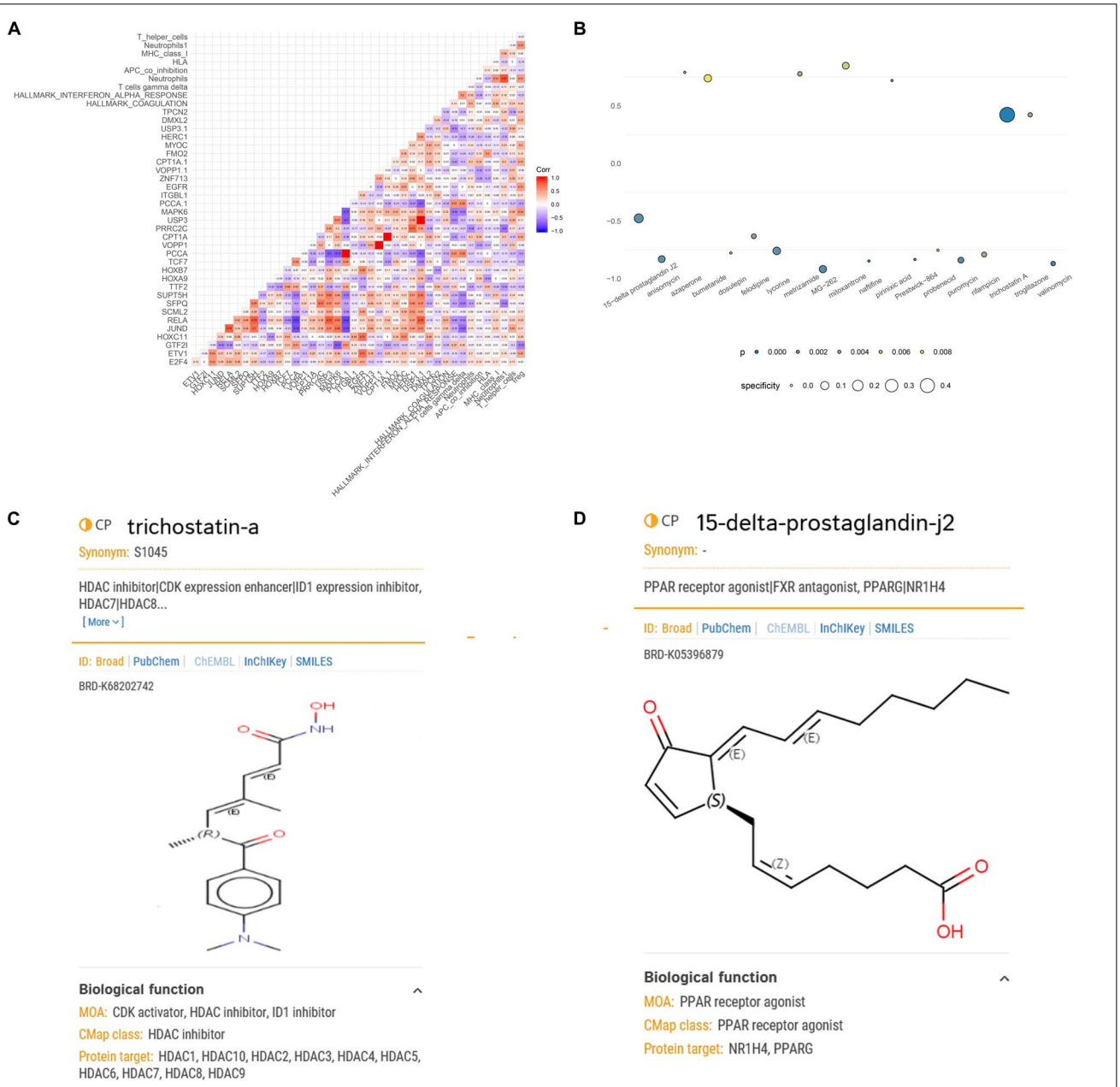


FIGURE 6 | Correlation analysis of the key biomarkers and identification of potential small-molecule drugs for spinal cord injury (SCI). **(A)** Heat map showing the results of correlation analysis (Pearson analysis) of enhancer RNAs (eRNAs), transcription factors (TFs), target genes, immune cells/pathways, and hallmark pathways. **(B)** Heat map showing perturbagens from the Connectivity Map (CMap) that might be capable of targeting SCI-related eRNAs. **(C)** Structural formula of trichostatin A. **(D)** Structural formula of 15-delta prostaglandin J2.

serious neurological dysfunction (Forner et al., 2016). Primary injuries of SCI during the early stages are usually accompanied by secondary ischemia, tissue edema, and ischemia–reperfusion injury. Because of the non-renewable features of nerve cells and extremely high deformity rates, SCI has brought economic and psychological burdens (Smith et al., 2015). Research reported that primary damages induced by early stage of SCI caused massive neuronal apoptosis and keratinocyte regeneration. Moreover, formation of glial scars also inhibited nerve fiber growth, as well as exogenous and endogenous factors (Selvarajah et al., 2015). Currently, studies of SCI are mainly concentrated in mice, whereas research on gene expression changes in peripheral whole blood is relatively inadequate (Jin et al., 2015; Ropper et al., 2015).

In this study, on the basis of comprehensive bioinformatics analysis, DETFs, SCI-related eRNAs, and target genes were all identified. Furthermore, according to 27 SCI-related DEeRNAs and significant DETF links, link of SFPQ (TF) and VOPP1 (eRNA) was extracted ($R = 0.990$, $p < 0.001$, positive). Based on

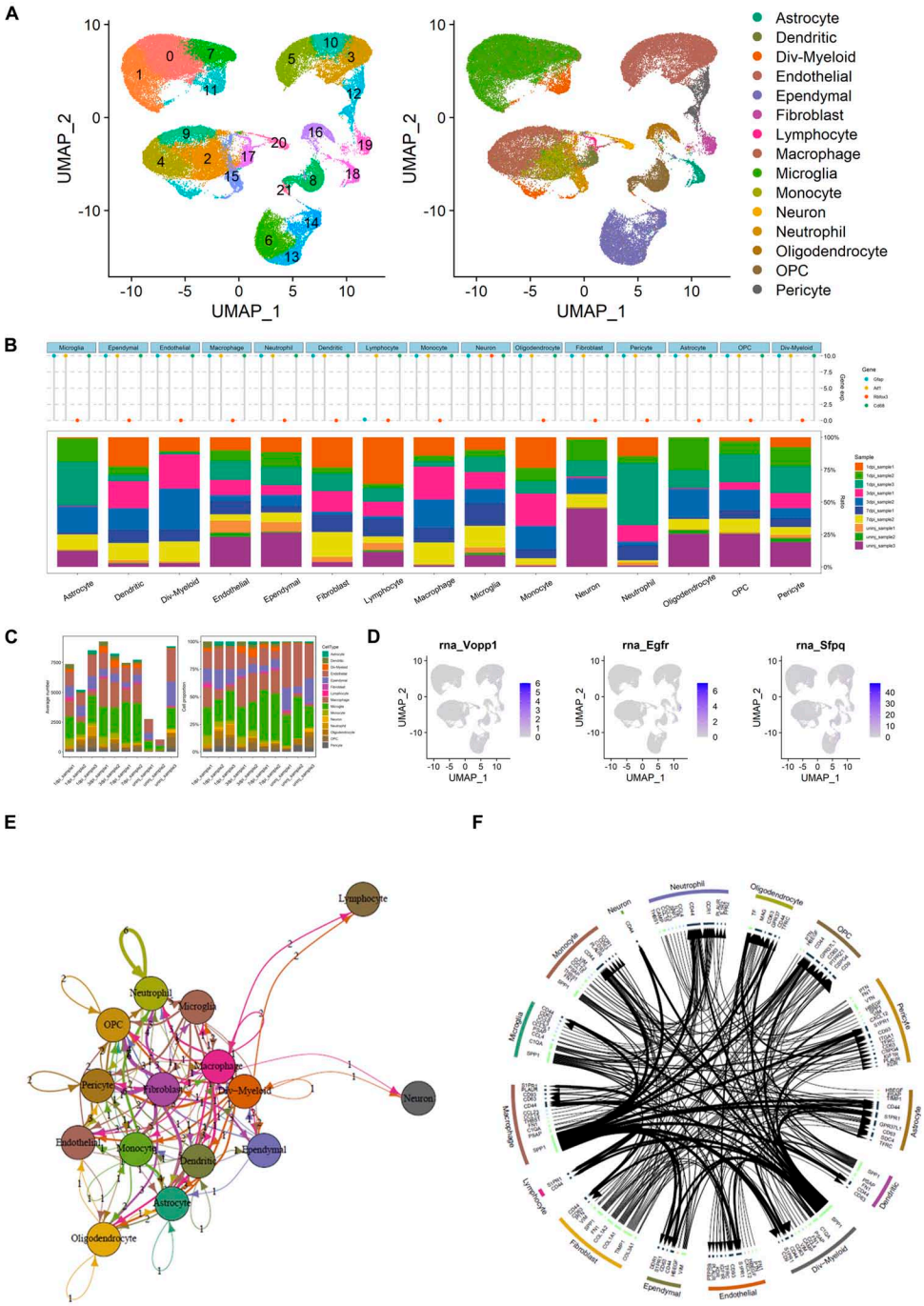


FIGURE 7 | Integrated analysis of single-cell RNA sequencing (scRNA-seq) and cellular communication. **(A)** Uniform Manifold Approximation and Projection (UMAP) scatter plots and cellular feature plots showing the cell types and marker genes of each cell type. **(B)** Dot plot showing gene expression levels of the most significant differentially expressed genes (DEGs) of multiple inflammation cells and neuron. In addition, the proportion of these cells in different spinal cord injury (SCI) samples was also shown in the bar plot. **(C)** Bar plots showing the average number and cell proportion of these cells. **(D)** Feature plots of vesicular overexpressed in cancer prosurvival protein 1 (VOPP1), epidermal growth factor receptor (EGFR), and Splicing factor proline and glutamine rich (SFPQ). **(E)** Network showing the intersected cellular communication genes. **(F)** Circle plot showing the intersected cellular communication genes.

eight regulation relationships of key DEeRNAs and DETGs, the link of VOPP1 (eRNA) and EGFR (target gene) was extracted ($R = 0.974$, $p < 0.001$, positive). And by co-analyzing VOPP1 and 22 types of immune cells or immune-related pathways, Th cells ($R = -0.987$, $p < 0.001$, negative) were postulated to be the most significant immune cells. On the basis of comprehensive

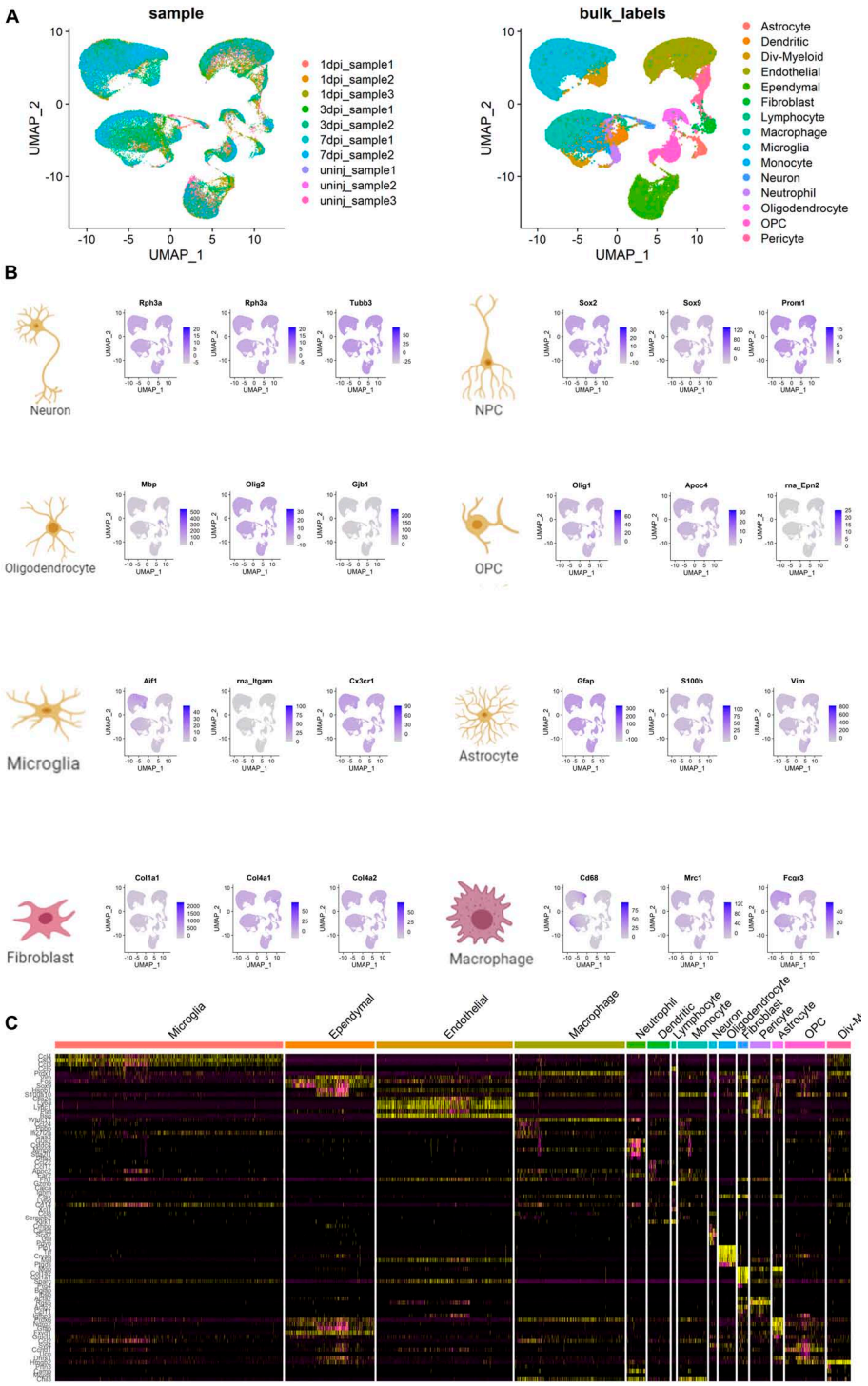


FIGURE 8 | Cell type annotation and cellular communication analysis. **(A)** Uniform Manifold Approximation and Projection (UMAP) scatter plots and cellular feature plots showing the cell types and marker genes for each cell type. **(B)** Cellular feature plots showing marker genes of different cell types. **(C)** Heat map showing the most significant differentially expressed genes (DEGs) of neuron and inflammation cells.

consideration of GSVA and correlation results of VOPP1 and 50 hallmark gene sets, VOPP1 and hallmark coagulation ($R = 0.937$, $p < 0.001$, positive) showed the strongest interaction. Eventually, positively regulated by SFPQ, VOPP1 strengthened the transient expression of EGFR. Th cells and hallmark coagulation were the downstream immune cells/pathways of VOPP1 in SCI.

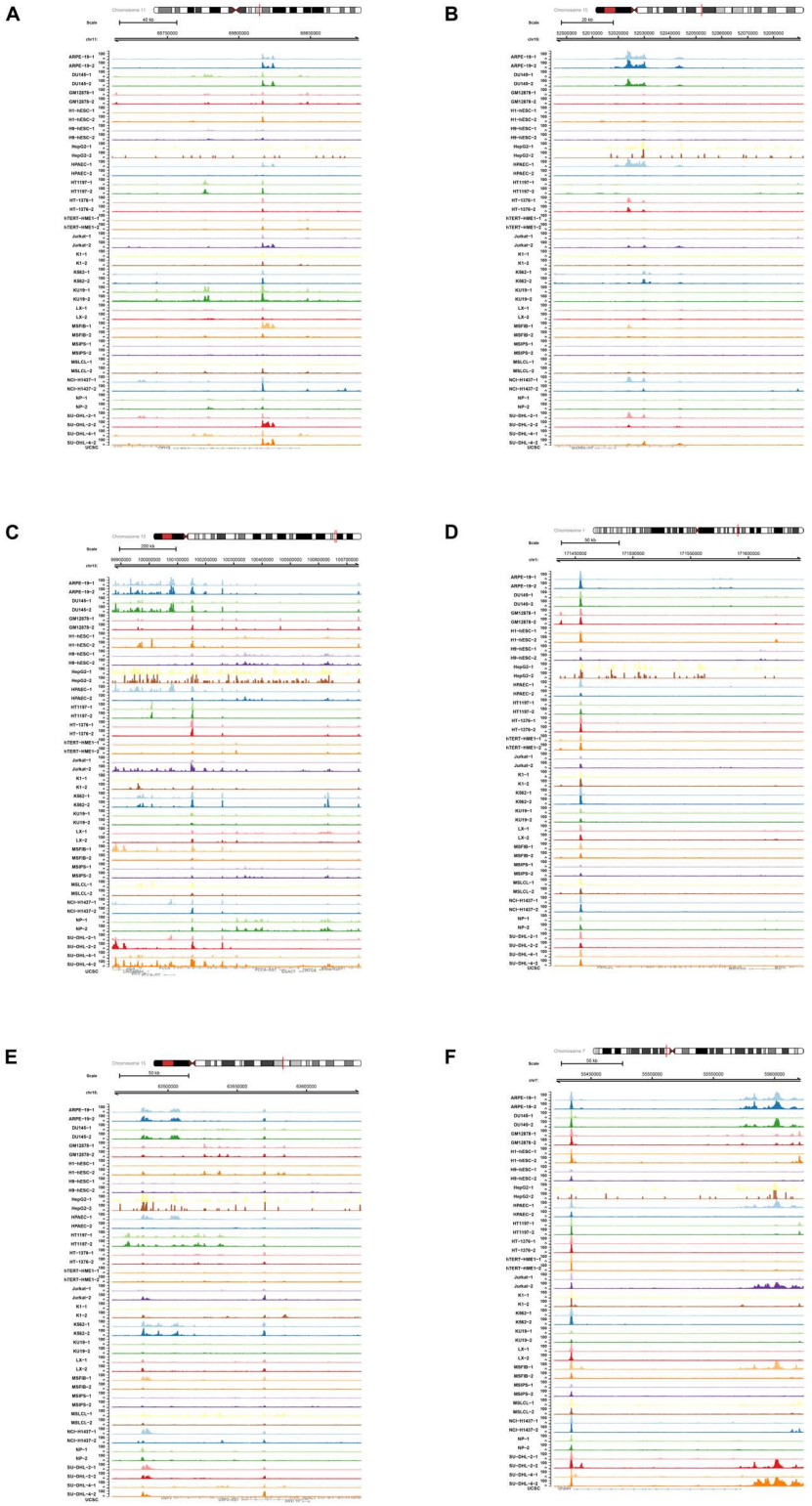


FIGURE 9 | Chromatin immunoprecipitation sequence (ChIP-seq) analysis of key differentially expressed enhancer RNAs (DEeRNAs). **(A)** In ChIP-seq data of H3K27ac, multiple binding peaks were found in CPT1A sequences. **(B)** In ChIP-seq data of H3K27ac, multiple binding peaks were found in MAPK6 sequences. **(C)** In ChIP-seq data of H3K27ac, multiple binding peaks were found in PCCA sequences. **(D)** In ChIP-seq data of H3K27ac, multiple binding peaks were found in PRRC2C sequences. **(E)** In ChIP-seq data of H3K27ac, multiple binding peaks were found in USP3 sequences. **(F)** In ChIP-seq data of H3K27ac, multiple binding peaks were found in vesicular overexpressed in cancer prosurvival protein 1 (VOPP1) sequences.

Neurocytes with enhanced expression of SFPQ and EGFR signal the lymphocytes overexpressing VOPP1 to infiltrate peripheral blood, which may play important roles in apoptosis and promote the development of inflammation in patients with SCI based on scRNA-seq and cellular communication analysis. In addition, based on ChIP-seq and ATAC-seq analysis, multiple binding peaks in SFPQ and H3K27 groups were identified in the chromosomal position of VOPP1, further validating the binding relationship.

VOPP1, also known as glioblastoma-amplified secreted protein (Dunham et al., 2012) and EGFR-coamplified and overexpressed protein (ECOP) (Fang et al., 2020), is upregulated in multiple human tumors, such as gastric cancer and squamous cell carcinoma (Baras et al., 2011; Gao et al., 2015). Overwhelming evidence demonstrates that VOPP1 acts as a critical regulator of nuclear factor kappa B (NF-κB) signaling pathway, and it could be implicated in apoptosis resistance (Li et al., 2017). The nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor alpha (IκBα) is correlated with NF-κB that prevents NF-κB from binding cognate promoters and also leads to steady-state cytoplasm localization. When stimulated by activation signals including tumor necrosis factor alpha (TNFα), IκBα is quickly degraded by phosphorylation-dependent ubiquitination (Karin and Ben-Neriah, 2000). Degradation of IκBα causes NF-κB to translocate to the nucleus and leads to the activation of multiple genes that are significant in various processes, including apoptosis, immune-related and inflammation response, and cellular proliferation (Park and James, 2005). Inhibition of VOPP1 expression in PBMCs alleviated the inflammatory response in advanced sepsis patients (Li et al., 2018). Taken together, increased VOPP1 expression could confer resistance to apoptosis and promote the development of inflammation. Further study of VOPP1 in SCI pathological process may bring insights into refining therapeutic regimens, as well as other disease states that are associated with elevated VOPP1 activity, such as asthma, arthritis, chronic/acute inflammation, and diverse tumors (Karin et al., 2002).

SFPQ refers to a multifunctional protein, which contributes to substantial biological processes, including DNA synthesis, gene expression regulation, DNA repair, RNA splicing, and apoptosis (Shav-Tal and Zipori, 2002). Importantly, SFPQ plays a critical role in neuronal differentiation and survival (Lowery et al., 2007). Furthermore, central nervous system (CNS) injury stimulates SFPQ expression, hindering neurogenesis and nerve regeneration, which may inhibit the repair of injured spinal cord (Elsaeidi et al., 2014). SFPQ is rich in paraspeckles, RNA-protein structures identified in the interchromatin space of nucleus, which increase in response to pro-inflammatory stimuli (Fox and Lamond, 2010). Additionally, SFPQ were significantly upregulated in chronic inflammatory diseases, including Crohn's disease and ulcerative colitis (Hasler et al., 2011). Hence, this provides novel options for researchers in the field of SCI because SFPQ not only shows SCI-specific effects of inflammation that serves as a diagnostic biomarker but also can be a potential target for therapeutic intervention of SCI.

Based on integrated multinomial bioinformatics analysis and other studies, EGFR was suggested to be a target gene of VOPP1 (Wang et al., 1998; Eley et al., 2002). EGFR belongs to the receptor tyrosine kinase (RTK) superfamily, consisting of three other members, ErbB2/Neu/HER-2, ErbB3/HER-3, and ErbB4/HER-4 (Romano and Bucci, 2020). Being a neurotrophin receptor to initiate cellular signaling and regulate neuronal processes, EGFR has many important roles in the CNS (Hennigan et al., 2007). Moreover, EGFR is important in neural stem cells self-renewal, and loss of EGFR signaling induces these cells to differentiate preferentially into glia, which may promote glia scar formation in SCI and influence functional recovery (Robson et al., 2018). Furthermore, EGFR is crucial in regulating the differentiation of precursors to astrocytes, and increased EGFR expression levels determine their differentiation to astrocytes (Burrows et al., 1997). EGFR also plays an important role in astrocytes' morphology. Blockade of its expression causes disorganization of astrocytes in CNS development, losing their processes surrounding neurons, which induces degeneration of abundant axons of nerve (Liu and Neufeld, 2007). Modulation of EGFR expression may be propitious to activate regeneration and counteract neurodegeneration (Ceresoli, 2012). Therefore, EGFR is critical for differentiation, growth, and repair of the injured tissue in the spinal cord, which could be a potential target for therapy of patients with SCI.

Immune response in peripheral blood plays an important role in maintaining spinal cord homeostasis. Dysfunction or disorder caused by SCI in vegetative innervation of lymphatic and endocrine systems leads to long-term abnormal inflammation responses (Zhang Y. et al., 2013). Differentiating to functionally distinct Th subsets, Th cells are crucial in normal immune surveillance and proper immune regulation (Murphy and Reiner, 2002). Importantly, SCI is associated with immune depression syndrome, owing to the dysregulated hypothalamic–pituitary–adrenal (HPA) axis and dysfunctioned sympathetic nervous system (Jones, 2014). Based on bioinformatics analysis and other studies, a rapid decrease in Th cells was identified in patients with SCI, contributing to immunosuppression in the acute phase of SCI (Gao et al., 2021). We postulated that overexpression of VOPP1 decreased the counts of Th cells in peripheral blood of SCI patients. It is an important reason that acute SCI patient becoming chronic, revealing novel targets for future SCI immunotherapy.

SCI is correlated with a remarkable thrombophilia and aberrant coagulation. In addition, the relation between deep venous thrombosis (DVT) and SCI has been well established (Yao, 1992). Numerous experimental research with fibrinogen scanning has shown the presence of DVT in nearly 100% of patients with acute SCI (Maynard et al., 1997). Sequelae from SCI affect the three components of Virchow triad, which are responsible for DVT, blood flow deceleration, vascular endothelial damage, release of procoagulants, and coagulation cascade activation, resulting in hypercoagulability state. Furthermore, patients with SCI often show dehydration and secondary increase of blood viscosity, aggravating the stasis and hypoxia in injured spinal cord (Ersoz et al., 1999). On the basis of bioinformatics analysis and other studies, coagulation was postulated to contribute to the pathogenesis and progress of SCI

via the positive regulation of VOPP1, which is a promising entry point of therapeutic interventions for SCI (Ghlissi et al., 2019).

TSA refers to a highly specific hydroxamic acid that can inhibit histone deacetylase (HDAC) enzymes. Importantly, HDACs repress repressor element 1 silencing transcription factor (REST) that is able to counteract neuronal differentiation traits and Sp1, a TF that can mediate neuronal antioxidant signaling pathways. TSA leads to histone hyperacetylation, accompanied by activation of antioxidant gene expression and neuronal maturation, which indicate TF derepression of REST and Sp1 (Hakimi et al., 2002; Ryu et al., 2003). Moreover, TSA can block immune cell proliferation and suppress pro-Th1 factor IFN-γ, which could cause the transformation of Th1 to Th2 phenotype, indicating neuroimmunoprotective effects (Dangond et al., 2004). Importantly, TSA can also inhibit CNS immune cell infiltration and lipid breakdown products absorbing of microglial cells, which are consistent with anti-microglial activation of TSA recently reported (Konishi et al., 2002). Furthermore, a recent treatment showed TSA was able to decrease nitrosylation of spinal cord tissues, protecting nerve cells from free radical attack after SCI (Dasgupta et al., 2003). TSA increases histone acetylation and E2F-dependent transcripts in injured spinal cords, indicating its significant neuroprotective effect. Taken together, the general effect of TSA is to adjust dysregulated homeostatic processes, promoting histone acetylation and harnessing much more favorable genes than pathogenicity gene such as chemokine and pro-Th1, which may be effective for SCI treatment.

There were still several limitations to this study. First, the data were obtained from public sources statistically imperfect with limited samples. Second, information on other confounding variables in this study, such as smoking, was not available. Third, a prospective study is needed to evaluate the significance of these key biomarkers in terms of long-term clinical outcomes and possible applications of molecular drugs for SCI therapy. Finally, further experiment is an absolute necessity in demonstrating the regulatory mechanisms of key eRNAs implicated in SCI. Therefore, ChIP-seq data of H3K27ac from online databases were obtained and analyzed, which broadened the scope of validation and supplemented the specific regulatory mechanisms of eRNA action involved in the pathogenesis of SCI. ATAC-seq data of VOPP1 (key eRNA) were also utilized to validate the eRNA regulatory mechanisms. Moreover, the cell subtype localizations of the key eRNAs and TFs were identified by scRNA-seq validation fluorescence immunohistochemistry. Additionally, a comprehensive transcriptome bioinformatics analysis of spatial transcriptome and scRNA-seq, fluorescence immunohistochemistry, and eRNA-related direct mechanism experiments would be the further research directions.

CONCLUSION

Based on integrated multinomial bioinformatics analysis, we found that SFPQ was the most significant TF and VOPP1 was the most significant key eRNA in the progression of SCI patients. In addition, during this pathological process, VOPP1 upregulated the transient expression of EGFR. Th cells and hallmark coagulation were the potential downstream pathways of VOPP1. Moreover, this study provided candidate small-molecule compounds as potential targets for the treatment of SCI patients.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), ArrayExpress (https://www.ebi.ac.uk/arrayexpress/), and Cistrome Cancer database (http://cistrome.org/).

AUTHOR CONTRIBUTIONS

RH, SW, RZ, SX, ZH, LC, and JZ: conception and design, collection and assembly of data, data analysis and interpretation, the manuscript writing, and final approval of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported in part by the National Natural Science Foundation of China (Nos. 81501203 and 81820108013), the Shanghai Municipal Health Commission (No. 201940306), the Henan Medical Science and Technology research project (No. 201602031), and the Key Project of Provincial and Ministerial Co-construction of Henan Medical Science and Technology (No. SBGJ202002031). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.728242/full#supplementary-material

Supplementary Figure 1 | In SFPQ Chromatin immunoprecipitation sequence (ChIP-seq) data, multiple binding peaks were found in VOPP1 sequences.

Supplementary Figure 2 | In H3K27 Chromatin immunoprecipitation sequence (ChIP-seq) data, multiple binding peaks were identified in VOPP1 sequences.

Supplementary Figure 3 | In SFPQ Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) data, multiple binding peaks were identified in VOPP1 sequences.

Supplementary Figure 4 | Single-cell RNA-seq validation. **(A)** A total of 25 numbered clusters were identified by t-NSE. **(B)** Feature plots illustrating the distribution and expression of SFPQ. **(C)** Feature plots illustrating the distribution and expression of VOPP1. **(D)** Feature plots illustrating the distribution and expression of EGFR. **(E)** Heat map showing the expression level of the top 5 marker genes of each cluster.

Supplementary Table 1 | Baseline information of 27 normal peripheral blood mononuclear cell (PBMC) samples and 25 spinal cord injury PBMC samples.

REFERENCES

Anderson, M. A., O'shea, T. M., Burda, J. E., Ao, Y., Barlatey, S. L., Bernstein, A. M., et al. (2018). Required growth facilitators propel axon regeneration across complete spinal cord injury. *Nature* 561, 396–400. doi: 10.1038/s41586-018-0467-6

Athar, A., Fullgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., et al. (2019). ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* 47, D711–D715. doi: 10.1093/nar/gky964

Badhiwala, J. H., Wilson, J. R., and Fehlings, M. G. (2019). Global burden of traumatic brain and spinal cord injury. *Lancet Neurol.* 18, 24–25. doi: 10.1016/S1474-4422(18)30444-7

Baras, A. S., Solomon, A., Davidson, R., and Moskaluk, C. A. (2011). Loss of VOPP1 overexpression in squamous carcinoma cells induces apoptosis through oxidative cellular injury. *Lab. Investig.* 91, 1170–1180. doi: 10.1038/labinvest.2011.70

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., et al. (2011). NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.* 39, D1005–D1010. doi: 10.1093/nar/gkq1184

Beck, K. D., Nguyen, H. X., Galvan, M. D., Salazar, D. L., Woodruff, T. M., and Anderson, A. J. (2010). Quantitative analysis of cellular inflammation after traumatic spinal cord injury: evidence for a multiphasic inflammatory response in the acute to chronic environment. *Brain* 133, 433–447. doi: 10.1093/brain/awp322

Bhattacharya, S., Andorf, S., Gomes, L., Dunn, P., Schaefer, H., Pontius, J., et al. (2014). ImmPort: disseminating data to the public for the future of immunology. *Immunol. Res.* 58, 234–239. doi: 10.1007/s12026-014-8516-1

Blackwood, E. M., and Kadonaga, J. T. (1998). Going the distance: a current view of enhancer action. *Science* 281, 60–63. doi: 10.1126/science.281.5373.60

Borton, D., Bonizzato, M., Beauparlant, J., Digiovanna, J., Moraud, E. M., Wenger, N., et al. (2014). Corticospinal neuroprostheses to restore locomotion after spinal cord injury. *Neurosci. Res.* 78, 21–29. doi: 10.1016/j.neures.2013.10.001

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218. doi: 10.1038/nmeth.2688

Burrows, R. C., Wancio, D., Levitt, P., and Lillien, L. (1997). Response diversity and the timing of progenitor cell maturation are regulated by developmental changes in EGFR expression in the cortex. *Neuron* 19, 251–267. doi: 10.1016/S0896-6273(00)80937-X

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Ceresoli, G. L. (2012). Role of EGFR inhibitors in the treatment of central nervous system metastases from non-small cell lung cancer. *Curr. Cancer Drug. Targets* 12, 237–246. doi: 10.2174/156800912799277430

Chen, K. B., Uchida, K., Nakajima, H., Yayama, T., Hirai, T., Watanabe, S., et al. (2011). Tumor necrosis factor-alpha antagonist reduces apoptosis of neurons and oligodendroglia in rat spinal cord injury. *Spine* 36, 1350–1358. doi: 10.1097/BRS.0b013e3181f014ec

Chen, W. T., Lu, A., Craessaerts, K., Pavie, B., Sala Frigerio, C., Corthout, N., et al. (2020). Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. *Cell* 182, 976–991.e919. doi: 10.1016/j.cell.2020.06.038

Dangond, F., Hwang, D., Camelo, S., Pasinelli, P., Frosch, M. P., Stephanopoulos, G., et al. (2004). Molecular signature of late-stage human ALS revealed by expression profiling of postmortem spinal cord gray matter. *Physiol. Genomics* 16, 229–239. doi: 10.1152/physiolgenomics.00087.2001

Dasgupta, S., Zhou, Y., Jana, M., Banik, N. L., and Pahan, K. (2003). Sodium phenylacetate inhibits adoptive transfer of experimental allergic encephalomyelitis in SJL/J mice at multiple steps. *J. Immunol.* 170, 3874–3882. doi: 10.4049/jimmunol.170.7.3874

Delile, J., Rayon, T., Melchionda, M., Edwards, A., Briscoe, J., and Sagner, A. (2019). Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development* 146:dev173807. doi: 10.1242/dev.173807

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

Eley, G. D., Reiter, J. L., Pandita, A., Park, S., Jenkins, R. B., Maihle, N. J., et al. (2002). A chromosomal region 7p11.2 transcript map: its development and application to the study of EGFR amplicons in glioblastoma. *Neuron. Oncol.* 4, 86–94. doi: 10.1093/neuonc/4.2.86

Elsaeidi, F., Bemben, M. A., Zhao, X. F., and Goldman, D. (2014). Jak/Stat signaling stimulates zebrafish optic nerve regeneration and overcomes the inhibitory actions of Socs3 and Sfpq. *J. Neurosci.* 34, 2632–2644. doi: 10.1523/JNEUROSCI.3898-13.2014

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Ersoz, G., Ficicilar, H., Pasin, M., Yorgancioglu, R., and Yavuzer, S. (1999). Platelet aggregation in traumatic spinal cord injury. *Spinal Cord* 37, 644–647. doi: 10.1038/sj.sc.3100903

Fang, Z., Wu, L., Dai, H., Hu, P., Wang, B., Han, Q., et al. (2020). The role of vesicular overexpressed in cancer pro-survival protein 1 in hepatocellular carcinoma proliferation. *Cancer Biomark* 28, 9–20. doi: 10.3233/CBM-190574

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* 7, 1728–1740. doi: 10.1038/nprot.2012.101

Ferreira, M. R., Santos, G. A., Biagi, C. A., Silva Junior, W. A., and Zambuzzi, W. F. (2021). GSVA score reveals molecular signatures from transcriptomes for biomaterials comparison. *J. Biomed. Mater. Res. A.* 109, 1004–1014. doi: 10.1002/jbm.a.37090

Forner, S., Martini, A. C., Andrade, E. L., and Rae, G. A. (2016). Neuropathic pain induced by spinal cord injury: Role of endothelin ETA and ETB receptors. *Neurosci. Lett.* 617, 14–21. doi: 10.1016/j.neulet.2016.02.005

Fox, A. H., and Lamond, A. I. (2010). Paraspeckles. *Cold Spring Harb. Perspect. Biol.* 2:a000687. doi: 10.1101/cshperspect.a000687

Gao, C., Pang, M., Zhou, Z., Long, S., Dong, D., Yang, J., et al. (2015). Epidermal growth factor receptor-coamplified and overexpressed protein (VOPP1) is a putative oncogene in gastric cancer. *Clin. Exp. Med.* 15, 469–475. doi: 10.1007/s10238-014-0320-7

Gao, T. Y., Huang, F. F., Xie, Y. Y., Wang, W. Q., Wang, L. D., Mu, D., et al. (2021). Dynamic changes in the systemic immune responses of spinal cord injury model mice. *Neural. Regen. Res.* 16, 382–387. doi: 10.4103/1673-5374.290910

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

GBD 2016 Neurology Collaborators (2019). Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 18, 459–480.

Ghlissi, Z., Krichen, F., Kallel, R., Amor, I. B., Boudawara, T., Gargouri, J., et al. (2019). Sulfated polysaccharide isolated from Globularia alypum L.: Structural characterization, in vivo and in vitro anticoagulant activity, and toxicological profile. *Int. J. Biol. Macromol.* 123, 335–342. doi: 10.1016/j.ijbiomac.2018.11.044

Grubert, F., Srivas, R., Spacek, D. V., Kasowski, M., Ruiz-Velasco, M., Sinnott-Armstrong, N., et al. (2020). Landscape of cohesin-mediated chromatin loops in the human genome. *Nature* 583, 737–743. doi: 10.1038/s41586-020-2151-x

Guenther, M. G., Frampton, G. M., Soldner, F., Hockemeyer, D., Mitalipova, M., Jaenisch, R., et al. (2010). Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* 7, 249–257. doi: 10.1016/j.stem.2010.06.015

Hakimi, M. A., Bochar, D. A., Chenoweth, J., Lane, W. S., Mandel, G., and Shiekhattar, R. (2002). A core-BRAF35 complex containing histone deacetylase mediates repression of neuronal-specific genes. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7420–7425. doi: 10.1073/pnas.112008599

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* 14:7. doi: 10.1186/1471-2105-14-7

Hasler, R., Kerick, M., Mah, N., Hultschig, C., Richter, G., Bretz, F., et al. (2011). Alterations of pre-mRNA splicing in human inflammatory bowel disease. *Eur. J. Cell Biol.* 90, 603–611. doi: 10.1016/j.ejcb.2010.11.010

Hayashi, M., Ueyama, T., Nemoto, K., Tamaki, T., and Senba, E. (2000). Sequential mRNA expression for immediate early genes, cytokines, and neurotrophins in spinal cord injury. *J. Neurotrauma.* 17, 203–218. doi: 10.1089/neu.2000.17.203

Hennigan, A., O'callaghan, R. M., and Kelly, A. M. (2007). Neurotrophins and their receptors: roles in plasticity, neurodegeneration and neuroprotection. *Biochem. Soc. Trans.* 35, 424–427. doi: 10.1042/BST0350424

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Jin, Y., Bouyer, J., Haas, C., and Fischer, I. (2015). Evaluation of the anatomical and functional consequences of repetitive mild cervical contusion using a model of spinal concussion. *Exp. Neurol.* 271, 175–188. doi: 10.1016/j.expneurol.2015.06.001

Jones, T. B. (2014). Lymphocytes and autoimmunity after spinal cord injury. *Exp. Neurol.* 258, 78–90. doi: 10.1016/j.expneurol.2014.03.003

Kang, Y., Kim, Y. W., Kang, J., and Kim, A. (2021). Histone H3K4me1 and H3K27ac play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers. *FASEB J.* 35:e21781. doi: 10.1096/fj.202100488R

Karin, M., and Ben-Neriah, Y. (2000). Phosphorylation meets ubiquitination: the control of NF-[kappa]B activity. *Annu. Rev. Immunol.* 18, 621–663. doi: 10.1146/annurev.immunol.18.1.621

Karin, M., Cao, Y., Greten, F. R., and Li, Z. W. (2002). NF-kappaB in cancer: from innocent bystander to major culprit. *Nat. Rev. Cancer* 2, 301–310. doi: 10.1038/nrc780

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182–U165. doi: 10.1038/nature09033

Kohl, M., Wiese, S., and Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. *Methods Mol. Biol.* 696, 291–303. doi: 10.1007/978-1-60761-987-1_18

Konishi, Y., Lehtinen, M., Donovan, N., and Bonni, A. (2002). Cdc2 phosphorylation of BAD links the cell cycle to the cell death machinery. *Mol. Cell* 9, 1005–1016. doi: 10.1016/S1097-2765(02)00524-5

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Li, J. M., Zhang, H., and Zuo, Y. J. (2018). MicroRNA-218 alleviates sepsis inflammation by negatively regulating VOPP1 via JAK/STAT pathway. *Eur. Rev. Med. Pharmacol. Sci.* 22, 5620–5626.

Li, W., Hu, Y., Oh, S., Ma, Q., Merkurjev, D., Song, X., et al. (2015). Condensin I and II Complexes License Full Estrogen Receptor alpha-Dependent Enhancer Activation. *Mol. Cell* 59, 188–202. doi: 10.1016/j.molcel.2015.06.002

Li, W., Notani, D., and Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* 17, 207–223. doi: 10.1038/nrg.2016.4

Li, Y. J., Zhang, W., Xia, H., Zhang, B. S., Chen, P., Zhao, Y. L., et al. (2017). miR-218 suppresses epithelial-to-mesenchymal transition by targeting Robo1 and Ecop in lung adenocarcinoma cells. *Fut. Oncol.* 13, 2571–2582. doi: 10.2217/fon-2017-0398

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322. doi: 10.1038/nature08514

Liu, B., and Neufeld, A. H. (2007). Activation of epidermal growth factor receptors in astrocytes: from development to neural injury. *J. Neurosci. Res.* 85, 3523–3529. doi: 10.1002/jnr.21364

Lowery, L. A., Rubin, J., and Sive, H. (2007). Whitesnake/sfpq is required for cell survival and neuronal development in the zebrafish. *Dev. DYN* 236, 1347–1357. doi: 10.1002/dvdy.21132

Maynard, F. M., Bracken, M. B., Creasey, G., Ditunno, J. F., Donovan, W. H., Ducker, T. B., et al. (1997). International standards for neurological and functional classification of spinal cord injury. *Spinal. Cord* 35, 266–274. doi: 10.1038/sj.sc.3100432

Murphy, K. M., and Reiner, S. L. (2002). The lineage decisions of helper T cells. *Nat. Rev. Immunol.* 2, 933–944. doi: 10.1038/nri954

Nakae, A., Nakai, K., Yano, K., Hosokawa, K., Shibata, M., and Mashimo, T. (2011). The animal model of spinal cord injury as an experimental pain model. *J. Biomed. Biotechnol.* 2011:939023. doi: 10.1155/2011/939023

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Park, S., and James, C. D. (2005). ECop (EGFR-coamplified and overexpressed protein), a novel protein, regulates NF-kappaB transcriptional activity and associated apoptotic response in an IkappaBalpha-dependent manner. *Oncogene* 24, 2495–2502. doi: 10.1038/sj.onc.1208496

Phillips, A., Matin, N., West, C., Zheng, A., Galea, L., Dorrance, A., et al. (2015). Autonomic Dysreflexia Impairs Cerebrovascular Health and Cognition in Experimental Spinal Cord Injury. *FASEB J.* 29, 59–70. doi: 10.1096/fasebj.29.1_supplement.800.10

Pineau, I., and Lacroix, S. (2007). Proinflammatory cytokine synthesis in the injured mouse spinal cord: multiphasic expression pattern and identification of the cell types involved. *J. Comp. Neurol.* 500, 267–285. doi: 10.1002/cne.21149

Pouw, M. H., Hosman, A. J., Van Kampen, A., Hirschfeld, S., Thietje, R., and Van De Meent, H. (2011). Is the outcome in acute spinal cord ischaemia different from that in traumatic spinal cord injury? A cross-sectional analysis of the neurological and functional outcome in a cohort of 93 paraplegics. *Spinal. Cord* 49, 307–312. doi: 10.1038/sc.2010.114

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Robson, J. P., Wagner, B., Glitzner, E., Heppner, F. L., Steinkellner, T., Khan, D., et al. (2018). Impaired neural stem cell expansion and hypersensitivity to epileptic seizures in mice lacking the EGFR in the brain. *FEBS J.* 285, 3175–3196. doi: 10.1111/febs.14603

Romano, R., and Bucci, C. (2020). Role of EGFR in the Nervous System. *Cells* 9:1887. doi: 10.3390/cells9081887

Ropper, A. E., Zeng, X., Anderson, J. E., Yu, D., Han, I., Haragopal, H., et al. (2015). An efficient device to experimentally model compression injury of mammalian spinal cord. *Exp. Neurol.* 271, 515–523. doi: 10.1016/j.expneurol.2015.07.012

Rudman, M. D., Choi, J. S., Lee, H. E., Tan, S. K., Ayad, N. G., and Lee, J. K. (2018). Bromodomain and extraterminal domain-containing protein inhibition attenuates acute inflammation after spinal cord injury. *Exp. Neurol.* 309, 181–192. doi: 10.1016/j.expneurol.2018.08.005

Ryu, H., Lee, J., Olofsson, B. A., Mwidau, A., Dedeoglu, A., Escudero, M., et al. (2003). Histone deacetylase inhibitors prevent oxidative neuronal death independent of expanded polyglutamine repeats via an Sp1-dependent pathway. *Proc. Natl. Acad. Sci. U.S.A.* 100, 4281–4286. doi: 10.1073/pnas.0737363100

Selvarajah, S., Hammond, E. R., and Schneider, E. B. (2015). Trends in Traumatic Spinal Cord Injury. *JAMA* 314:1643. doi: 10.1001/jama.2015.11194

Shav-Tal, Y., and Zipori, D. (2002). PSF and p54(nrb)/NonO–multi-functional nuclear proteins. *FEBS Lett.* 531, 109–114. doi: 10.1016/S0014-5793(02)03447-6

Smith, E., O'reilly, A., Synnott, K., Morris, S., and Timlin, M. (2015). Review of Time to Surgical Decompression in Traumatic Spinal Cord Injured Patients. *Ir. Med. J.* 108, 265–267.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3:Article3. doi: 10.2202/1544-6115.1027

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 23, 3251–3253. doi: 10.1093/bioinformatics/btm369

Wang, H., Zou, J., Zhao, B., Johannsen, E., Ashworth, T., Wong, H., et al. (2011). Genome-wide analysis reveals conserved and divergent features of Notch1/RBPJ binding in human and murine T-lymphoblastic leukemia cells. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14908–14913. doi: 10.1073/pnas.1109023108

Wang, X. Y., Smith, D. I., Frederick, L., and James, C. D. (1998). Analysis of EGF receptor amplicons reveals amplification of multiple expressed sequences. *Oncogene* 16, 191–195. doi: 10.1038/sj.onc.1201476

Wang, Y., Wang, R., Zhang, S., Song, S., Jiang, C., Han, G., et al. (2019). iTALK: an R Package to Characterize and Illustrate Intercellular Communication. *bioRxiv* [preprint]. doi: 10.1101/507871

West, J. A., Davis, C. P., Sunwoo, H., Simon, M. D., Sadreyev, R. I., Wang, P. I., et al. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell.* 55, 791–802. doi: 10.1016/j.molcel.2014.07.012

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., et al. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153, 307–319. doi: 10.1016/j.cell.2013.03.035

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19:15. doi: 10.1186/s13059-017-1382-0

Xiao, B., Liu, L., Li, A., Xiang, C., Wang, P., Li, H., et al. (2020). Identification and Verification of Immune-Related Gene Prognostic Signature Based on ssGSEA for Osteosarcoma. *Front. Oncol.* 10:607622. doi: 10.3389/fonc.2020.607622

Yao, J. S. (1992). Deep vein thrombosis in spinal cord-injured patients. Evaluation and assessment. *Chest* 102, 645S–648S. doi: 10.1378/chest.102.6_Supplement.645S

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, T., Jiang, M., Chen, L., Niu, B., and Cai, Y. (2013). Prediction of gene phenotypes based on GO and KEGG pathway enrichment scores. *Biomed. Res. Int.* 2013:870795. doi: 10.1155/2013/870795

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 47, D721–D728. doi: 10.1093/nar/gky900

Zhang, Y., Guan, Z., Reader, B., Shawler, T., Mandrekar-Colucci, S., Huang, K., et al. (2013). Autonomic dysreflexia causes chronic immune suppression after spinal cord injury. *J. Neurosci.* 33, 12970–12981. doi: 10.1523/JNEUROSCI.1974-13.2013

Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., et al. (2019). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* 47, D729–D735. doi: 10.1093/nar/gky1094

Zhu, Y., Soderblom, C., Krishnan, V., Ashbaugh, J., Bethea, J. R., and Lee, J. K. (2015). Hematogenous macrophage depletion reduces the fibrotic scar and increases axonal growth after spinal cord injury. *Neurobiol. Dis.* 74, 114–125. doi: 10.1016/j.nbd.2014.10.024

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Huang, Wang, Zhu, Xian, Huang, Cheng and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 12 October 2021
doi: 10.3389/fcell.2021.746932

Identification of an Immune-Related Prognostic Gene CLEC5A Based on Immune Microenvironment and Risk Modeling of Ovarian Cancer

*Jiacheng Shen[1], Tingwei Liu[1], Jia Lv[2] and Shaohua Xu[1]**

[1] Department of Gynecology, Shanghai First Maternity and Infant Hospital, School of Medicine, Tongji University, Shanghai, China, [2] Department of Obstetrics and Gynecology, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China

Objective: To understand the immune characteristics of the ovarian cancer (OC) microenvironment and explore the differences of immune-related molecules and cells to establish an effective risk model and identify the molecules that significantly affected the immune response of OC, to help guide the diagnosis.

Methods: First, we calculate the TMEscore which reflects the immune microenvironment, and then analyze the molecular differences between patients with different immune characteristics, and determine the prognostic genes. Then, the risk model was established by least absolute shrinkage and selection operator (LASSO) analysis and combined with clinical data into a nomogram for diagnosis and prediction. Subsequently, the potential gene CLEC5A influencing the immune response of OC was identified from the prognostic genes by integrative immune-stromal analysis. The genomic alteration was explored based on copy number variant (CNV) and somatic mutation data.

Results: TMEscore was a prognostic indicator of OC. The prognosis of patients with high TMEscore was better. The risk model based on immune characteristics was a reliable index to predict the prognosis of patients, and the nomogram could comprehensively evaluate the prognosis of patients. Besides, CLEC5A was closely related to the abundance of immune cells, immune response, and the expression of immune checkpoints in the OC microenvironment. OC cells with high expression of CLEC5A increased the polarization of M2 macrophages. CLEC5A expression was significantly associated with TTN and CDK12 mutations and affected the copy number of tumor progression and immune-related genes.

Conclusion: The study of immune characteristics in the OC microenvironment and the risk model can reveal the factors affecting the prognosis and guide the clinical hierarchical treatment. CLEC5A can be used as a potential key gene affecting the immune microenvironment remodeling of OC, which provides a new perspective for improving the effect of OC immunotherapy.

Keywords: ovarian cancer, immune microenvironment remodeling, CLEC5A, multi-omics, immunotherapy, prognostic modeling

OPEN ACCESS

Edited by:
Chuan-Xing Li,
Karolinska Institutet (KI), Sweden

Reviewed by:
Zhi Feng,
Modern Medical Research Center, First People's Hospital of Changzhou, China
Xuanmao Jiao,
Baruch S. Blumberg Institute, United States

****Correspondence:***
Shaohua Xu
xushaohua@tongji.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *25 July 2021*
Accepted: *16 September 2021*
Published: *12 October 2021*

Citation:
Shen J, Liu T, Lv J and Xu S (2021) Identification of an Immune-Related Prognostic Gene CLEC5A Based on Immune Microenvironment and Risk Modeling of Ovarian Cancer. Front. Cell Dev. Biol. 9:746932. doi: 10.3389/fcell.2021.746932

INTRODUCTION

Ovarian cancer (OC) is a kind of malignant tumor with a poor prognosis, and its mortality rate ranks second in gynecological cancer mortality (Bray et al., 2018). In recent years, with the improvement of the molecular basis of immune recognition and immune regulation of tumor cells, many kinds of cancer have been treated with immunotherapy. The immune microenvironment of tumor tissue can reflect the response rate to immunotherapy and chemotherapy (Rosenberg et al., 2016; Jiang et al., 2018). Understanding the immune regulatory network in the OC microenvironment will promote the effective use of immunotherapy.

The tumor microenvironment (TME) is a highly complex system. Tumor cells coexist with immune cells and non-immune cells and establish an interaction with them, which affects the tumor's development and outcome. The immune cell components in TME are the basis for determining tumor fate and the cells' ability to invade and metastasize (Melaiu et al., 2019; Vitale et al., 2019; Xiao et al., 2019). Their composition and interactions within the TME are closely related to clinical outcomes in cancer patients. Recent studies have shown that TME played a crucial role in cancer progression and treatment response, and advances in TME understanding have also contributed to the development of advanced cancer treatment methods (Jiang et al., 2018; Zeng et al., 2018). The composition of resident cell types in TME varies from different cancer patients. The number of $CD8^+$ T cells, $CD4^+$ T cells, macrophages, and tumor-associated fibroblasts in TME is associated with the clinical outcome of many malignant tumors, such as gastric cancer, urothelial cancer, lung cancer, and breast cancer (Fridman et al., 2017; Mantovani et al., 2017; Nishino et al., 2017; Mariathasan et al., 2018). There has been evidence that OC was an immunogenic tumor, and some studies have confirmed the prognostic value of the immune system in OC (Kandalaft et al., 2011). Therefore, understanding the differences of the immune microenvironment in OC may be an essential step in finding prognostic markers, stratifying patients before treatment, and prolonging life expectancy in OC patients.

In this study, we evaluated the TME characteristics of patients by calculating the TMEscore. Functional enrichment analysis showed that the TME signature could reflect the prognosis of OC patients. However, in the clinic, it is not realistic to calculate the TMEscore for each patient. Therefore, we integrate multiple differentially expressed prognostic genes based on the TMEscore into a model that can reflect the characteristics of TME, rather than using a single prognostic gene as an indicator, which will significantly improve the predictive value. We established a risk model by least absolute shrinkage and selection operator (LASSO) regression and further established a nomogram including risk score to predict the prognosis of OC patients. The risk model can not only reflect the TME characteristics but also predict the sensitivity of patients to platinum drugs and guide the clinical treatment of OC patients. In addition, we combined with the immune and stromal score, screened out the potential gene CLEC5A which was most likely to participate in the immune response of OC from these prognostic genes, and analyzed its expression in OC. Then, we further explored the effect of CLEC5A on M2 macrophage polarization through a co-culture experiment. This study opens up a new field of vision for the exploratory study of TME in OC and the prediction of the prognosis of OC patients.

Abbreviations: TCGA, The Cancer Genome Atlas; ICGC, International Cancer Genome Consortium; OC, ovarian cancer; TME, tumor microenvironment; DEG, differentially expressed gene; GO, Genetic ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; OS, overall survival; ROC, receiver operating characteristic; AUC, area under the curve; LASSO, least absolute shrinkage and selection operator; TAM, tumor-associated macrophages, CAM, cell adhesion molecules; TF, transcription factor; ncRNA, noncoding RNA; DCA, decision curve analysis; HR, hazard ratio; CI, confidence interval; PPI, protein–protein interaction; VIF, variance inflation factor; TMB, tumor mutational burden.

MATERIALS AND METHODS

Data Extraction and Processing

RNA-Seq Data

All the data in the current study were from the databases. The RNA-seq expression profile data and clinical information of OC patients were downloaded from the The Cancer Genome Atlas (TCGA) database.[1] The overall parameters for downloading data were as follows: "Primary Site" was ovary, "Program" was TCGA, "Project" was TCGA-OV, and additional parameters for transcriptome data were the following: "Data Category" was transcriptome profiling, "Data Type" was gene expression quantification, "Experimental Strategy" was RNA-Seq, "Workflow Type" was HISeq-FPKM, and additional parameters for clinical data download were the following: "Data Category" was clinical, and "Data Type" was clinical supply. All samples were excluded for patients without survival information.

The RNA-seq expression profile data and clinical information of OC patients were downloaded from the International Cancer Genome Consortium (ICGC) OV-AU (Ovarian cancer-Australia) database,[2] and samples were excluded for patients without survival information, which are used for subsequent validation. The statistical information of the preprocessed dataset is shown in **Supplementary Table 1**, and the detail clinical information of each patient with OC from these two databases is shown in **Supplementary Tables 2, 3**.

The transcriptome-level data of datasets GSE62873, GSE69207, and GSE146553 downloaded from the GEO database[3] were used to verify the correlation between the expression level of CLEC5A and the proportion of immune cells after removing batch effects.

Noncoding RNA/Transcription Factor–mRNA Interaction Data

The noncoding RNA (ncRNA)–mRNA interaction and transcription factor (TF)-mRNA interaction data were

[1] https://tcga-data.nci.nih.gov/tcga/

[2] https://icgc.org/

[3] https://www.ncbi.nlm.nih.gov/geo/

downloaded from the RAID database (v2.0)[4] and TRRUST v2 database.[5] The interaction pairs of all ncRNAs/TFs and module genes were counted, and then the interaction pairs of each ncRNA/TFs and genes in or outside the module were counted.

Genomic Variation Data

The copy number variant (CNV) data of OC patients was downloaded from the TCGA database using the "TCGABIOLinks" package, and the data type was Masked Number Segment. The SNP6 GRCh38 Remapped Probeset File[6] was used as the marker file, and the CNV interval was mapped to the corresponding gene. Somatic single-nucleotide variant (SSNV) data was the Mutect2 version in the TCGA database,[7] which was the whole-exome sequencing data.

Generation of TMEScore, Stromal Score, and Immune Score

Based on TCGA-OV RNA-seq data and clinical information, the "edger voom" algorithm was used to remove heteroscedasticity. Then, the proportions of 22 immune cell types in the OC were estimated by CIBERSORT (Newman et al., 2015). Next, we performed three unsupervised methods (consensus cluster, elbow method, and gap statistics) using the "factoextra" package and "ConsensusClusterPlus" package to explore the best number of clusters, and all three clustering methods were iterated 1,000 times to increase the stability of clustering. After the best number of clusters was selected, the k-means method was used to divide samples into two TME clusters and obtained the differentially expressed genes (DEGs) between two clusters using the "limma" package. We applied the consensus clustering method to explore sample classification again and used the chi-square test to compare the consistency of the two classifications (p = 2.089e-16). The random forest algorithm was performed using the "randomForest" package to reduce redundant genes and identify the core DEGs. The Cox regression model was used to classify the core genes and finally calculated the TMEscore of each patient using the formula as follows, which was similar to the algorithm mentioned in the previous study (Zeng et al., 2019).

$$\boldsymbol{TMEscore} = \sum \log_2(\mathbf{X}+\mathbf{1}) - \sum \log_2(\mathbf{Y}+\mathbf{1})$$

where X is the gene with the positive Cox coefficient, and Y means the gene with the negative Cox coefficient. The detailed procession for generating TMEscore is shown in **Supplementary Figure 1**.

The proportion of immune-stromal components of the TME which were represented in the form of the immune score and stromal score, respectively, in each OC sample was estimated utilizing the "estimate" package in R software.

Identification of Differentially Expressed Genes Based on TMEscore, Stromal Score, and Immune Score

The OC samples were subdivided into the high- or low-score groups based on the TMEscore, stromal score, and immune score, respectively. The "limma" package was used to analyze the DEGs, and DEGs with filtering criteria of adj. p-values < 0.05 and $|$ Fold Change$| > 1.5$ were selected for further analysis.

Survival and Functional Enrichment Analysis

Survival analysis refers to the method of analyzing and inferring the survival time of organisms or people according to the data obtained from experiments or surveys, exploring the relationship between survival time, outcome, and many influencing factors. Survival analysis of DEGs was performed by R using the "survminer" package. Genes with a log-rank p-value < 0.05 were considered as prognostic genes.

Functional enrichment analysis of related genes was performed using the "ClusterProfiler" package to identify significantly enriched Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. An adj. p-value < 0.05 was considered statistically significant. As for single-gene function analysis, the "GSEA" package was used in R software and the FDR q-value < 0.05 was set as the threshold to identify significantly important function terms.

Construction of Protein–Protein Interaction and Pivot Network

The STRING database[8] was used to build the protein–protein interaction (PPI) network of prognostic genes, and Cytoscape software (V3.6.0) was used to rebuild it. To identify closely connected modules in the network, the key module of the network was mined using the "MCODE" with criteria as MCODE score ≥ 3 and degree > 10. The rank of hub genes was identified by "Cytohubba" in Cytoscape.

Pivot consists of participants that significantly regulate modules in OC tumorigenesis, including ncRNA and TF. The pivot is defined as follows: (1) the pivot has at least two interactions with the module gene and (2) the p-value of the significance analysis of the interaction between the pivot in which each module should be less than or equal to 0.05 (Wu et al., 2015). The hypergeometric test method was used for the significance analysis (see **Supplementary Material**).

Construction of the Prognostic Risk Model and External Validation

The LASSO Cox regression model analysis was performed by the "glmnet" package to optimize prognostic genes in patients with OC (Simon et al., 2011). We resampled the dataset 1,000 times, and the best value of the penalty parameter λ was determined by 10-fold cross-validations. Meanwhile, variables with potential collinearity were excluded from the regression model (correlation

[4]http://www.rna-society.org/raid2/

[5]http://www.grnpedia.org/trrust/

[6]https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files

[7]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[8]https://string-db.org/

< 0.5, variance inflation factor (VIF) <2). After core prognostic genes were selected, the multivariate Cox regression model was used to determine the regression coefficients (β) of each gene to establish a risk score formula:

$$\textit{The Risk Score} = \sum(\beta * \textit{expression level})$$

All samples were subdivided into high- or low-risk according to the Risk Score of each sample. To evaluate the prognostic risk model's predictive value, the time-dependent receiver operating characteristic (ROC) curve was performed using the "survivalROC" package. Finally, the risk model was validated in the ICGC OV-AU dataset.

Independence of Risk Model and Construction of Nomogram

To determine the independence of the prognostic risk model in predicting the prognosis of OC patients, we performed a univariate and multivariate Cox regression analysis of several variables, which included risk score and calculated hazard ratio (HR), as well as 95% confidence intervals (CI). By combining all independent prognostic factors, a nomogram was constructed to evaluate the probability of 2-, 3-, 4-, and 5-year overall survival (OS) for patients with OC. The graph of prediction probability and observation rate was then drawn using the "rms" package, and the calibration curve of the prognostic model was graphically evaluated, which overlapped with the diagonal line, indicating that the model was utterly consistent. The decision curve analysis (DCA) was performed by the "rmda" package to test the clinical efficacy of the nomogram.

Genomic Variation Analysis

Somatic variation data were retained in the mutation note format. Differentially mutated genes were identified by the "maftools" package with $p < 0.05$ as the significant threshold. The online tool GISTIC2.0[9] was used to analyze CNV data. Specific amplifications and deletions of chromatin sites in each immune subtype were then selected for subsequent analysis.

Clinical Specimens and Immunohistochemistry

The samples were collected from the Shanghai First Maternal and Infant Hospital and have obtained the informed consents (the ethical certification number: KS1748) with permission from the Medical Ethics Committee of Shanghai First Maternity and Infant Hospital. The diagnoses of acquired samples were all carefully certified and checked by experienced pathologists. The samples were fixed in 10% formalin, dehydrated, embedded and sectioned, and then immunohistochemically stained using VECTASTAIN Elite ABC Rabbit IgG Mini-PLUS Kit (PK-8501, Vector Laboratories, United States) with anti-CLEC5A antibody (ab203200, Abcam, Cambridge, United Kingdom) and hematoxylin (Wuhan Servicebio Technology Co., Ltd., Wuhan, China).

[9]https://cloud.genepattern.org/

Cell Culture and Co-culture Assay

The human leukemic cell line: THP-1 cell and human OC cell lines: SKOV3, HEY (American Type Culture Collection, Manassas, VA, United States), were cultured in RPMI 1640 medium (Wuhan Servicebio Technology Co., Ltd., Wuhan, China) containing 10% fetal bovine serum (Biological Industries, Beit HaEmek, Israel) and 1% penicillin/streptomycin (Wuhan Servicebio Technology Co., Ltd., Wuhan, China) at 37°C in a humidified atmosphere with 5% CO_2. The THP-1 cells (5×10^4 cells/100 μl) were seeded in a lower chamber and differentiated into macrophage by adding 100 ng/ml of phorbol 12-myristate-13-acetate (PMA, MCE, Chengdu, China) for 48 h. After THP-1 cells were differentiated into M0 macrophages, the M0 macrophages were treated with 20 ng/ml IL-4 (Peprotech, Rocky Hill, NJ, United States) plus 20 ng/ml IL-10 (Peprotech, Rocky Hill, NJ, United States) for 48 h to differentiate into M2 macrophages (Murray et al., 2014). The OC cells were transfected with NC or CLEC5A overexpression plasmid (ordered from Public Protein/Plasmid Library, China) using Lipofectamine 2000 Reagent (Invitrogen, United States) for 24 h, and the transfection efficiency was verified by real-time-qPCR (RT-qPCR). After transfection, the OC cells (3×10^4 cells/100 μl) were then added to the upper insert (0.4 μm pore; 6-well Transwell, LABSELECT, Zhejiang, China) for co-culture for 24 h. In addition, we also transfected plasmids into M2 macrophages to detect the effect of CLEC5A on the polarization of M2 macrophages.

RNA Extraction and Real-Time PCR Analysis

Total RNA from cells was extracted using the TRIzol reagent (TAKARA BIO INC., Japan) and then synthesized to complementary DNA (cDNA) using the 5 × ALL-IN-One RT Master Mix Kit (Applied Biological Materials Inc., Canada). TB Green Premix Ex Taq Kit (TAKARA BIO INC., Japan) was used for real-time PCR. The ΔΔCt values were normalized to GAPDH, and relative quantification of gene expression was compared to the NC group. The primers used in this study are listed in **Supplementary Table 7** and synthesized by Sangon Biotech (Shanghai) Co., Ltd., (Shanghai, China).

Statistical Analysis

Univariate survival analysis was performed using the Cox risk regression model, and a significance threshold of log-rank $p < 0.05$ was set to screen prognostic genes. The chi-square test was used to determine the significance of bias in the distribution of two clustering methods. The Wilcoxon rank test was used to determine significance in comparisons of two groups of continuous variables, the Kruskal–Wallis rank test was used for comparisons of more than two groups, and the Benjamini–Hochberg method was used to control the FDR. All the above analyses were performed using R version 3.6.0. Unless otherwise specified, $^{****}p < 0.0001$; $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and ns, no significance.

RESULTS

Quantification of Immune Infiltration and Identification of Molecular Characteristics in Patients With Ovarian Cancer

The workflow design of this study is shown in **Figure 1**. By integrating the expression data and clinical data (OS >1 month, precise pathological stage), 339 OC samples were obtained for the generation of the TMEscore (**Figure 2A**). The survival analysis showed that a high TMEscore was significantly associated with a longer survival time ($p = 0.0082$) (**Figure 2B**). However, there was no statistical significance between TMEscore and pathological stage ($p = 0.82$, one of the OC patients was a stage I patient) (**Supplementary Figure 2A**). Analysis of the proportion of immune cells in tumor tissue of two groups divided by the TMEscore showed that the proportions of M1 macrophages, $CD8^+$ T cells, and activated $CD4^+$ memory T cells were increased in the high TMEscore group (**Figures 2C,D**), as well as the IFN-gamma related genes (**Supplementary Figure 2B**). The stromal score and immune score calculated by ESTIMATE were also significantly correlated with the TMEscore (**Figure 2E**). To identify the difference of the molecular level between the different degrees of immune infiltration, 747 genes were screened as DEGs for follow-up analysis according to the high- and low-TMEscore group, including 661 upregulated genes and 86 downregulated genes (**Figure 2F** and **Supplementary Figure 2C**). These DEGs were mainly enriched in immune-related functions and pathways, such as leukocyte migration, immune response, phagocytosis, cytokine–cytokine receptor interaction, and cell adhesion molecules (CAMs; **Figures 2G,H**).

Identification of Prognostic Genes and Construction of the Protein–Protein Interaction Network

To further explore the prognosis-related molecular characteristics, 83 prognostic genes associated with OS ($p < 0.05$) were screened and the top six significant genes are shown in **Figure 3A**. These 83 prognostic genes still showed significant enrichment of immune-related functions or pathways such as lymphocyte-mediated immunity, B cell-mediated immunity, cytokine–cytokine receptor interaction, and CAMs (**Supplementary Figure 2D**). External validation of these prognostic genes was performed in the ICGC OV-AU dataset, in which 58 of the 83 genes were expressed. Further survival analysis showed that seven genes were identified as consistent with the results in the TCGA-OV dataset (**Supplementary Figure 2E**). Some genes have been validated in OC, so these genes were likely prognostic biomarkers in the OC immune microenvironment (Leffers et al., 2009; Ignacio et al., 2018; Nowak et al., 2018). To better understand the interactions among prognostic genes, all prognostic genes were analyzed in the STRING database to construct the PPI network (**Figure 3B**). Further analysis using the "MCODE" found two modules (**Figure 3B**). The module genes were significantly associated with immune regulation, such as CXCL9, CXCL11, CXCL13, CCR1, and CD27, as well as NAD metabolism-related genes, including IDO1 and CD38.

Construction of the Pivot Network of Noncoding RNA and Transcription Factor

Based on the 51,913 pairs of ncRNA–mRNA interaction from the RAID 2.0 database as the background of interaction, 27 ncRNAs regulating the prognostic genes were screened, including several ncRNAs associated with OC, such as SNORA64, ATP6V0CP3, SNHG11, and PPARGC1A (**Supplementary Table 4**). The TFs regulating the prognostic genes were screened based on the 9,396 pairs of TF–mRNA interaction in the TRRUST v2 database. There were 17 TFs that regulate prognostic genes, including several TFs associated with OC, such as *NFKB1*, *MYC*, *IRF1*, and *STAT1* (**Supplementary Table 5**). The interactions among prognostic genes, ncRNAs, and TFs were visualized by Cytoscape, as shown in **Figure 3C**.

Construction of Risk Model to Predict Prognosis in Ovarian Cancer Patients and External Validation

Least absolute shrinkage and selection operator regression analysis was performed with 10-fold cross-validation to further select the genes from 83 prognostic genes (**Figure 3D**). After excluding intergenic collinearity by correlation coefficient and VIF, nine genes were eventually selected to construct a risk model (**Supplementary Figure 3A**). According to the formula Risk Score = (−0.0780* APOL4) + (−0.2073* BTN3A3) + (0.1201* CCDC80) + (0.1958* CLEC5A) + (0.1917* COTL1) + (−0.1102* FOXA2) + (−0.2001* HLA-DOB) + (−0.3214* PLA2G2D) + (−0.1822* UBB), each patient's risk score was calculated. Then, 339 patients with risk scores were classified into the high- or low-risk group (**Figures 4A–C**), and the K–M survival curves of the two groups had a significant difference ($p < 0.0001$) (**Figure 3E**). The prognostic ability of the model was evaluated by ROC. As a result, the areas under the curve (AUC) of the risk model were 0.680, 0.653, 0.704, 0.728, and 0.753 for the 2-, 3-, 4-, 5-, and 6-year survival times, indicating that the prediction of the risk model had a good performance and the longer patients survive the better the fitness (**Figure 4D**). Further analysis of the relationship between the risk score and the response status of OC patients to platinum drugs found that the risk scores of platinum-resistant patients were significantly higher than those of platinum-sensitive patients ($p = 0.0032$) (**Figure 4E**).

For external validation in ICGC databases, each patient was scored using the same risk score formula and grouped by the same cutoff value. Consistent with the result of the TCGA-OV cohort, the OS of patients in the low-risk group was significantly higher than that in the high-risk group ($p = 0.012$) (**Figure 4F**). The ROC analysis showed that AUCs of the prognostic model were 0.717, 0.693, 0.691, and 0.692 at 2-, 3-, 4-, and 5- years, respectively, suggesting that this risk model was able to predict OS in OC patients (**Figure 4G**).

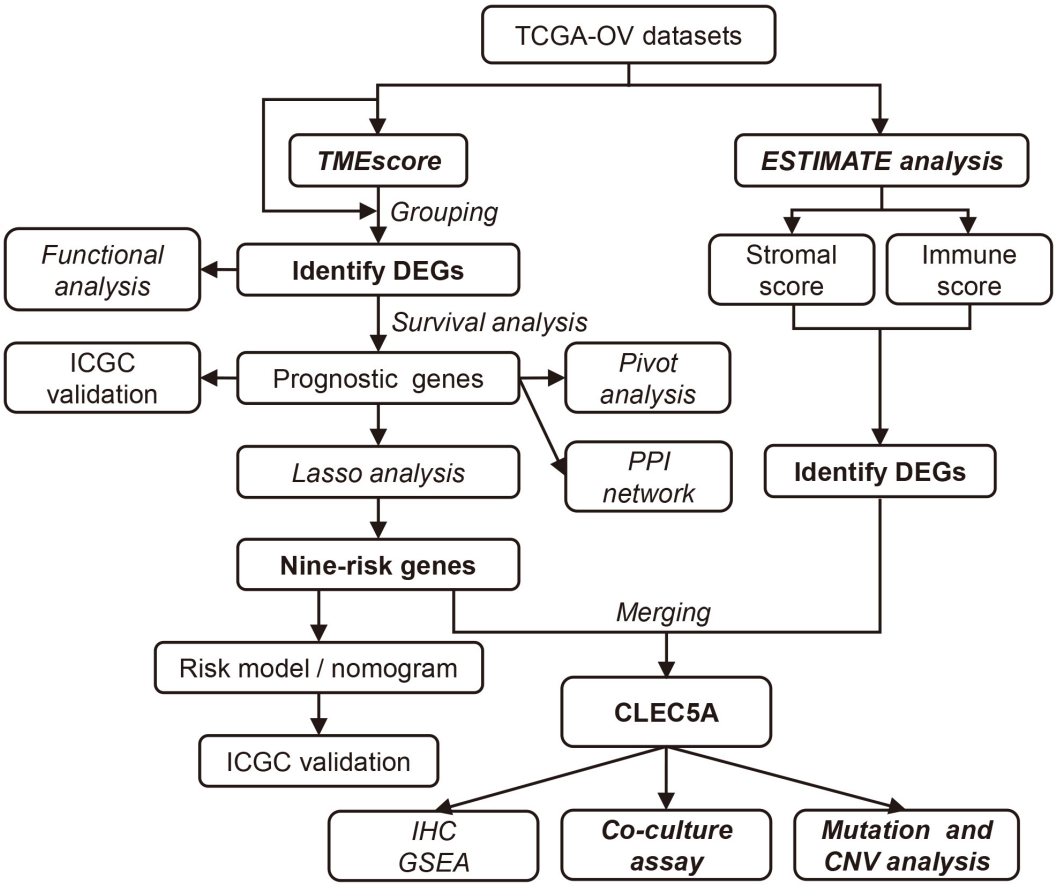


FIGURE 1 | The workflow of study design.

Independent Predictive Value of Risk Model and Nomogram Constructing

To evaluate the independent predictive value of the risk model, we performed univariate and multivariate Cox regression analyses on 325 OC patients with complete clinical data in the TCGA-OV dataset. We reassessed the risk model on this patient subset, showing still good benefits (**Supplementary Figure 2F**). Univariate Cox regression analysis showed that age and prognostic risk models had a certain predictive value for prognosis. In contrast, pathological stage, histological grade, and TMEscore were not independent prognostic factors (**Figure 4H** and **Supplementary Table 6**). We incorporated the age and risk model into multivariate Cox regression analysis, and the results showed that age and risk model were independent prognostic elements associated with OS (**Figure 4H** and **Supplementary Table 6**). To create a method to predict the individual survival likelihood of OC patients to guide clinical diagnosis and treatment, we subsequently developed a nomogram including two independent prognostic factors (the age and risk score) to predict the probability of the 2-, 3-, 4-, and 5-year OS in the TCGA-OV cohort (**Figure 4I**). Calibration curves used to visualize the nomogram's performance showed that the nomogram had an excellent predictive value (**Figure 4J** and **Supplementary Figure 3B**). The DCA curve was used to test the clinical efficacy, and the results showed that the net benefit rate of the nomogram-combined risk model and age increased significantly (**Figure 4K**). Moreover, the nomogram (age + risk score) also has a good predictive value in the ICGC dataset (**Supplementary Figure 3C**).

CLEC5A Is the Key Prognostic Gene of Ovarian Cancer Associated With Immunity

To find the key genes that affect the immune microenvironment of OC in the prognostic genes, we combined the stromal score and immune score obtained by ESTIMATE analysis based on TMEscore and finally screened CLEC5A and PLA2G2D (**Figure 5A**). There were differences in the expression level of CLCE5A between tumor tissues and normal tissues, but there was no significant difference between PLA2G2D (**Figure 5B**), and CLEC5A significantly affected the survival of patients (**Figure 5C**), indicating that CLEC5A was more suitable as a prognostic gene than PLA2G2D. There was no significant correlation between the expression of CLEC5A and other clinical features of OC, such as FIGO stage and histological grade (**Supplementary Figure 4A**). Immunohistochemistry showed

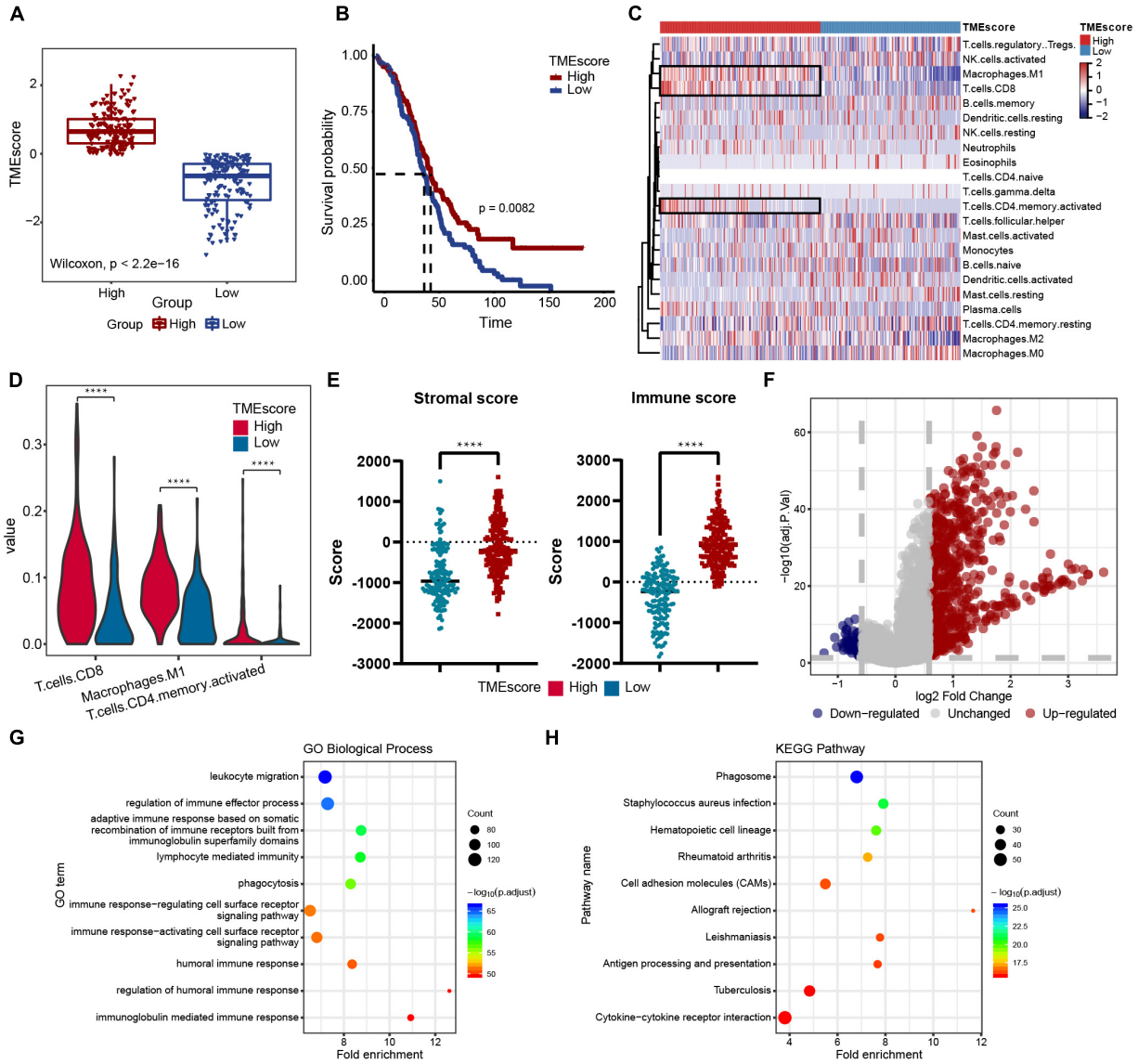


FIGURE 2 | Immune features of the ovarian cancer (OC) microenvironment and identification of molecular characteristic. **(A)** TMEscore of samples in the TCGA-OV dataset (high group: 181 samples, low group: 158 samples). Wilcoxon's test, $p < 2.2e\text{-}16$. **(B)** Overall survival (OS) Kaplan–Meier curves of TMEscore, log-rank test, $p = 0.0082$. **(C)** Heatmap of the 22 immune cell proportions in two TME group. **(D)** Violin plot of abundance of $CD8^+$ T cells, M1 macrophages, and activated $CD4^+$ memory T cells in the two groups. **(E)** Correlation between TMEscore and immune-stromal score. **(F)** Volcano plot of all differentially expressed genes (DEGs); the red dots represent upregulated genes, the blue dots represent downregulated genes, and the gray dots represent unchanged genes. **(G,H)** Genetic ontology (GO) biological process and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of DEGs; the fold enrichment represents the ratio of the gene ratio to background ratio. ****$p < 0.0001$.

that the expression of CLEC5A in OC tissue was significantly higher than that in normal ovarian tissue (**Figure 5D**). To further analyze the role of CLEC5A in the immune microenvironment of OC, we first divided CLEC5A-related genes into three groups. The functions of these three groups of genes were significantly related to the immune response, the process of the immune system, and the response to stimulation (**Figure 5E**). Then, we performed the single-gene GSEA analysis of CLEC5A, and the results showed that CLEC5A significantly regulated various immune cell pathways; in addition, it was also significantly related to the NF-κB pathway, Toll-like receptor pathway, cytokine, and chemokine pathway (**Figure 5F**).

CLEC5A Has the Potential to Remodel the Tumor Microenvironment of Ovarian Cancer and Influence the Immune Response

According to the expression level of CLEC5A, the patients were divided into high- and low-expression groups. The

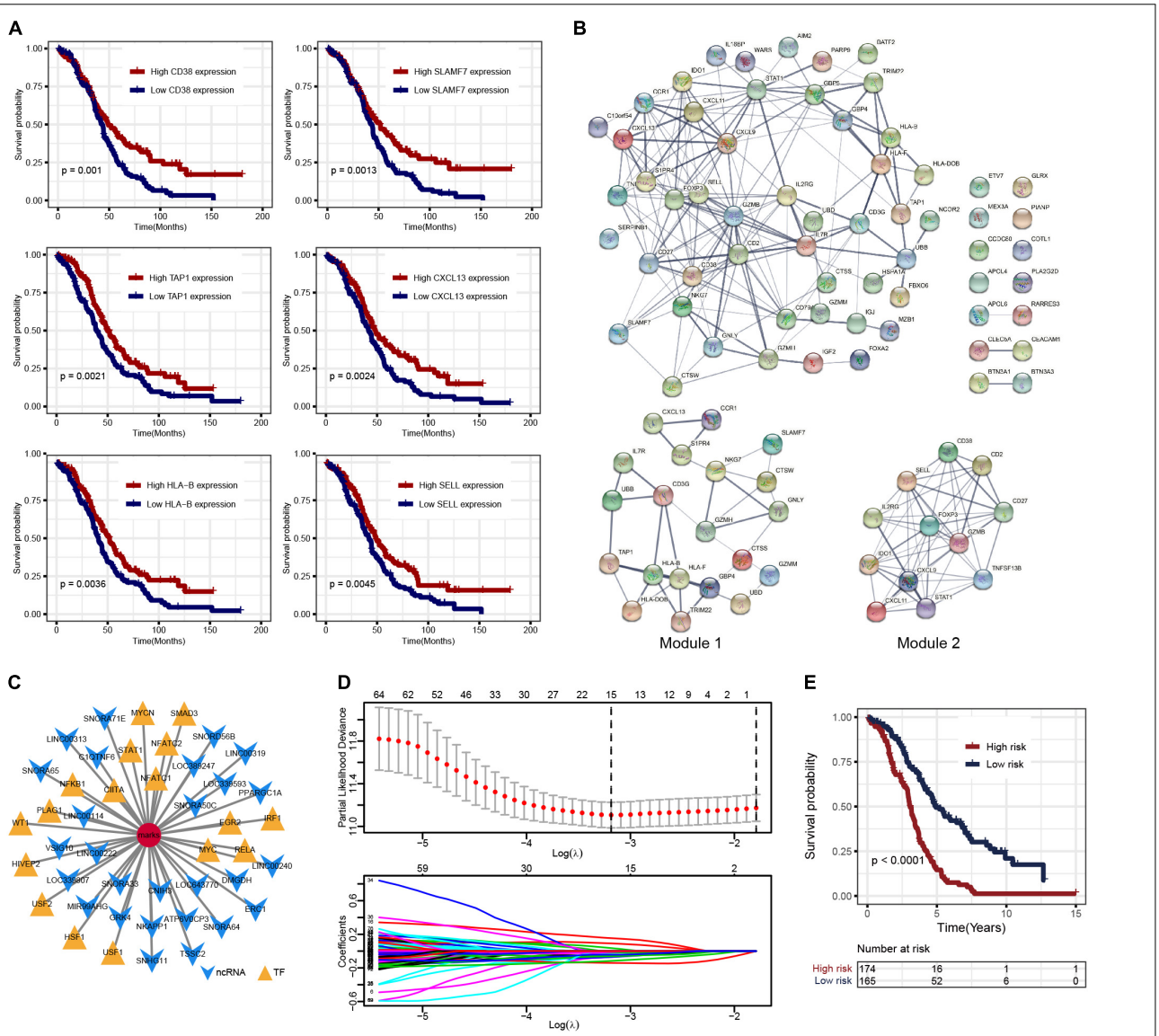


FIGURE 3 | (A) Overall survival Kaplan–Meier curves of DEGs (show top six genes). **(B)** The protein–protein interaction (PPI) network of prognostic genes and the two key modules. **(C)** Noncoding RNA (ncRNA) and transcription factor (TF)-pivot of prognostic genes; red circular mark represents 83 prognostic genes, yellow arrows represent ncRNAs, and green triangles represent TF. **(D)** Tenfold cross-validation for tuning parameter selection in the least absolute shrinkage and selection operator (LASSO) model. **(E)** The K–M survival curves show the OS based on the high- and low-risk patients.

relationship between the expression level of CLEC5A and immune cells in the immune microenvironment was further analyzed. It was found that the abundance of B cells naïve, T cells follicular helper, and NK cells activated in TME of patients with high CLEC5A expression was significantly lower than that of patients with low expression, while the abundance of M2 macrophages was much higher than that of patients with low expression (**Figures 6A,B**). The analysis of three datasets (GSE62873, GSE69207, and GSE146553) in the GEO database also showed the correlation between CLEC5A expression and M2 macrophages (**Supplementary Figure 4B**). To verify the correlation between CLEC5A and M2 macrophages, we carried out the co-culture assay of OC cells and M2 macrophages. Co-culture assay showed that OC cells overexpressing CLEC5A could enhance the M2 polarization of M2 macrophages (**Figure 6D** and **Supplementary Figure 4C**). Moreover, overexpression of CLEC5A in M2 macrophages can also enhance its own M2 polarization (**Figure 6E**), as in a previous study (Tong et al., 2020). The correlation analysis between CLEC5A and immune checkpoint gene suggested that CLEC5A was positively correlated with CD80, CD86, PD-L1, CTLA4, CD47, and PD-L2 (**Figure 6C**). RT-PCR analysis showed that the expression of immune checkpoints such as PD-L1, PD-L2, and CD47 in OC cells overexpressing CLEC5A was also increased (**Figure 6F** and **Supplementary Figure 4D**), suggesting that OC patients with high expression of CLEC5A may have a better

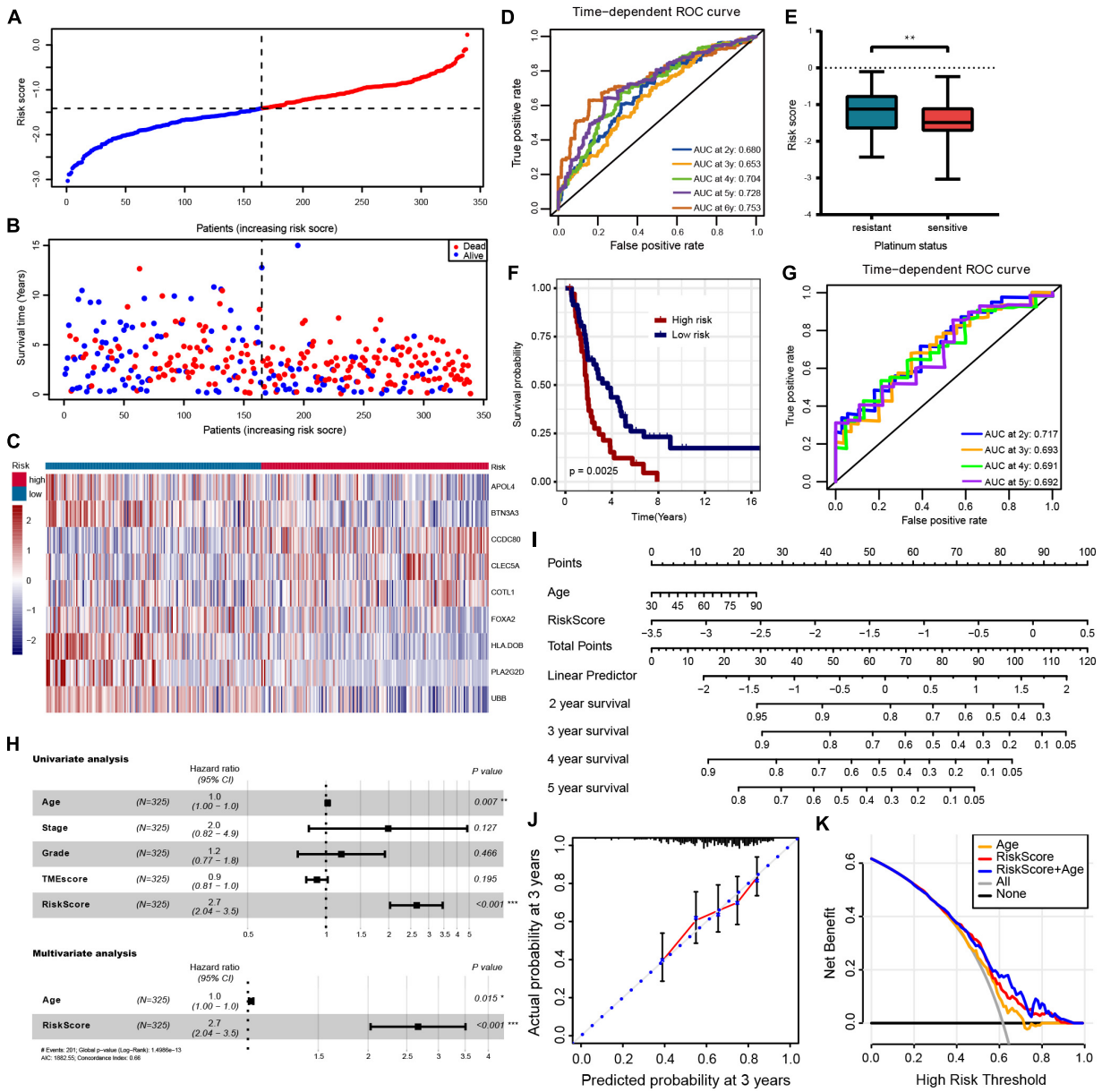


FIGURE 4 | (A–C) The risk score distribution, patients' survival status and gene expression profiles are displayed; the black-dotted line divided patients into high- and low-risk groups. **(D)** Time-dependent receiver operating characteristic (ROC) curve analysis of survival prediction by the prognostic model. **(E)** The risk score of platinum-resistant patients with OC is significantly higher than that of platinum-sensitive patients. **(F,G)** OS K–M curves and time-dependent ROC curves for the prognostic model in the ICGC OV-AU cohort. **(H)** Univariate and multivariate associations of the prognostic model and clinical characteristics with OS. **(I)** The nomogram is applied by adding up the points identified on the point scale for each variable. The total points projected on the bottom scales indicate the probability of 2-, 3- 4-, and 5-year OS. **(J)** The calibration curve for predicting 3-year OS for patients with OC; the *Y*-axis represents actual survival, as measured by K–M analysis, and the *X*-axis represents the prognostic model-predicted survival. **(K)** The net benefit (*Y*-axis) of age, risk model, and the nomogram; the black solid line represents the assumption that all patients survived, the gray solid line represents the assumption that no patients survived, and the orange, red, or blue solid lines represent the age, risk model, and nomograms. ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

therapeutic effect on immune checkpoint inhibitors. Mutation analysis showed that the tumor mutational burden (TMB) was significantly increased in OC patients with high expression of CLEC5A (**Figure 7C**). Moreover, the mutations of TTN, CDK12, TAF1L, and DNMT1 genes in these patients were significantly higher than those in patients with low expression of CLEC5A, and the co-mutations occurred frequently (**Figures 7A,B**). Through CNV analysis, we further found that the copy number of CTHRC1, DCAF13, RIMS2, and SLC25A32 was significantly increased in patients with high CLEC5A expression (**Figure 7D**). Most of these genes were proved to be closely related to the occurrence and development of tumors in previous studies

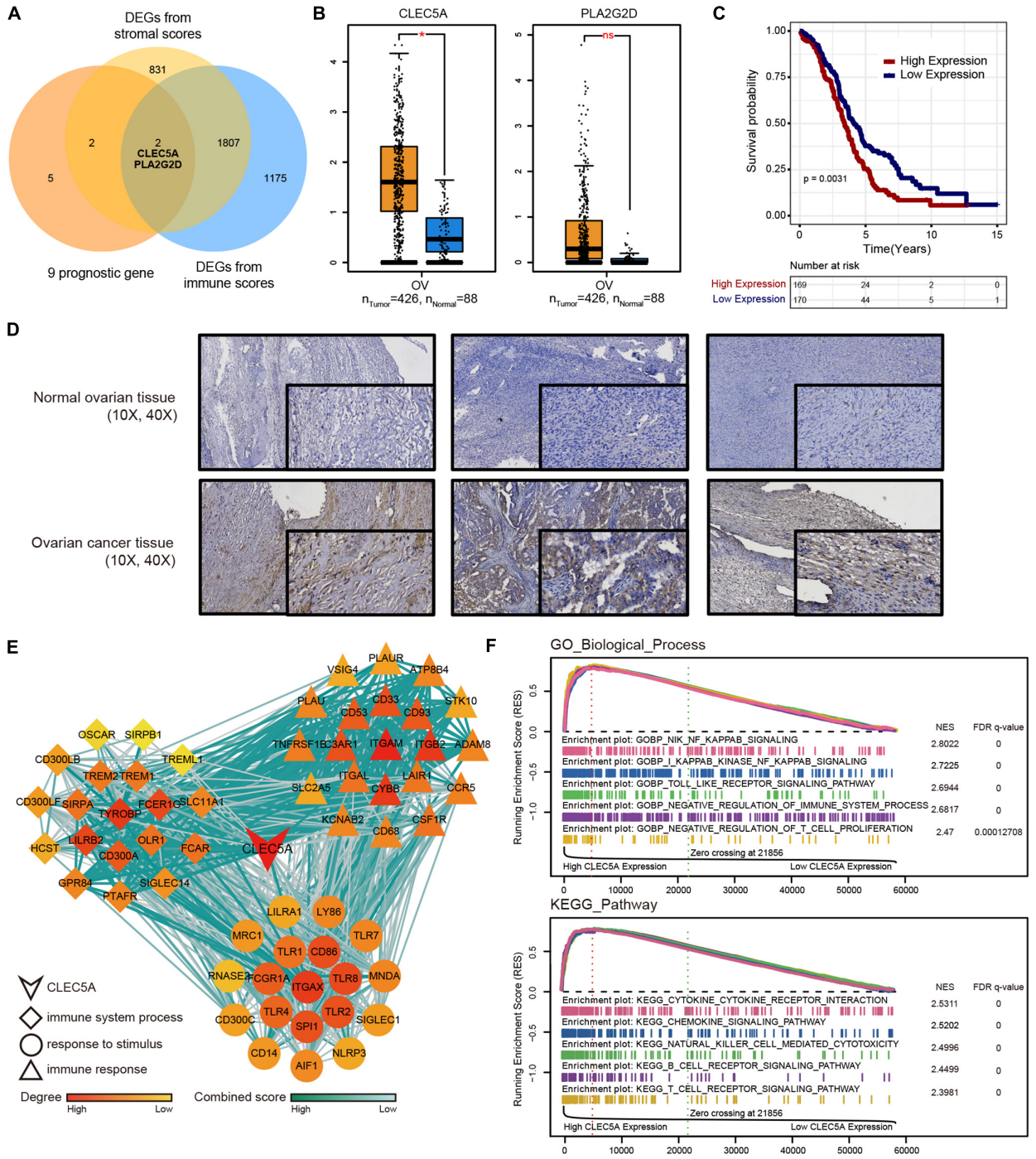


FIGURE 5 | (A) Venn diagram of DEGs from TMEscore and immune-stromal score. **(B)** Expression difference of CLEC5A and PLA2G2D in normal ovarian tissue and OC. **(C)** OS Kaplan–Meier curves of CLEC5A. **(D)** Expression of CLEC5A in normal ovarian tissue and OC (immunohistochemical staining). **(E)** The functions of three groups of CLEC5A-related genes. **(F)** Single-gene GSEA analysis of CLEC5A (GO-BP and KEGG pathway).*, $p < 0.05$; ns, no significance.

(Wang K. et al., 2019; Ding et al., 2020; Hill et al., 2020; Liu et al., 2020; Mei et al., 2020; Santoro et al., 2020). In the high CLEC5A group, the low copy number genes were significantly related to NK cell killing, Fc-gamma signaling pathway, and immune response (**Figure 7E**).

DISCUSSION

Ovarian cancer is common in women and is considered one of the most lethal malignancies. The high mortality of OC is usually due to the failure of early diagnosis. Although the current treatment

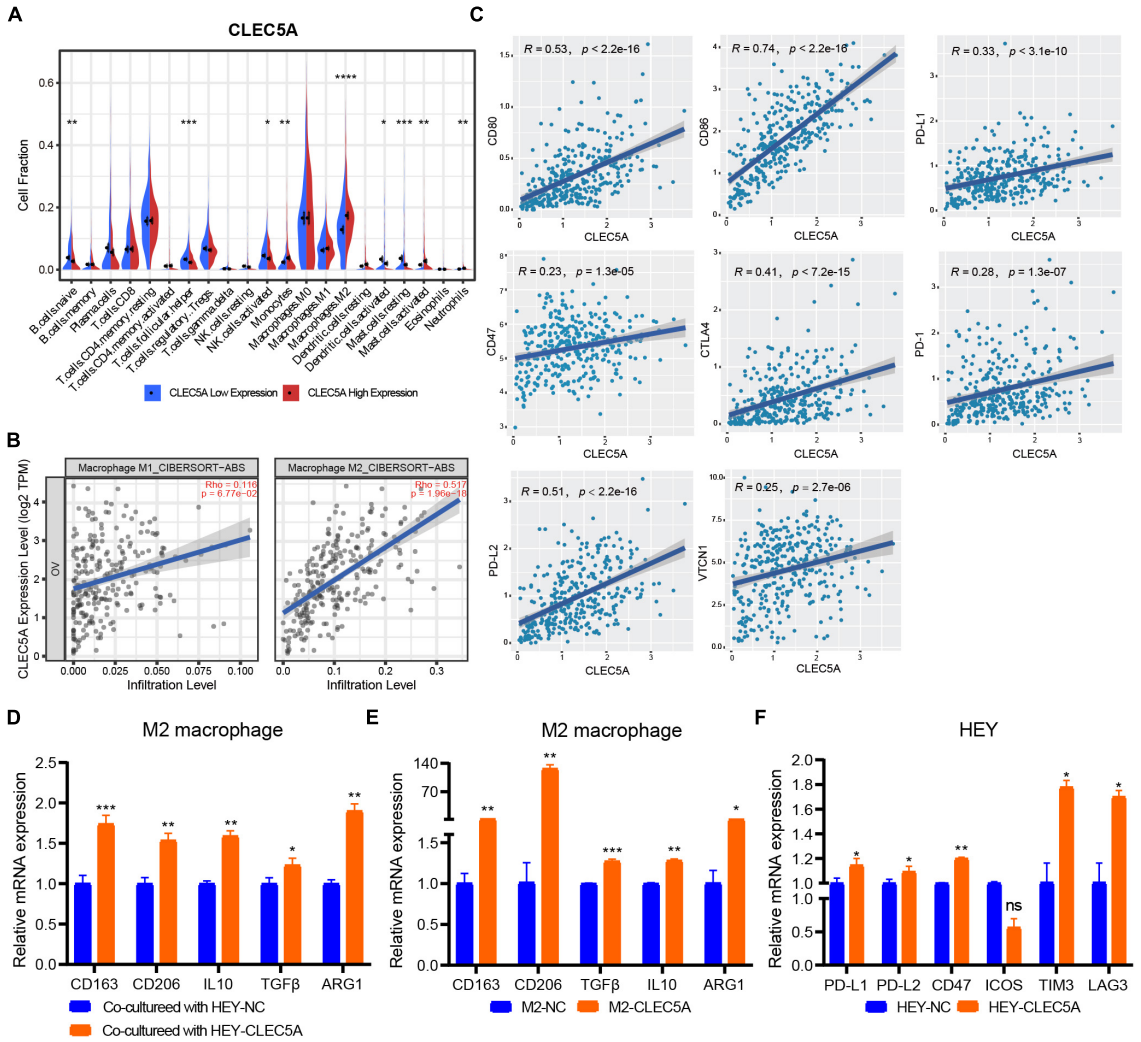


FIGURE 6 | (A) Correlation analysis of CLEC5A expression and immune cell components in the TME of OC (TCGA-OV dataset). **(B)** Correlation analysis between the expression of CLEC5A and the infiltration of M1 and M2 macrophages in TME of OC. **(C)** Correlation analysis of CLEC5A expression and immune checkpoint gene. **(D)** When the M2 macrophages were co-culture with OC cell line HEY overexpressing CLEC5A, the M2 polarization level of macrophages increased. **(E)** After overexpression of CLEC5A in M2 macrophages, the M2 polarization level increased. **(F)** The expression of immune checkpoint genes PD-L1, PD-L1, CD47, TIM3, and LAG3 increased in OC cell line HEY overexpressing CLEC5A. ****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; ns, no significance.

plan is reasonably effective in the early stages, most patients ultimately recur and eventually develop chemoresistance. An increasing number of studies have shown that immune cell infiltration in the TME plays a vital role in regulating the occurrence and therapeutic efficacy of OC (Au et al., 2017; Curtis et al., 2018; Singel et al., 2019). In recent years, several immunoscore-based models have been reported to quantify the immune environment in tumor tissue and provide a statistically powerful indicator for the prognosis of patients with various solid tumors (Galon et al., 2012, 2014; Jiang et al., 2018). Earlier studies by Zeng et al. (2019) suggested that the TMEscore had good value in predicting the survival of patients with gastric cancer and guiding more effective immunotherapy strategies. Hence, we assessed the TME in OC patients using the TMEscore in this study. Our findings suggested that the assessment of OC immune status by the TMEscore provided a predictor of survival in OC patients. The TME-relevant genes highly expressed in the high TMEscore group significantly related to immune cell activation and proliferation, response to IFN-γ, and cellular killer.

Previous studies have shown that M1 tumor-associated macrophages (TAMs), $CD8^+$ T cells, and $CD4^+$ T cells played a crucial role in controlling tumor metastasis and prognosis (Borst et al., 2018; Yang et al., 2018; van der Leun et al., 2020). The present study supported the previous views that our results showed a significant increase in the proportion of M1 macrophages, $CD8^+$ T cells, and $CD4^+$ T cells in the high TMEscore group patients with OC, and the survival rates of these patients increased significantly. Besides, the chemokines secreted by the M1 macrophages and the cytolytic molecules associated with cell killing were highly expressed in the high TMEscore

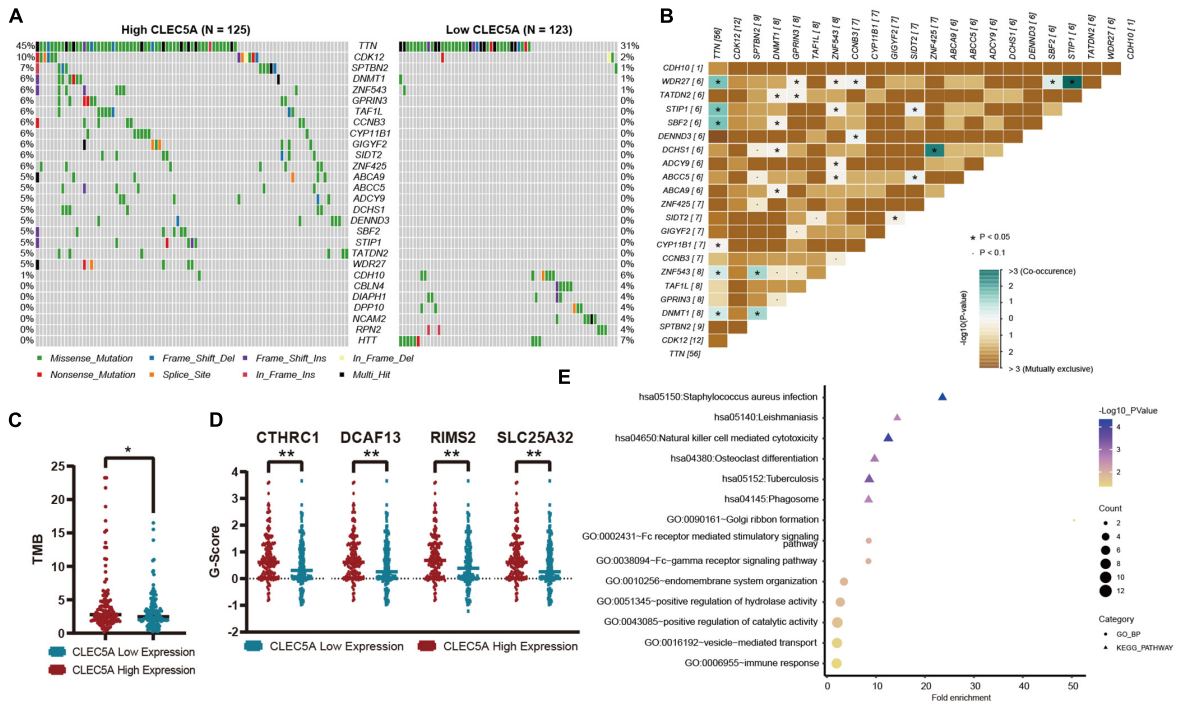


FIGURE 7 | (A) Mutation genes with significant difference between high- and low-expression CLEC5A groups. **(B)** Analysis of co-mutation and exclusive mutations in the CLEC5A overexpression group. **(C)** The difference of tumor mutational burden (TMB) levels in high- and low-expression CLEC5A groups. **(D)** Genes with significantly increased copy number variation in CLEC5A overexpression group. **(E)** Functional analysis of genes with decreased copy number variation in CLEC5A overexpression group. $^{**}p < 0.01$; $^{*}p < 0.05$.

group. We further analyzed the PPI networks and found a close relationship between these genes involved in antitumor immunity. Several ncRNAs (SNORA64, ATP6V0CP3, SNHG11, PPARGC1A) and TFs (*NFKB1*, *IRF1*, and *STAT1*) regulating prognostic genes found in pivot analysis were associated with OC (Permuth-Wey et al., 2011; Huo et al., 2013; Pavan et al., 2013; Dong et al., 2014; Au et al., 2016; Zhao and Fan, 2019). Taken together, these results demonstrated that the TMEscore could reflect the immune microenvironment of the tumor and the prognosis of patients with OC.

Although a single biomarker for predicting prognosis has been widely reported, a single gene cannot reflect the complexity of TME well. It is necessary to establish a multi-gene prognostic risk model (Long et al., 2018; Wang R. et al., 2019). In our study, we combined the TMEscore reflecting the prognosis with LASSO for the first time to develop a risk model that reflected both TME characteristics and OS in OC patients. We identified nine genes to construct the risk model which had a good performance for OS prediction according to the survival and ROC analysis. Besides, the significant consistency between the risk score groups and the TMEscore groups (χ^2 contingency tests, $p = 0.001$) demonstrates that the risk model can reflect the TME signature of patients with OC to some extent. The relationship between risk score and platinum response indicated that this risk model can be used to predict the sensitivity to platinum drugs in patients with OC and guide the clinical treatment. In addition, we confirmed that the risk score was independent of other relative factors such as pathological stage, histological grade, and TMEscore in patients with OC. Then, the nomogram based on age and risk score was constructed to precisely predict the survival probability in OC patients. The calibration curves and DCA curve showed that the actual survival was closely related to predicted survival, indicating that the nomogram could predict the OS of patients with OC well. Therefore, clinicians can offer individualized treatment for OC patients according to their nomogram predictions.

To select the genes that are most closely related to tumor immunity from these key prognostic genes, we further combined with the DEGs based on ESTIMATE analysis to screen CLEC5A as the key prognostic gene. Previous studies on CLEC5A are mostly related to the immune response caused by microbial viruses, and most of the studies related to tumors focus on glioblastoma, but the research in OC has been a vacancy (Fan et al., 2019; Sung et al., 2019; Tang et al., 2019). Our study found that compared with normal ovarian tissue, the expression of CLEC5A in OC increased significantly, and the higher the expression of CLEC5A, the lower the survival rate of patients. Previous studies have shown that CLEC5A expression in glioma patients was significantly correlated with immunosuppression and survival rate and could be used as a marker of M2 macrophages (Tong et al., 2020). The same pattern was also found in OC that M2 macrophage infiltration abundance was significantly increased in OC patients with high expression of CLEC5A. When OC cells overexpressing CLEC5A were co-cultured with M2 macrophages, the polarization

of M2 macrophages increased. Moreover, M2 macrophages overexpressing CLEC5A also showed enhanced M2 polarization. The expression of immune checkpoint genes in OC patients was also closely related to CLEC5A, which indicates that CLEC5A may affect the response of OC patients to immune drugs. The functional analysis of CLEC5A showed that it was closely related to various immune cell pathways and the NF-κB pathway, which provides some clues for the follow-up mechanism research. The TMB of patients with high expression of CLEC5A was increased, and there were high-frequency mutations of TTN, CDK12, TAF1L, and DNMT1, which have been proved to be closely related to the occurrence and development of tumors in previous studies (Tien et al., 2017; Zhong et al., 2019; Yang et al., 2020). In addition, in patients with high expression of CLEC5A, sites with reduced genome copy number significantly affected the immune response pathway. These results suggest that CLEC5A could be used as a potential prognostic marker for OC patients.

In conclusion, TMEscore is an indicator of OC prognosis, and patients with a high TMEscore often have a good prognosis. The risk model based on the prognostic genes analyzed by TME signature is a dependable indicator for predicting the OS of OC patients, and the nomogram containing prognostic models can comprehensively evaluate the prognosis of patients. In addition, we confirmed for the first time that CLEC5A was a potential factor affecting the immune microenvironment of OC, and more research is needed to explore its importance in microenvironment remodeling, to further improve the effect of OC immunotherapy.

DATA AVAILABILITY STATEMENT

The RNA-seq expression profile data and clinical information of OC patients were downloaded from the TCGA database (https://tcga-data.nci.nih.gov/tcga/).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Shanghai First Maternity and Infant Hospital (the ethical certification number: KS1748). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JS and TL collected and assembled the data, performed the data analysis, and wrote the manuscript. JL performed the immunohistochemistry. SX provided the conception and helped with the manuscript and data review finally. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 81772762), Clinical Science and Technology Innovation Project of Shanghai Shenkang Hospital Development Center (Grant No. SHDC12019X34), Natural Science Foundation of Shanghai (Grant No. 21ZR1450900), Shanghai Science and Technology Planning Project (Grant No. 21Y11907000), and the National Key R&D Program of China (Grant No. 2017YFA0104603).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.746932/full#supplementary-material

REFERENCES

Au, K. K., Le Page, C., Ren, R., Meunier, L., Clement, I., Tyrishkin, K., et al. (2016). STAT1-associated intratumoural TH1 immunity predicts chemotherapy resistance in high-grade serous ovarian cancer. *J. Pathol. Clin. Res.* 2, 259–270. doi: 10.1002/cjp2.55

Au, K. K., Peterson, N., Truesdell, P., Reid-Schachter, G., Khalaj, K., Ren, R., et al. (2017). CXCL10 alters the tumour immune microenvironment and disease progression in a syngeneic murine model of high-grade serous ovarian cancer. *Gynecol. Oncol.* 145, 436–445. doi: 10.1016/j.ygyno.2017.03.007

Borst, J., Ahrends, T., Babala, N., Melief, C. J. M., and Kastenmuller, W. (2018). CD4(+) T cell help in cancer immunology and immunotherapy. *Nat. Rev. Immunol.* 18, 635–647. doi: 10.1038/s41577-018-0044-0

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Curtis, M., Mukherjee, A., and Lengyel, E. (2018). The tumor microenvironment takes center stage in ovarian cancer metastasis. *Trends Cancer* 4, 517–519. doi: 10.1016/j.trecan.2018.06.002

Ding, X., Huang, R., Zhong, Y., Cui, N., Wang, Y., Weng, J., et al. (2020). CTHRC1 promotes gastric cancer metastasis via HIF-1alpha/CXCR4 signaling pathway. *Biomed. Pharmacother.* 123:109742. doi: 10.1016/j.biopha.2019.109742

Dong, X., Men, X., Zhang, W., and Lei, P. (2014). Advances in tumor markers of ovarian cancer for early diagnosis. *Indian J. Cancer* 51(Suppl. 3), e72–e76. doi: 10.4103/0019-509X.154049

Fan, H. W., Ni, Q., Fan, Y. N., Ma, Z. X., and Li, Y. B. (2019). C-type lectin domain family 5, member A (CLEC5A, MDL-1) promotes brain glioblastoma tumorigenesis by regulating PI3K/Akt signalling. *Cell Prolif.* 52:e12584. doi: 10.1111/cpr.12584

Fridman, W. H., Zitvogel, L., Sautes-Fridman, C., and Kroemer, G. (2017). The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol.* 14, 717–734. doi: 10.1038/nrclinonc.2017.101

Galon, J., Mlecnik, B., Bindea, G., Angell, H. K., Berger, A., Lagorce, C., et al. (2014). Towards the introduction of the "Immunoscore" in the classification of malignant tumours. *J. Pathol.* 232, 199–209. doi: 10.1002/path.4287

Galon, J., Pages, F., Marincola, F. M., Thurin, M., Trinchieri, G., Fox, B. A., et al. (2012). The immune score as a new possible approach for the classification of cancer. *J. Transl. Med.* 10:1. doi: 10.1186/1479-5876-10-1

Hill, H. A., Qi, X., Jain, P., Nomie, K., Wang, Y., Zhou, S., et al. (2020). Genetic mutations and features of mantle cell lymphoma: a systematic review and meta-analysis. *Blood Adv.* 4, 2927–2938. doi: 10.1182/bloodadvances.2019001350

Huo, Z. H., Zhong, H. J., Zhu, Y. S., Xing, B., and Tang, H. (2013). Roles of functional NFKB1 and beta-TrCP insertion/deletion polymorphisms in mRNA expression and epithelial ovarian cancer susceptibility. *Genet. Mol. Res.* 12, 3435–3443. doi: 10.4238/2013.March.11.6

Ignacio, R. M. C., Lee, E. S., Wilson, A. J., Beeghly-Fadiel, A., Whalen, M. M., and Son, D. S. (2018). Chemokine network and overall survival in TP53 wild-type and mutant ovarian cancer. *Immune Netw.* 18:e29. doi: 10.4110/in.2018.18.e29

Jiang, Y., Zhang, Q., Hu, Y., Li, T., Yu, J., Zhao, L., et al. (2018). ImmunoScore signature: a prognostic and predictive tool in gastric cancer. *Ann. Surg.* 267, 504–513. doi: 10.1097/SLA.0000000000002116

Kandalaft, L. E., Powell, D. J. Jr., Singh, N., and Coukos, G. (2011). Immunotherapy for ovarian cancer: what's next? *J. Clin. Oncol.* 29, 925–933.

Leffers, N., Gooden, M. J., Mokhova, A. A., Kast, W. M., Boezen, H. M., Ten Hoor, K. A., et al. (2009). Down-regulation of proteasomal subunit MB1 is an independent predictor of improved survival in ovarian cancer. *Gynecol. Oncol.* 113, 256–263. doi: 10.1016/j.ygyno.2008.12.030

Liu, J., Li, H., Mao, A., Lu, J., Liu, W., Qie, J., et al. (2020). DCAF13 promotes triple-negative breast cancer metastasis by mediating DTX3 mRNA degradation. *Cell Cycle* 19, 3622–3631. doi: 10.1080/15384101.2020.1859196

Long, J., Zhang, L., Wan, X., Lin, J., Bai, Y., Xu, W., et al. (2018). A four-gene-based prognostic model predicts overall survival in patients with hepatocellular carcinoma. *J. Cell. Mol. Med.* 22, 5928–5938. doi: 10.1111/jcmm.13863

Mantovani, A., Marchesi, F., Malesci, A., Laghi, L., and Allavena, P. (2017). Tumour-associated macrophages as treatment targets in oncology. *Nat. Rev. Clin. Oncol.* 14, 399–416. doi: 10.1038/nrclinonc.2016.217

Mariathasan, S., Turley, S. J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., et al. (2018). TGFbeta attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* 554, 544–548. doi: 10.1038/nature25501

Mei, D., Zhu, Y., Zhang, L., and Wei, W. (2020). The role of CTHRC1 in regulation of multiple signaling and tumor progression and metastasis. *Mediators Inflamm.* 2020:9578701. doi: 10.1155/2020/9578701

Melaiu, O., Lucarini, V., Cifaldi, L., and Fruci, D. (2019). Influence of the tumor microenvironment on NK cell function in solid tumors. *Front. Immunol.* 10:3038. doi: 10.3389/fimmu.2019.03038

Murray, P. J., Allen, J. E., Biswas, S. K., Fisher, E. A., Gilroy, D. W., Goerdt, S., et al. (2014). Macrophage activation and polarization: nomenclature and experimental guidelines. *Immunity* 41, 14–20.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Nishino, M., Ramaiya, N. H., Hatabu, H., and Hodi, F. S. (2017). Monitoring immune-checkpoint blockade: response evaluation and biomarker development. *Nat. Rev. Clin. Oncol.* 14, 655–668. doi: 10.1038/nrclinonc.2017.88

Nowak, M., Glowacka, E., Lewkowicz, P., Banasik, M., Szyllo, K., Zimna, K., et al. (2018). Sub-optimal primary surgery leads to unfavorable immunological changes in ovarian cancer patients. *Immunobiology* 223, 1–7. doi: 10.1016/j.imbio.2017.10.021

Pavan, S., Olivero, M., Cora, D., and Di Renzo, M. F. (2013). IRF-1 expression is induced by cisplatin in ovarian cancer cells and limits drug effectiveness. *Eur. J. Cancer* 49, 964–973. doi: 10.1016/j.ejca.2012.09.024

Permuth-Wey, J., Chen, Y. A., Tsai, Y. Y., Chen, Z., Qu, X., Lancaster, J. M., et al. (2011). Inherited variants in mitochondrial biogenesis genes may influence epithelial ovarian cancer risk. *Cancer Epidemiol. Biomarkers Prev.* 20, 1131–1145. doi: 10.1158/1055-9965.EPI-10-1224

Rosenberg, J. E., Hoffman-Censits, J., Powles, T., van der Heijden, M. S., Balar, A. V., Necchi, A., et al. (2016). Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *Lancet* 387, 1909–1920. doi: 10.1016/S0140-6736(16)00561-4

Santoro, V., Kovalenko, I., Vriens, K., Christen, S., Bernthaler, A., Haegebarth, A., et al. (2020). SLC25A32 sustains cancer cell proliferation by regulating flavin adenine nucleotide (FAD) metabolism. *Oncotarget* 11, 801–812.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1–13. doi: 10.18637/jss.v039.i05

Singel, K. L., Emmons, T. R., Khan, A. N. H., Mayor, P. C., Shen, S., Wong, J. T., et al. (2019). Mature neutrophils suppress T cell immunity in ovarian cancer microenvironment. *JCI Insight* 4:e122311. doi: 10.1172/jci.insight.122311

Sung, P. S., Huang, T. F., and Hsieh, S. L. (2019). Extracellular vesicles from CLEC2-activated platelets enhance dengue virus-induced lethality via CLEC5A/TLR2. *Nat. Commun.* 10:2402. doi: 10.1038/s41467-019-10360-4

Tang, X., Xu, P., Wang, B., Luo, J., Fu, R., Huang, K., et al. (2019). Identification of a specific gene module for predicting prognosis in glioblastoma patients. *Front. Oncol.* 9:812. doi: 10.3389/fonc.2019.00812

Tien, J. F., Mazloomian, A., Cheng, S. G., Hughes, C. S., Chow, C. C. T., Canapi, L. T., et al. (2017). CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. *Nucleic Acids Res.* 45, 6698–6716. doi: 10.1093/nar/gkx187

Tong, L., Li, J., Choi, J., Pant, A., Xia, Y., Jackson, C., et al. (2020). CLEC5A expressed on myeloid cells as a M2 biomarker relates to immunosuppression and decreased survival in patients with glioma. *Cancer Gene Ther.* 27, 669–679. doi: 10.1038/s41417-019-0140-8

van der Leun, A. M., Thommen, D. S., and Schumacher, T. N. (2020). CD8(+) T cell states in human cancer: insights from single-cell analysis. *Nat. Rev. Cancer* 20, 218–232. doi: 10.1038/s41568-019-0235-4

Vitale, I., Manic, G., Coussens, L. M., Kroemer, G., and Galluzzi, L. (2019). Macrophages and metabolism in the tumor microenvironment. *Cell Metab.* 30, 36–50. doi: 10.1016/j.cmet.2019.06.001

Wang, K., Li, L., Fu, L., Yuan, Y., Dai, H., Zhu, T., et al. (2019). Integrated bioinformatics analysis the function of RNA binding proteins (RBPs) and their prognostic value in breast cancer. *Front. Pharmacol.* 10:140. doi: 10.3389/fphar.2019.00140

Wang, R., Ye, X. H., Zhao, X. L., Liu, J. L., and Zhang, C. Y. (2019). Development of a five-gene signature as a novel prognostic marker in ovarian cancer. *Neoplasma* 66, 343–349. doi: 10.4149/neo_2018_180705N447

Wu, D., Huang, Y., Kang, J., Li, K., Bi, X., Zhang, T., et al. (2015). ncRDeathDB: a comprehensive bioinformatics resource for deciphering network organization of the ncRNA-mediated cell death system. *Autophagy* 11, 1917–1926. doi: 10.1080/15548627.2015.1089375

Xiao, Z., Dai, Z., and Locasale, J. W. (2019). Metabolic landscape of the tumor microenvironment at single cell resolution. *Nat. Commun.* 10:3763. doi: 10.1038/s41467-019-11738-0

Yang, L., Wang, S., Zhang, Q., Pan, Y., Lv, Y., Chen, X., et al. (2018). Clinical significance of the immune microenvironment in ovarian cancer patients. *Mol. Omics* 14, 341–351. doi: 10.1039/c8mo00128f

Yang, Y., Zhang, J., Chen, Y., Xu, R., Zhao, Q., and Guo, W. (2020). MUC4, MUC16, and TTN genes mutation correlated with prognosis, and predicted tumor mutation burden and immunotherapy efficacy in gastric cancer and pan-cancer. *Clin. Transl. Med.* 10:e155. doi: 10.1002/ctm2.155

Zeng, D., Li, M., Zhou, R., Zhang, J., Sun, H., Shi, M., et al. (2019). Tumor microenvironment characterization in gastric cancer identifies prognostic and immunotherapeutically relevant gene signatures. *Cancer Immunol. Res.* 7, 737–750. doi: 10.1158/2326-6066.CIR-18-0436

Zeng, D., Zhou, R., Yu, Y., Luo, Y., Zhang, J., Sun, H., et al. (2018). Gene expression profiles for a prognostic immunoscore in gastric cancer. *Br. J. Surg.* 105, 1338–1348. doi: 10.1002/bjs.10871

Zhao, Q., and Fan, C. (2019). A novel risk score system for assessment of ovarian cancer based on co-expression network analysis and expression level of five lncRNAs. *BMC Med. Genet.* 20:103. doi: 10.1186/s12881-019-0832-9

Zhong, S., Yan, H., Chen, Z., Li, Y., Shen, Y., Wang, Y., et al. (2019). Overexpression of TAF1L promotes cell proliferation, migration and invasion in esophageal squamous cell carcinoma. *J. Cancer* 10, 979–989. doi: 10.7150/jca.26504

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Shen, Liu, Lv and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 01 November 2021
doi: 10.3389/fcell.2021.739715

Prediction of Gastric Cancer-Related Proteins Based on Graph Fusion Method

Hao Zhang†, Ruisi Xu†, Meng Ding and Ying Zhang**

Endoscopy Center, China-Japan Union Hospital of Jilin University, Changchun, China

Gastric cancer is a common malignant tumor of the digestive system with no specific symptoms. Due to the limited knowledge of pathogenesis, patients are usually diagnosed in advanced stage and do not have effective treatment methods. Proteome has unique tissue and time specificity and can reflect the influence of external factors that has become a potential biomarker for early diagnosis. Therefore, discovering gastric cancer-related proteins could greatly help researchers design drugs and develop an early diagnosis kit. However, identifying gastric cancer-related proteins by biological experiments is time- and money-consuming. With the high speed increase of data, it has become a hot issue to mine the knowledge of proteomics data on a large scale through computational methods. Based on the hypothesis that the stronger the association between the two proteins, the more likely they are to be associated with the same disease, in this paper, we constructed both disease similarity network and protein interaction network. Then, Graph Convolutional Networks (GCN) was applied to extract topological features of these networks. Finally, Xgboost was used to identify the relationship between proteins and gastric cancer. Results of 10-cross validation experiments show high area under the curve (AUC) (0.85) and area under the precision recall (AUPR) curve (0.76) of our method, which proves the effectiveness of our method.

Keywords: gastric cancer, protein, proteomics data, graph convolutional network, Xgboost

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Ningyi Zhang,
Harbin Institute of Technology, China
Hong Ju,
Heilongjiang Vocational College of Biology Science and Technology, China

****Correspondence:***
Meng Ding
309478506@qq.com
Ying Zhang
515070789@qq.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *11 July 2021*
Accepted: *02 August 2021*
Published: *01 November 2021*

Citation:
Zhang H, Xu R, Ding M and Zhang Y (2021) Prediction of Gastric Cancer-Related Proteins Based on Graph Fusion Method. Front. Cell Dev. Biol. 9:739715. doi: 10.3389/fcell.2021.739715

INTRODUCTION

Gastric cancer is a worldwide disease with high incidence rate and mortality rate, especially in East Asia (Villanueva, 2011). According to the data from GLOBOCAN in 2018, there were 1,033,701 new cases of and 782,685 deaths from gastric cancer in the world (Bray et al., 2018). At present, the early diagnosis of gastric cancer is limited; patients are usually diagnosed in advanced stage. Therefore, early diagnosis is the key to improve the prognosis of patients, which is also the goal pursued by many researchers (Jin et al., 2015). Biomarkers refer to the substances that can reflect the physiological, biochemical, immune, genetic, and other molecular changes in the organism (Szász et al., 2016; Liang et al., 2019; Cheng et al., 2021). The levels of biomarkers in patients' samples (such as blood, plasma, saliva, and urine) can reflect the health or disease status of patients, as well as the response to anticancer treatment. Due to the strong heterogeneity of gastric cancer, the use of proteomics technology to find new specific biomarkers will greatly improve the sensitivity and accuracy of the diagnosis of patients (Gullo et al., 2018).

Although many researchers tend to reveal diseases pathogenic mechanism by genomics (Peng and Zhao, 2020; Zhao et al., 2020b; Zhou et al., 2020), changes in protein quality in diseases reflect the progression of the disease and are also the product of genes (Zhao et al., 2020c). Unlike those studies that research diseases through gene expression (Zhao et al., 2021b), protein quantification is more accurate and has the potential to become a biomarker. Researchers have used various protein separation techniques, such as two-dimensional gel electrophoresis (2-DE) (Gygi et al., 2000), two-dimensional fluorescence difference gel electrophoresis (2D-DIGE) (Tannu and Hemby, 2006), isobaric tags for relative and absolute quantitation (iTRAQ), hydrophilic interaction liquid chromatography (HILIC) screening of potential target proteins of new gastric cancer biomarkers, and then Western blotting and enzyme-linked immunosorbent assay or immunohistochemistry (IHC) methods are further validated, and biomarkers that play a key role in the occurrence of malignant tumors can be discovered.

Ryu et al. (2003) used the tumor proteomics technology of antibody microarrays to identify inflammatory protein markers of gastric cancer. They found that 14 proteins have different expressions between normal gastric mucosa and tumor gastric mucosa. The proteome can be regarded as the functional cell equivalent of the genome. Proteomics is useful in discovering biomarkers and improving the diagnostic efficiency of early gastric cancer and has obvious advantages. At present, the prognosis and treatment methods of gastric cancer are guided by genome. Surgical resection is still the most common strategy of gastric cancer, but due to the high risk of disease progression in stage II or III patients, it becomes important to increase adjuvant therapy. The strong heterogeneity of gastric cancer makes the therapeutic effect heterogeneous. Therefore, although the TNM system can help the prognosis of gastric cancer, many researchers tend to discover biomarkers to predict treatment outcomes more accurately (Pang et al., 2018). For example, Balluff et al. (2011) used matrix-assisted laser desorption/ionization (MALDI) imaging technology to analyze tissue samples and found that cysteine-rich intestinal protein 1 (CRIP1) and human neutrophil peptide-1 (HNP-1) were prognostic factors for gastric cancer. Human epidermal growth factor receptor 2 (HER2) is an important biomarker in gastric tumors, which can be specifically targeted for treatment with trastuzumab monoclonal antibody (mAb). For patients with advanced gastric cancer or gastroesophageal junction cancer, trastuzumab combined with chemotherapy can improve the survival rate of patients (Park et al., 2018).

There are still few proteins known to be related to gastric cancer. With the explosive growth of various types of omics data (Mo et al., 2020; Zhao et al., 2020a, 2021a), computational methods are widely used to identify disease-related biomolecules. Mining disease-related molecules based on the protein interaction networks has become a universal method. Sang et al. (2011) discovered genes and pathways of ciliopathy disease based on protein network. Seyfried et al. (2017) constructed protein network to identify protein-specific co-expression in Alzheimer's disease. With the development of Graph Convolutional Networks (GCN), an increasing number of researchers tend to use this method to process the complex topological features of the biological network. Its core point of view is to make the entire graph converge through the dissemination of node information and then make predictions on the basis of it. It has been widely used in prediction of biomolecular interaction (Tianyi et al., 2020). Therefore, we proposed a GCN-based method in this paper, named "GXGCP" (Gcn-Xgboost for Gastric Cancer-related Proteins identification) to identify gastric cancer-related proteins.

MATERIALS AND METHODS

There are four steps to implement GXGCP. Step 1 is to construct disease similarity network and protein interaction network. Step 2 is using GCN to extract topological features of disease similarity network and protein interaction network, respectively. Step 3 is to reduce the dimension of protein and gastric cancer features by principal component analysis (PCA). Step 4 is to identify gastric cancer-related proteins based on the features of protein and gastric cancer by Xgboost. The workflow of GXGCP is shown in **Figure 1**.

Construction of Network

We used SemFunsim (Cheng et al., 2014) to obtain diseases that are similar to gastric cancer. This method considers both disease semantic association and gene association. The detailed calculation process will not be repeated in this paper. A total of 327 diseases were found to be similar to gastric cancer. Based on the similarity, we constructed disease network, in which the edges are similarity and nodes are diseases. Therefore, the network has weight.

We downloaded protein interaction information from Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Mering et al., 2003). Based on the interaction, we constructed a protein network. If a protein can interact with the other one, there would be an edge to connect each other. Since the intensity of interaction between different proteins is different, this network also has weight.

Extracting Topological Features by Graph Convolutional Networks

To fully extract topological features of protein and disease network, GCN was applied (Han et al., 2019). The aim to implement GCN is to convert network topology into a vector output:

$$H^{(l+1)} = GCN(H^{(l)}, A) \tag{1}$$

where $H^{(0)}$ is node's feature in the network.

First, Laplace transform should be done on the network:

$$L = D - A \tag{2}$$

where D is the degree matrix of the network, and A is the adjacency matrix.

$$\hat{\mathrm{D}}_{\mathrm{ii}} = \sum_{j} \hat{\mathrm{A}}_{ij} \tag{3}$$

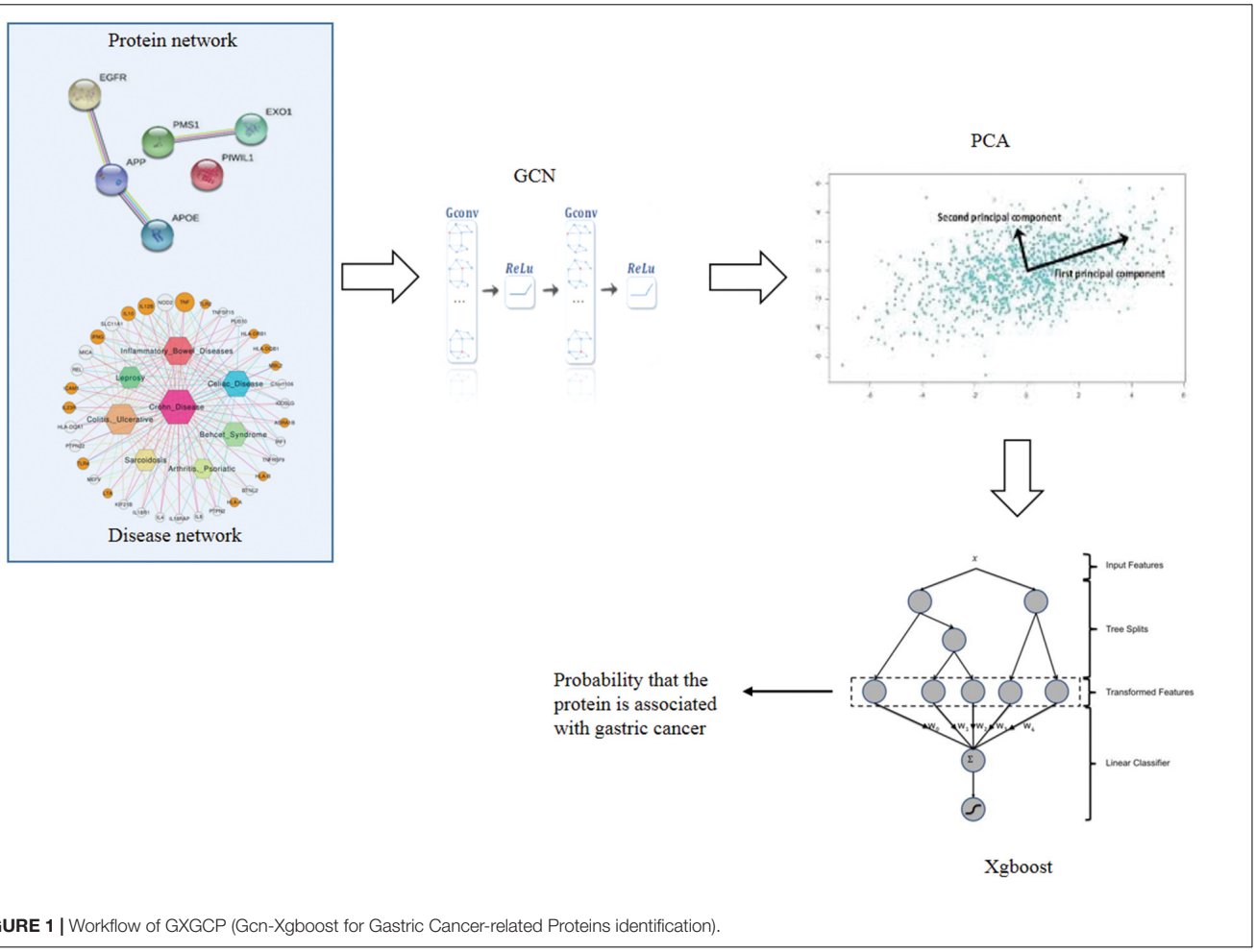


FIGURE 1 | Workflow of GXGCP (Gcn-Xgboost for Gastric Cancer-related Proteins identification).

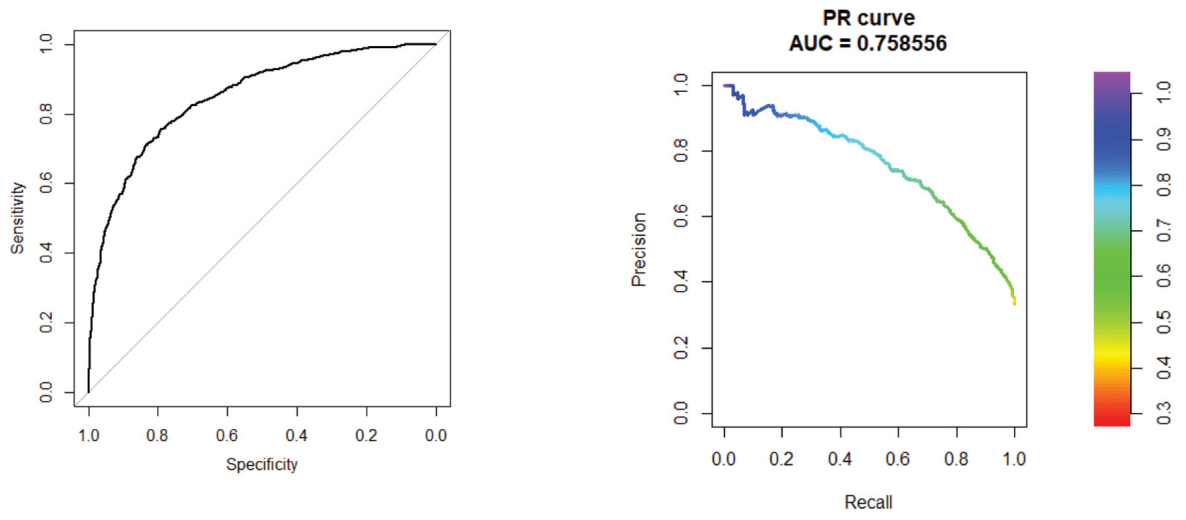


FIGURE 2 | Receiver operating characteristic (ROC) curve of 10-cross validation experiments. The AUC and AUPR curve of GXGCP are 0.85 and 0.76, respectively. The comparison results are listed in **Table 1**.

TABLE 1 | Comparison results.

Method	AUC	AUPR
GXGCP	0.85	0.76
RWXGCP	0.81	0.72
GXGCP without PCA	0.72	0.68
GSVMCP	0.74	0.65
GANNCP	0.76	0.71
GCNNCP	0.82	0.76
GDNNCP	0.80	0.74

Then, normalization should be implemented on the Laplacian matrix:

$$L^{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{4}$$

L^{sym} is defined as:

$$L_{i,j}^{sym} = \begin{cases} 1 & i = j \ and \ \deg(v_i) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} & i \neq j \ and \ v_i \ adjacent \ to \ v_j \\ 0 & otherwise \end{cases} \tag{5}$$

With the Laplacian matrix, we can perform spectral convolution on the network. We need to find a suitable convolution kernel so that $f()$ can reduce the loss of classification after the convolution transformation of the convolution kernel. The core of the machine learning task on the graph is to find a convolution kernel that can reduce the loss, regard $h(\lambda_1)$, ...$h(\lambda_n)$ as the parameters of the model, and apply the gradient descent method to update these parameters.

The final formula of GCN would be:

$$H^{(l+1)} = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(l)} W^{(l)}) \tag{6}$$

where $\sigma()$ is the activation function, and $W^{(l)}$ is the parameter to be trained.

Reduction Dimension by Principal Component Analysis

Since the dimension of metabolites and gastric cancer features are large, we used PCA to reduce the dimension. There are four steps to apply PCA (Tipping and Bishop, 1999; Cheng et al., 2019). The first step is feature centralization. That is, the data of each dimension are subtracted from the mean value of that dimension, and the mean value of each dimension becomes 0 after the transformation. The second step is to calculate covariance matrix. The third step is to calculate the eigenvalues and eigenvectors of the covariance matrix. The last step is to select the feature vector corresponding to the large feature value to obtain a new data set.

Classification of Gastric Cancer-Related Proteins by Xgboost

Xgboost is a sparse perception algorithm that can be used for parallel tree learning (Chen and Guestrin, 2016). Since the features of gastric cancer and proteins are sparse, Xgboost is very suitable for the classification.

Xgboost is a tree ensemble model. It sums the results of K (the number of trees) as the final predicted value.

$$\widehat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{7}$$

Assuming that a given sample set has n samples and m features, then

$$D = \{(x_i, y_i)\} \tag{8}$$

where x_i represents the i-th sample, y_i represents the i-th category label, and the space F of the regression tree (CART tree) is:

$$F = \{f(x) = w_q(x)\} \tag{9}$$

where q represents the structure of each tree, it maps the sample to the corresponding leaf node; T is the number of leaf nodes of the corresponding tree; f(x) corresponds to the structure q of the tree and the leaf node weight w. Therefore, the predicted value of Xgboost is the sum of the values of the leaf nodes corresponding to each tree.

Our goal is to learn these k trees, so we minimize the following objective function with regular terms:

$$L(\phi) = \sum_i l(\widehat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{10}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2$

EXPERIMENT RESULTS

We implemented 10-cross validation experiments to test the performance of GXGCP. We divided our data into 10 groups. We used nine of 10 groups' data to train the model and the data of the remaining one to test the model. After repeating this process 10 times, each group has been tested once. To show the accuracy of our model, we compared GXGCP with several other methods such as RWXGCP, GXGCP without PCA, GSVMCP, GANNCP, and GCNNCP. RWXGCP replaces the GCN part of GXGCP with random walk (RW). GSVMCP replaces the Xgboost part of GXGCP with support vector machine (SVM). GANNCP replaces the Xgboost part of GXGCP with artificial neural network (ANN). GCNNCP replaces the Xgboost part of GXGCP with convolutional neural network (CNN).

The area under the curve (AUC) and area under the precision recall (AUPR) curve of GXGCP are shown in **Figure 2**. The comparison results are listed in **Table 1**.

As shown in **Table 1**, GXGCP performed best among these five methods. These results show that GCN is more suitable for encoding network than RW, and Xgboost is more suitable for building model by sparse data than SVM and ANN.

CONCLUSION

Protein is the main executor of life activities. To decrypt the genome, you must first systematically understand the proteome. Identifying gastric cancer-related proteins can greatly help develop screening or testing tools for tumor detection, early

diagnosis or differential diagnosis, prognostic analysis, efficacy evaluation, etc. Due to the high cost of biological experiments, we proposed GXGCP that fuses GCN, Xgboost, and PCA to identify gastric cancer-related proteins. To verify the accuracy of our method, we did 10-cross validation experiments. The results show that the AUC of GXGCP reached 0.85 and AUPR reached 0.76. To show the superiority of GXGCP, we compared it with several other methods, and GXGCP performed best. Overall, we propose a novel, efficient, and accurate method for large-scale identification of gastric cancer-related proteins, which would greatly benefit the study of the pathogenic mechanism and clinical research of gastric cancer.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

HZ and RX wrote this manuscript and did experiments. MD and YZ provided important ideas. All authors read and approved the final manuscript.

REFERENCES

Balluff, B., Rauser, S., Meding, S., Elsner, M., Schöne, C., Feuchtinger, A., et al. (2011). MALDI imaging identifies prognostic seven-protein signature of novel tissue markers in intestinal-type gastric cancer. *Am. J. Pathol.* 179, 2720–2729. doi: 10.1016/j.ajpath.2011.08.032

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, (New York, NY: Association for Computing Machinery), 785–794.

Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., and Zhang, X. (2021). Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. *Brief. Bioinform.* 22, 1442–1450. doi: 10.1093/bib/bbab042

Cheng, L., Li, J., Ju, P., Peng, J., and Wang, Y. (2014). SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One* 9:e99415. doi: 10.1371/journal.pone.0099415

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational methods for identifying similar diseases. *Mol. Ther. Nucleic Acids* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Gullo, I., Carneiro, F., Oliveira, C., and Almeida, G. M. (2018). Heterogeneity in gastric cancer: from pure morphology to molecular classifications. *Pathobiology* 85, 50–63. doi: 10.1159/000473881

Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U.S.A.* 97, 9390–9395. doi: 10.1073/pnas.160270797

Han, P., Yang, P., Zhao, P., Shang, S., Liu, Y., Zhou, J., et al. (2019). "GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (New York, NY: Association for Computing Machinery), 705–713.

Jin, Z., Jiang, W., and Wang, L. (2015). Biomarkers for gastric cancer: Progression in early diagnosis and prognosis. *Oncol. Lett.* 9, 1502–1508. doi: 10.3892/ol.2015.2959

Liang, C., Changlu, Q., He, Z., Tongze, F., and Xue, Z. (2019). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48:7603. doi: 10.1093/nar/gkz843

Mering, C. V., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261. doi: 10.1093/nar/gkg034

Mo, F., Luo, Y., Fan, D. A., Zeng, H., Zhao, Y. N., Luo, M., et al. (2020). Integrated analysis of mRNA-seq and miRNA-seq to identify c-MYC, YAP1 and miR-3960 as major players in the anticancer effects of caffeic acid phenethyl ester in human small cell lung cancer cell line. *Curr. Gene Ther.* 20, 15–24. doi: 10.2174/1566523220666200523165159

Pang, L., Wang, J., Fan, Y., Xu, R., Bai, Y., and Bai, L. (2018). Correlations of TNM staging and lymph node metastasis of gastric cancer with MRI features and VEGF expression. *Cancer Biomark.* 23, 53–59. doi: 10.3233/cbm-181287

Park, J. S., Lee, N., Beom, S. H., Kim, H. S., Lee, C.-K., Rha, S. Y., et al. (2018). The prognostic value of volume-based parameters using 18 F-FDG PET/CT in gastric cancer according to HER2 status. *Gastric Cancer* 21, 213–224. doi: 10.1007/s10120-017-0739-0

Peng, J., and Zhao, T. (2020). Reduction in TOM1 expression exacerbates Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* 117, 3915–3916. doi: 10.1073/pnas.1917589117

Ryu, J. W., Kim, H. J., Lee, Y. S., Myong, N. H., Hwang, C. H., Lee, G. S., et al. (2003). The proteomics approach to find biomarkers in gastric cancer. *J. Korean Med. Sci.* 18, 505–509. doi: 10.3346/jkms.2003.18.4.505

Sang, L., Miller, J. J., Corbit, K. C., Giles, R. H., Brauer, M. J., Otto, E. A., et al. (2011). Mapping the NPHP-JBTS-MKS protein network reveals ciliopathy disease genes and pathways. *Cell* 145, 513–528. doi: 10.1016/j.cell.2011.04.019

Seyfried, N. T., Dammer, E. B., Swarup, V., Nandakumar, D., Duong, D. M., Yin, L., et al. (2017). A multi-network approach identifies protein-specific co-expression in asymptomatic and symptomatic Alzheimer's disease. *Cell Syst.* 4, 60–72.e4. doi: 10.1016/j.cels.2016.11.006

Szász, A. M., Lánczky, A., Nagy, Á, Förster, S., Hark, K., Green, J. E., et al. (2016). Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* 7, 49322–49333. doi: 10.18632/oncotarget.10337

Tannu, N. S., and Hemby, S. E. (2006). Two-dimensional fluorescence difference gel electrophoresis for comparative proteomics profiling. *Nat. Protoc.* 1, 1732–1742. doi: 10.1038/nprot.2006.256

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22, 2141–2150. doi: 10.1093/bib/bbaa044

Tipping, M. E., and Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural. Comput.* 11, 443–482. doi: 10.1162/089976699300016728

Villanueva, M. T. (2011). Combination therapy: update on gastric cancer in East Asia. *Nat. Rev. Clin. Oncol.* 8:690. doi: 10.1038/nrclinonc.2011.171

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep

learning approaches. *Brief. Bioinform.* 22:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2020c). Identifying protein biomarkers in blood for Alzheimer's disease. *Front. Cell Dev. Biol.* 8:472. doi: 10.3389/fcell.2020.00472

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021a). Prediction and collection of protein–metabolite interactions. *Brief. Bioinform.* bbab014. doi: 10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Zhou, M. J., Hu, Z. Q., Zhang, C. H., Wu, L. Q., Li, Z., and Liang, D. S. (2020). Gene therapy for hemophilia A: where we stand. *Curr. Gene Ther.* 20, 142–151. doi: 10.2174/1566523220666200806110849

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Zhang, Xu, Ding and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 02 November 2021
doi: 10.3389/fcell.2021.760079

Identification of Immune-Related Genes for Establishment of Prognostic Index in Hepatocellular Carcinoma

Yinfang Li[1†], *Ling Zou*[1†], *Xuejun Liu*[2], *Judong Luo*[3*] and *Hui Liu*[2*]

[1] Aliyun School of Big Data, Changzhou University, Changzhou, China, [2] School of Computer Science and Technology, Nanjing Tech University, Nanjing, China, [3] Department of Radiotherapy, The Affiliated Changzhou No. 2 People's Hospital of Nanjing Medical University, Changzhou, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Jingpu Zhang,
Henan University of Urban Construction, China
Lihong Peng,
Hunan University of Technology, China

***Correspondence:**
Judong Luo
judongluo@163.com
Hui Liu
hliu@njtech.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 17 August 2021
Accepted: 11 October 2021
Published: 02 November 2021

Citation:
Li Y, Zou L, Liu X, Luo J and Liu H (2021) Identification of Immune-Related Genes for Establishment of Prognostic Index in Hepatocellular Carcinoma. Front. Cell Dev. Biol. 9:760079. doi: 10.3389/fcell.2021.760079

Background: Immune checkpoint inhibitor (ICI) therapy has been proved to be a promising therapy to many types of solid tumors. However, effective biomarker for estimating the response to ICI therapy and prognosis of hepatocellular carcinoma (HCC) patients remains underexplored. The aim of this study is to build a novel immune-related prognostic index based on transcriptomic profiles.

Methods: Weighted gene co-expression network analysis (WGCNA) was conducted to identify immune-related hub genes that are differentially expressed in HCC cohorts. Next, univariate Cox regression analysis and least absolute shrinkage and selection operator (LASSO) analysis were used to detect hub genes associated to overall survival (OS). To validate the immune-related prognostic index, univariate and multivariate Cox regression analysis were performed. CIBERSORT and ESTIMATE were used to explore the tumor microenvironment and immune infiltration level.

Results: The differential expression analysis detected a total of 148 immune-related genes, among which 25 genes were identified to be markedly related to overall survival in HCC patients. LASSO analysis yielded 10 genes used to construct the immune-related gene prognostic index (IRGPI), by which a risk score is computed to estimate low vs. high risk indicating the response to ICI therapy and prognosis. Further analysis confirmed that this immune-related prognostic index is an effective indicator to immune infiltration level, response to ICI treatment and OS. The IRGPI low-risk patients had better overall survival (OS) than IRGPI high-risk patients on two independent cohorts. Moreover, we found that IRGPI high-risk group was correlated with high TP53 mutation rate, immune-suppressing tumor microenvironment, and these patients acquired less benefit from ICI therapy. In contrast, IRGPI-low risk group was associated with low TP53 and PIK3CA mutation rate, high infiltration of naive B cells and T cells, and these patients gained relatively more benefit from ICI therapy.

Keywords: hepatocellular carcinoma, immune related genes, prognostic biomarker, immune infiltration level, tumor microenvironment

1. INTRODUCTION

Liver cancer remains a global health challenge, with an estimated incidence of more than 1 million cases by 2025 (Llovet et al., 2021) around the world. Hepatocellular carcinoma (HCC) is the most common form of liver cancer and accounts for 90% cases, and its increasing mortality rate is receiving growing concern. Conventional treatment, such as surgery, radiotherapy, and chemotherapy, do not significantly prolong overall survival (OS) of HCC patients (Ghouri et al., 2017).

Immunotherapy is emerged as an effective therapy in the field of cancer treatment in recent years, and among them the most impressive is immune checkpoint blockade (Mellman et al., 2011). Its clinical advantage including but not limited to continuous anti-tumor immune response with relatively weak side effect, low recurrence rate, and even complete remission for some advanced cancers (Khalil et al., 2016). As a result, immune checkpoint blockade has been approved for first-line therapy of some cancers.

For HCC, ICI has shown strong anti-tumor activity in a portion of patients (Khalil et al., 2016; Hou et al., 2020). Especially, the combination of the anti-PDL1 antibody atezolizumab and the VEGF-neutralizing antibody bevacizumab has become the standard of care as a first-line therapy for HCC (Sangro et al., 2021). However, the fraction of HCC patients that benefit from immunotherapy remains very limited, while other immunotherapy such as adoptive T-cell transfer, vaccination, or virotherapy have not yet demonstrated consistent clinical activity. As the factors that influence ICI efficacy are multifaceted, such as the immune microenvironment (TME) and PD-L1 level, the establishment of gold-standard biomarker for immunotherapy benefit is challenging (Nishino et al., 2017).

In this paper, we resort to transcriptomic data and clinical outcome to establish an immune-related gene prognostic index (IRGPI) of HCC, by exploiting weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008). Among 655 differentially expressed genes related to immunity, 148 hub genes were identified by WGCNA. Next, we identified 25 genes significantly related to overall survival by univariate Cox regression analysis, and selected 10 genes by LASSO analysis to construct the immune-related prognostic index (IRGPI), a quantitative score indicative of low vs. high risk of prognosis. Further analysis confirmed that the IRGPI is closely associated to immune microenvironment, response to ICI treatment and OS. In particular, the low-risk patients had better overall survival (OS) than high-risk patients, for both TCGA and GEO cohorts. Moreover, high-risk group was correlated with high TP53 mutation rate, immune-suppressing tumor microenvironment, and these patients acquired less benefit from ICI therapy. In contrast, low-risk group was associated with low TP53 and PIK3CA mutation rate, high infiltration of naive B cells and T cells, and these patients gained relatively more benefit from ICI therapy. Finally, we also validated IRGPI outperformed routine biomarkers used to screen patients who can benefit from immunotherapy. The results suggested that IRGPI was a promising prognostic biomarker for patients.

2. MATERIALS AND METHODS

2.1. Data Resource and Preprocessing

The cohort for IRGPI establishment was obtained from TCGA (TCGA cohort), which includes 424 samples (374 cancer samples and 50 para-cancer samples). We utilized the UCSC Xena browser (Goldman et al., 2020) to download the RNA-seq data (FPKM normalized). The matched somatic mutation data and clinical outcome of the HCC patients were also obtained.

An independent cohort was obtained from the GSE14520 to verify the effectiveness of the IRGPI risk score. The validation cohort includes 221 HCC patients, and the transcriptomic data and clinical outcome were obtained from GEO database (Edgar et al., 2002). To further verify the predictive performance of the IRGPI on ICI therapy, we employed GSE140901 dataset, which has the gene expression profiles and clinical outcome of HCC patients who received anti-PD-L1 therapy.

The immune-related genes were downloaded from ImmPort database (Bhattacharya et al., 2014) and InnateDB database (Breuer et al., 2013). After removal of duplicate genes, the ImmPort database contained 1,811 immune genes and the InnateDB database contained 1,226 immune genes.

Associations between protein-coding genes (mRNA) and transcriptional factors (TFs), miRNA and lncRNA were downloaded from TRRUST v2 (Han et al., 2018), TargetScan (Lewis et al., 2005), miRDB (Wong and Wang, 2015), miRTarBase (Chou et al., 2018), and miRcode databases (Jeggari et al., 2012), respectively.

2.2. Differential Expression and Enrichment Analysis

The *limma* package (version 3.44.3) (Ritchie et al., 2015) was used to identify the differentially expressed mRNA. The log2 fold change$>$1 and FDR$<$0.05 were used as the criteria to screen differentially expressed genes (DEGs). Enrichment analysis on Gene Ontology (GO) and KEGG pathways were conducted using the *clusterProflier* R package (Yu et al., 2012). *Treemap* R package (Baehrecke et al., 2004) were chosen for visualization of significant functional annotations and pathways.

2.3. Identification of Hub Genes by WGCNA

Weighted gene co-expression network analysis (WGCNA) was performed to identify hub genes. First, the similarity matrix was constructed by using the expression data by calculating the Pearson correlation coefficient between gene pairs. Next, the similarity matrix was transformed into an adjacency matrix with a network type of signed and then transformed into a topological matrix with the topological overlap measure (TOM) describing the degree of association between genes. 1-TOM was used as the distance to cluster the genes, and then the dynamic pruning tree was built to identify the modules. Finally, we identified the modules by setting the merging threshold function at 0.25.

The modules with significantly different expression patterns between tumor and normal tissues were chosen for downstream analysis. We took the intersection between the modules and differentially expressed immune-related genes. To assess the potential biologic functions and involved pathways of

the differentially expressed immune-related genes, gene set enrichment analysis (GSEA) was conducted using *clusterProfiler* R package.

2.4. Construction of Immune-Related Gene Prognostic Index

To develop a prognostic index, univariate Cox regression analysis and least absolute shrinkage and selection operator (LASSO) were conducted to assess the associations between immune-related genes expressions and overall survival (OS). Cox and LASSO regression were carried out using *survival* R package (Therneau and Lumley, 2014) and *glmnet* R package (Hastie and Qian, 2014).

The IRGPI risk score was computed using each gene expression level multiplied by its linear regression coefficient obtained from the univariate Cox regression. According to the median risk score, patients were assigned to high-risk and low-risk groups. The Kaplan-Meier survival analysis was performed to compare overall survival between high-risk and low-risk groups. The receiver operating characteristic (ROC) curves were plotted by *survival* R package. To validate the independent prognostic value of IRGPI, univariate and multivariate Cox regression analysis was performed on IRGPI risk score and other clinicopathologic feature.

2.5. Construction of Regulatory Networks

First, we identified the differentially expressed long non-coding RNAs (DElnRNAs) and miRNAs (DEmiRNAs) by setting the standards that log2 fold change$>$1 and FDR$<$0.05. Next, multiple miRNA target databases, including TargetScan, miRTarBase, and miRDB, were used to seek DEmiRNAs targeting the hub immune-related DEGs. The miRcode database was used to find lncRNA-miRNA associations. The TRRUST database version 2 was used to search the TFs regulating the hub immune-related DEGs. Finally, Cytoscape was used to visualize the regulatory network containing the miRNA-mRNA, lncRNA-miRNA, and TFs-mRNA associations.

2.6. TME and ICI Efficacy Compared Between IRGPI Groups

To explore the immune microenvironment of HCC, the transcriptomic data were imported into CIBERSORT (Newman et al., 2015)to compute the proportion of 22 types of immune cells. We compared the relative proportions of these immune cell types between two IRGPI groups. The immune infiltrating cells and tumor purity were assessed by ESTIMATE tool (Yoshihara et al., 2013). Besides, we explored the correlation between canonical immune subtypes and IRGPI groups.

Somatic mutation landscape was built in two IRGPI groups by using the *Maftools* R package (Mayakonda et al., 2018). Correlation analysis was performed between IRGPI risk score and conventional immunotherapy biomarkers, including programmed death receptor ligand-1/2 (PD-L1/2), programmed cell death protein-1 (PD-1), cytotoxic T lymphocyte associated antigen-4 (CTLA-4), cytolytic activity (CYT), and tumor mutation burden (TMB).

For estimation of the potential clinical efficacy of ICI treatment in different IRGPI groups, we calculated tumor immune dysfunction and exclusion (TIDE) scores (Jiang et al., 2018) in different IRGPI groups. Tumor inflammation signature (TIS) is an 18-gene signature that reflects an ongoing adaptive Th1 and cytotoxic CD8+ T cell response and shows promising results in predicting response to anti-PD-1/PD-L1 agents (Ayers et al., 2017). IRGPI risk score was also used to predict the overall survival of patients received anti-PD-L1 agents. ROC curves and AUC value were used to estimate the performance of IRGPI risk score. Performance comparison was also performed between IRGPI risk score and TIDE and TIS scores.

2.7. Statistical Analysis

Independent *t*-test was performed to compare continuous variables between two groups. Categorical data were tested using the chi-square test. TIDE score between groups was compared by the WilCoxon test. Univariate survival analysis was performed by Kaplan-Meier survival analysis with the log-rank test. A two-sided p-value < 0.05 was considered significant.

3. RESULTS

3.1. WGCNA Analysis Identified Immune-Related Hub DEGs

The differential expression analysis on 374 tumors and 50 normal TCGA samples yielded a total of 6,209 differentially expressed genes (DEGs), in which 5,391 genes were up-regulated and 818 genes were down-regulated. By intersecting DEGs with the immune-related genes collected from ImmPort and InnateDB databases, 655 differentially expressed immune-related genes were obtained, in which 460 genes were up-regulated and 195 were down-regulated in tumor samples compared with normal samples.

To obtain the immune-related hub genes, WGCNA analysis was carried out. As shown in **Figures 1A,B**, the logarithm log(k) of the node with connectivity k was negatively correlated with the logarithm log[P(k)] of the probability of the node, and the correlation coefficient was $>$0.85. The optimal soft-thresholding power was set to 10. Eight modules were identified based on the average linkage hierarchical clustering and the optimal soft-thresholding power (**Figures 1C,D**). According to the Pearson correlation coefficient between a module and sample feature, we found green, yellow, and magenta modules were closely correlated with HCC tumors, the genes in these two modules were selected for further analysis.

Taking the intersection between the genes included in the two WGCNA-derived modules and the lists of differentially expressed immune-related genes, 148 immune-related hub genes (IRGs) were obtained, of which 122 genes were up-regulated and 26 were down-regulated.

3.2. Treemaps of Enriched Functional Annotations and Pathways

To uncover the role of the immune-related hub genes in the pathogenesis of HCC, GO, and KEGG enrichment analysis was carried out. The top 10 significantly enriched GO terms

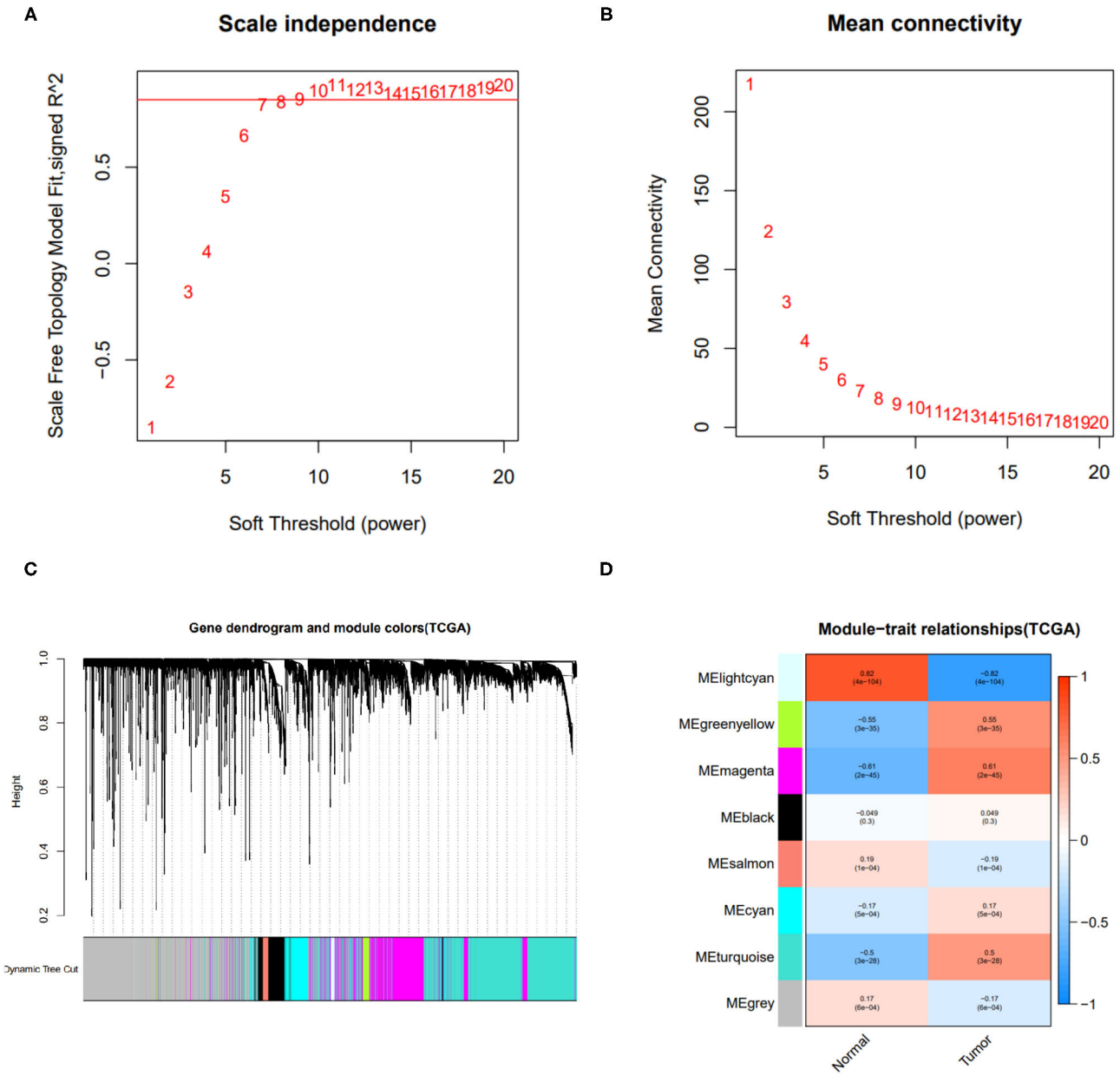


FIGURE 1 | Results of weighted gene co-expression network analysis (WGCNA). **(A)** The scale-free fit index of various soft-thresholding powers. **(B)** Mean connectivity of various soft-thresholding powers. **(C)** A dendrogram of the differentially expressed genes clustered based on different metrics. **(D)** Heatmap of associations between module eigengenes of normal and tumor tissues.

and KEGG pathways for IRGs are shown in **Figure 2**. As for cellular component, the set of IRGs are significantly enriched in the regulation of response to biotic stimulus, regulation of innate immune response, fc receptor signaling pathway, and regulation of morphogenesis of an epithelium. As for cellular component, they are significantly enriched in cytoplasmic vesicle lumen, secretory granule lumen, and peptidase complex. For molecular function, they are significantly enriched in receptor ligand activity signaling receptor activator activity and ubiquitin-like protein ligase binding. As for KEGG pathway analysis, they are significantly enriched in Epstein-Barr virus infection, pathways of neurodegeneration-multiple diseases and amyotrophic lateral sclerosis.

3.3. IRGPI Risk Score Is Predictive of Overall Survival

The screened 148 immune-related hub genes were further evaluated by univariate Cox regression. As a result, 25 genes

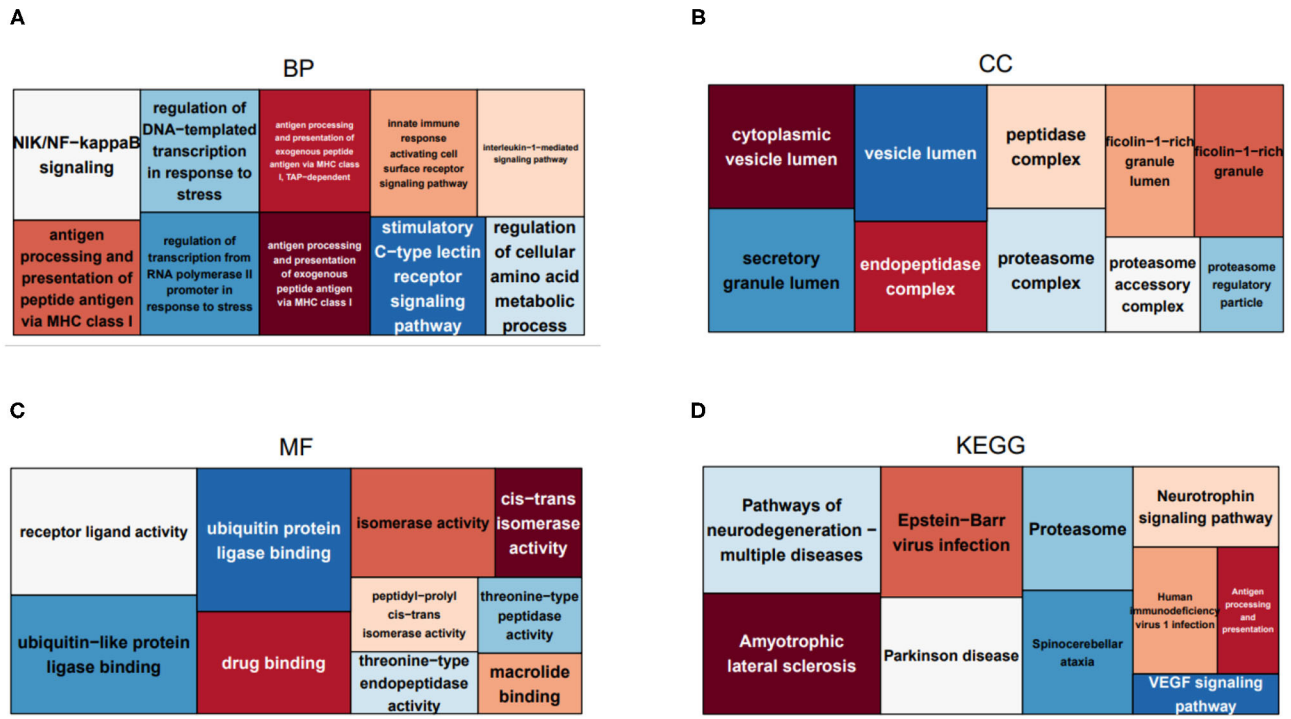


FIGURE 2 | Treemaps of GO and KEGG enrichment analysis, in which rectangle size is proportional to the statistical significance of functional annotations and pathways. **(A–C)** Treemaps of the enrichment analysis on GO biological process, cellular component, and molecular function, respectively. **(D)** Treemaps of enrichment analysis on KEGG pathways.

TABLE 1 | The 10 immune-related hub genes used to compute IRGPI risk score.

Gene	Coef	HR	HR.95L	HR.95H	*p*-value
CARS1	0.143645	1.68498	1.24329	2.283584	0.000768
CBS	−0.10811	0.656049	0.477454	0.901449	0.009325
CISD1	0.088366	1.631699	1.196967	2.224324	0.001953
GCLM	0.132237	1.445134	1.193399	1.749971	0.000163
SAT1	−0.06677	0.746933	0.576198	0.968259	0.027555
SLC7A11	0.086376	1.466116	1.232625	1.743836	1.54E-05
ACACA	0.116986	1.61311	1.228934	2.117383	0.00057
KEAP1	0.08436	1.54419	1.04808	2.275135	0.027987
SLC1A5	0.096401	1.363949	1.215323	1.530752	1.34E-07
G6PD	0.112358	1.415541	1.264641	1.584447	1.52E-09

showed statistical significance to overall survival on TCGA cohort. Next, LASSO regression was performed to further narrow the scope of OS-related hub genes, and finally 10 genes were selected for the establishment of prognostic index. Specifically, the IRGPI risk score was computed using gene expression level multiplied by the weights of the 10 genes. The weights are shown in **Table 1**.

The univariate Cox regression analysis was performed on IRGPI risk score and other clinicopathologic feature (**Figure 3A**), and verified that the IRGPI risk score and clinical stage were statistically significant factors with the prognosis of HCC (**Figure 3B**). Moreover, multivariate Cox regression analysis confirmed that IRGPI was an independent prognostic factor after adjusted for other clinicopathologic factors (**Figure 3B**).

Taking the median IRGPI score as the cutoff value to partition the TCGA cohort into low- and high- risk groups, the IRGPI low-risk patients showed obviously better OS than IRGPI high-risk patients ($p = 0.001$, log-rank test) (**Figure 3C**). The ROCAUC of the IRGPI prognosis model reach 0.790 for 1-year OS, 0.667 for 3-year OS, and 0.640 for 5-year OS (**Figure 3D**).

Furthermore, to evaluate the generalization of IRGPI in prognosis, we conducted survival analysis on validation cohort ($n = 221$). As showed in **Figure 3E**, the two IRGPI groups differed significantly in OS, and the IRGPI low-risk group had better prognosis than the IRGPI high-risk group ($p = 0.0001$, log-rank test). The ROCAUC of the prognostic model on validation cohort reach 0.806 for 1-year OS, 0.786 for 3-year OS, and 0.778 for 5-year OS (**Figure 3F**). Of note, the result on the validation cohort is consistent with that on the TCGA cohort, suggesting that the IRGPI could be a promising indicator for prognosis of HCC patients.

3.4. Regulatory Network of Immune-Related Hub Genes

The regulatory network acts as an important role in tumorigenesis and development of HCC, we attempted to

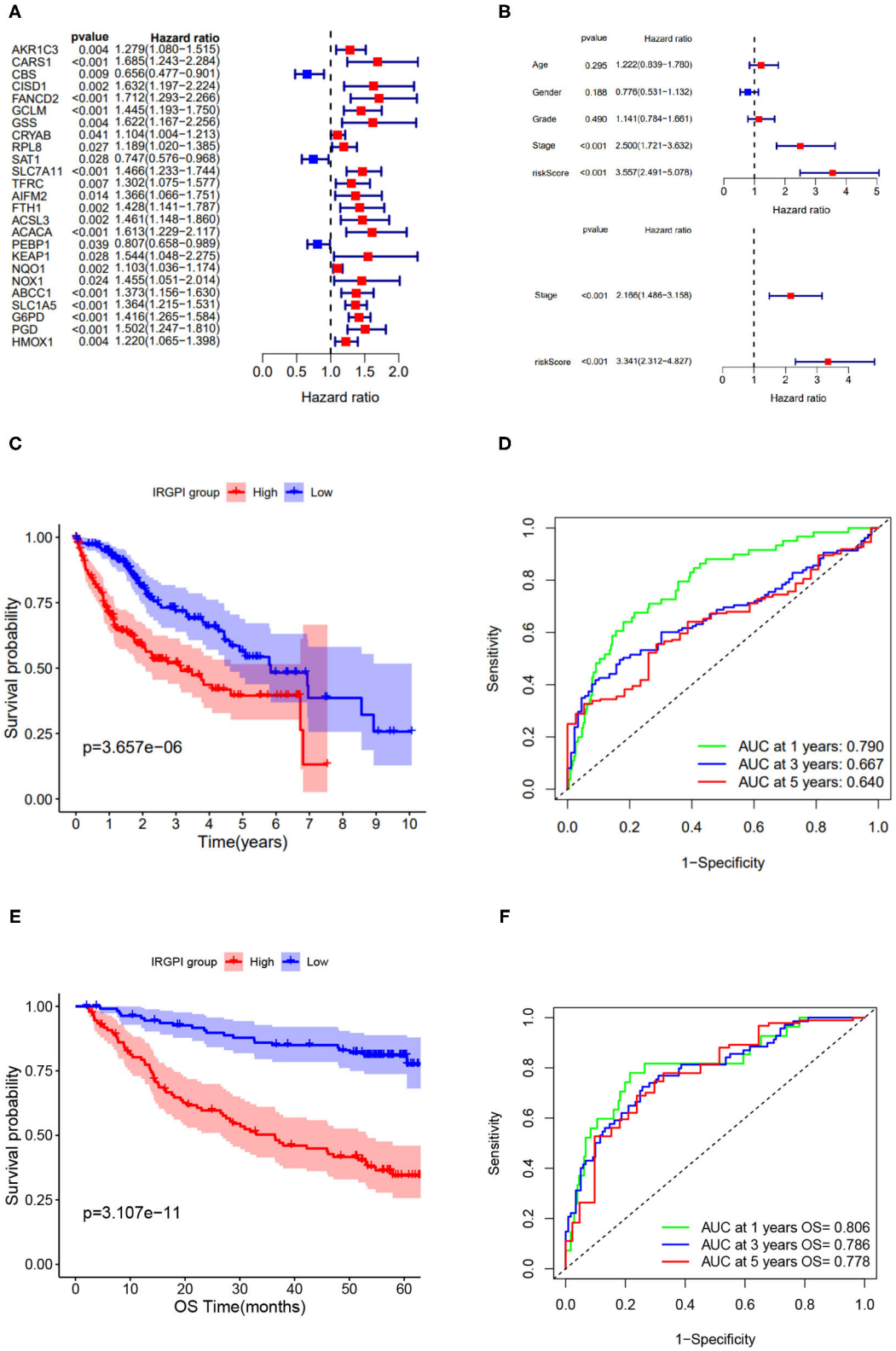


FIGURE 3 | Screening of immune-related prognostic genes *via* Cox and LASSO regression and survival analysis of IRGPI low- and high-risk groups. **(A)** Univariate Cox regression of 25 immune-related hub genes. **(B)** Univariate and multivariate Cox regression analysis on IRGPI and other clinicopathologic variables.

(Continued)

FIGURE 3 | (C) Kaplan-Meier curve of the survival analysis between IRGPI groups on TCGA cohort. **(D)** ROC curve and AUC values of IRGPI risk score in predicting 1-, 3-, and 5-year OS on TCGA cohort. **(E)** Kaplan-Meier curve of the survival analysis between IRGPI groups on GSE14520 cohort. **(F)** ROC curve and AUC values of IRGPI risk score in predicting 1-, 3-, and 5-year OS on GSE14520 cohort.

elucidate the endogenous regulatory mechanism of the immune-related hub genes. We built a regulatory network composed of mRNAs, miRNAs, lncRNAs, and transcription factors (TFs). To be specific, the regulatory network included 5 miRNAs, 6 lncRNAs, and 4 TFs associated to eight out of the immune-related hub genes, as shown in **Figure 4**.

We have conducted extensive literature study to verify these regulatory relationships. For example, the lncRNA FAM99A has been found to be up-regulated in HCC and closely related to clinical prognosis (Sun et al., 2020). There is increasing evidence that lncRNA GAS5 acts as a tumor suppressor, which is downregulated in certain tumor tissues and combined with miRNA to regulate related signaling pathways (Cheng et al., 2018). The lncRNA MALAT1 has been reported to be associated with diabetes-induced microvascular dysfunction, activate p38/MAPK signaling and regulate retinal endothelial cell function under diabetic condition (Liu et al., 2014). Another study indicated that MALAT1 regulated hyperglycemia induced inflammatory process in endothelial cells (Puthanveetil et al., 2015). The NFYA TF has been reported to be up-regulated in HCC and associated to tumors with mutant p53 (Bezzecchi et al., 2020). The CREB1 is a key transcription factor that mediates transcriptional responses to a variety of growth factors. CREB1 has been reported to be related with metastasis, tumor stage and poor outcome in gastric cancer (Wang et al., 2015), and the knockdown of CREB1 could inhibit liver cancer cell migration (Yang et al., 2013). PPARD has been reported to accelerate colorectal tumorigenesis, progression, and invasion (Liu et al., 2019), it could effectively predict the prognosis of HCC patients as an independent prognostic signature (Sun et al., 2021). It has been reported that hsa-miR-122-5p levels are connected with cholesterol levels in a viral hepatitis-free human population and associate with fatty liver and lipoprotein metabolism (Raitoharju et al., 2016).

3.5. TME and Somatic Mutations Differ in Two IRGPI Groups

We explored the tumor microenvironment difference by dissecting the composition of immune cells in different IRGPI groups. As shown in **Figure 5A**, T cells follicular helper, neutrophils, activated memory CD4 T cells, T cells gamma delta, dendritic cells resting, and M0 macrophages were more abundant in the IRGPI high-risk group, while naive B cells, resting memory CD4 T cells, and mast cells resting were more abundant in the IRGPI low-score group. The tumor microenvironment revealed that the IRGPI low-risk group reflects stronger immune response, thereby this group has better prognosis (**Figures 3C,E**).

The GSEA analysis was also performed to investigate the gene sets enriched in different IRGPI groups. The gene set in the IRGPI low-risk group was enriched in glycine serine and threonine metabolism, primary bile acid biosynthesis, and PPAR signaling pathway, while the gene set in the IRGPI high-risk group was enriched in cytosolic DNA pathway and immune related pathways (**Figure 5B**).

Somatic mutations can help to gain further insight into the tumorigenesis of HCC, thereby we plot the somatic mutation landscapes of the IRGPI low- and high-risk groups. It can be found significantly higher mutation counts in the IRGPI high-risk group than in the IRGPI low-risk group. Missense variations were the most common mutation type, followed by nonsense and frameshift deletions. We selected the top 20 genes with the highest mutation rates in two IRGPI groups (**Figures 5C,D**), and found the mutation rates of TP53, CTNNB1, TTN, MUC16, ALB, and PCLO were higher than 10% in both groups. The mutation of the MUC4 and LRP1B genes was more common in the IRGPI high-risk group, while the mutation of AXIN1 genes was more common in the IRGPI low-risk group.

Also, we assessed the immune infiltrating and tumor purity using ESTIMATE tool. The stromal score was considered to be correlated with the fraction of stromal cells, and the immune score reflected the infiltration of immune cells in solid tumor. As in previous studies, immune score has been confirmed to be correlated with prognosis in patients with several tumors (Yoshihara et al., 2013). In particular, Liu et al. has reported that patients with high immune scores had poor prognosis than those with low scores in HCC (Liu et al., 2020). Accordingly, we found that the immune score in IRGPI high-risk group was higher than those in IRGPI low-risk group (**Figure 5E**, $p < 0.05$), which is consistent with previous study.

In addition, as tumor cell proliferation leads to the overgrowth of population of clonally derived tumor cells, we explored the relationship between two IRGPI groups and tumor proliferation. As shown in **Figure 5F**, the cell proliferation was more active in the IRGPI high-risk group ($p = 3.28e\text{-}19$), and thereby leads to worse prognosis.

3.6. IRGPI Risk Score Closely Correlated to Immunotherapy Biomarkers

A few biomarkers have been used in clinical immunotherapy, including PD-L1/2, PD-1, CTLA-4, CYT, and TMB. Among these biomarkers, the immune checkpoint genes PD-L1/2, and CTLA4 are co-expressed in HCC (Shrestha et al., 2018). Beyond PD-L1/2 and CTLA4 level, the CYT value reflects the activity of cytotoxic T cells (CTLs) and NK cells due to their powerful ability to lyse tumor cells. A recent study found that CYT-high HCC has stronger immunogenicity and a more favorable TME than CYT-low HCC, which would result in better clinical outcomes (Takahashi et al., 2020). Also, TMB refers to the number of somatic mutations (non-synonymous mutations) that occur on an average of 1Mb base in the exon region. For immunotherapy, the higher TMB in cancer cells, the more antigens may be produced and thus stronger anti-tumor response in ICI therapy.

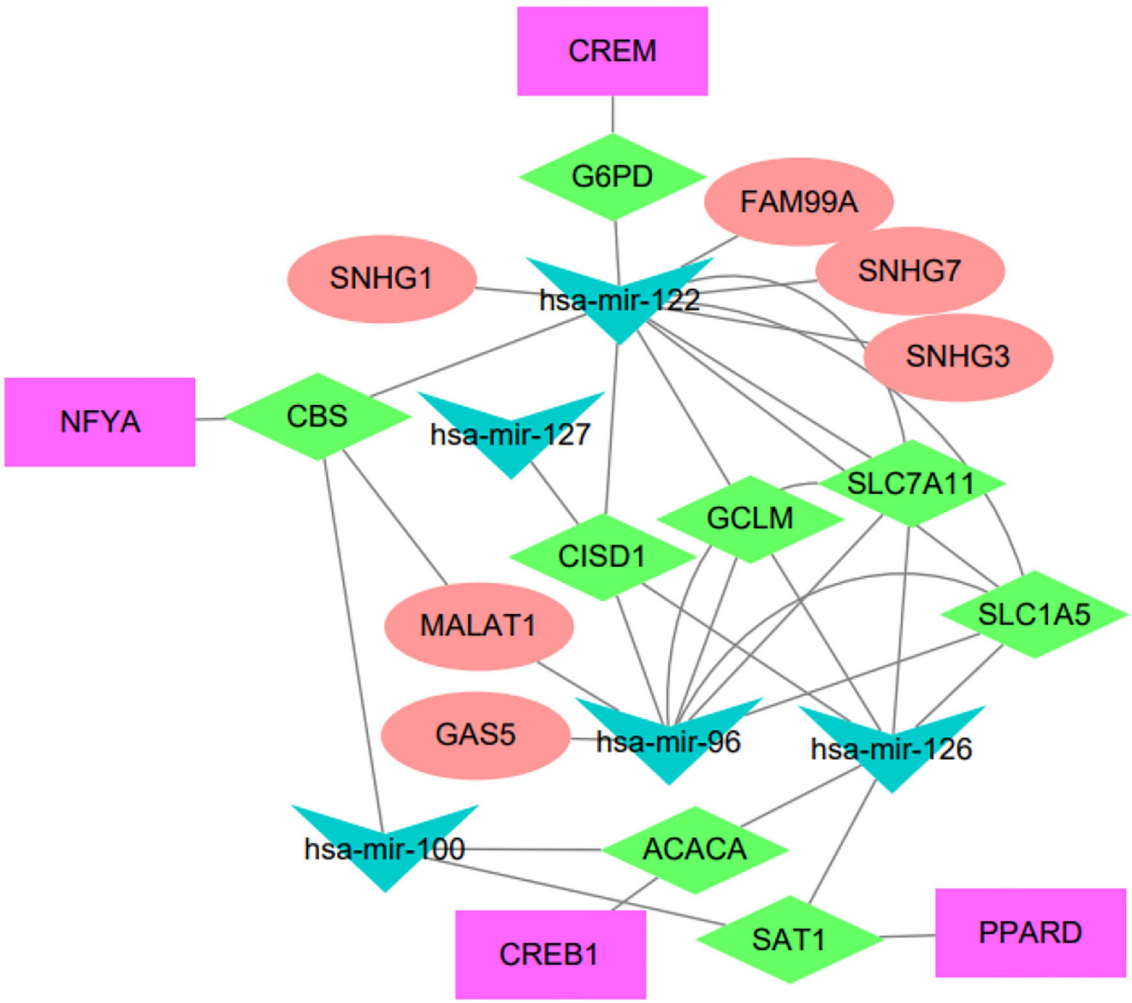


FIGURE 4 | Regulatory network of the IRGPI-related genes. The network summarizes regulation relationships between transcription factors (purple rectangles), mRNAs (green diamonds), miRNAs (blue triangles), and lncRNAs (pink eclipse).

In fact, high TMB is associated with improved response to immune checkpoint blockade in HCC (Yang et al., 2020).

We explored the relationship between the IRGPI scores and these biomarkers. As shown in **Figures 6A–F**, the IRGPI scores were positively related to the immune biomarkers. The Pearson correlation coefficients between IRGPI risk scores and PD-L1 is 0.24 with p-value = 4.6e-06 (PD-1: $r = 0.43$, $p = 2.2e\text{-}16$; PD-L2: $r = 0.2$; $p = 8.1e\text{-}05$; CTLA-4: $r = 0.32$, $p = 6e\text{-}10$; CYT: $r = 0.11$, $p = 0.038$; TMB: $r = 0.16$, $p = 0.0019$). The p-values of all the correlations were smaller than 0.05, suggesting that the IRGPI score is significantly correlated with immune biomarkers.

3.7. IRGPI Grouping Correlated to Immune Subtypes

The immune subtypes describe the immune landscape of tumors according to the tumor and stromal compartments. A consensus immune subtyping summarized four subtypes: wound healing (C1), IFN-gamma dominant(C2), inflammatory(C3), lymphocyte depleted(C4) (Thorsson et al., 2018). Thorsson et al. have showed that C3 had the best prognosis, C2 and C1 had less favorable outcomes, while C4 conferred the least favorable outcome. Correspondingly, it can be found from **Figure 7A** that C3 subtype had more IRGPI low-risk patients, while other three subtypes had more IRGPI high-risk patients (p = 0.001, chi-square test).

Another molecular subtyping has consistently reported three immune subtypes, namely, iCluster1, iCluster2, and iCluster3 (Colaprico et al., 2016). We then focused on the distribution of the molecular subtypes in the IRGPI groups. In our study, the IRGPI low-risk group comprised 30% iCluster1 samples, 41% iCluster2 samples, 29% iCluster3 samples, while the IRGPI high-risk group comprised 38% iCluster1 samples, 23% iCluster2 samples, 39% iCluster3 samples (**Figure 7B**, $p<0.05$, chi-square test). There were more samples in iCluster2 and fewer samples in iCluster1 and iCluster3 of the IRGPI low-risk group than in the IRGPI high-risk group, which is consistent with the prognosis of the molecular subtypes (Colaprico et al., 2016).

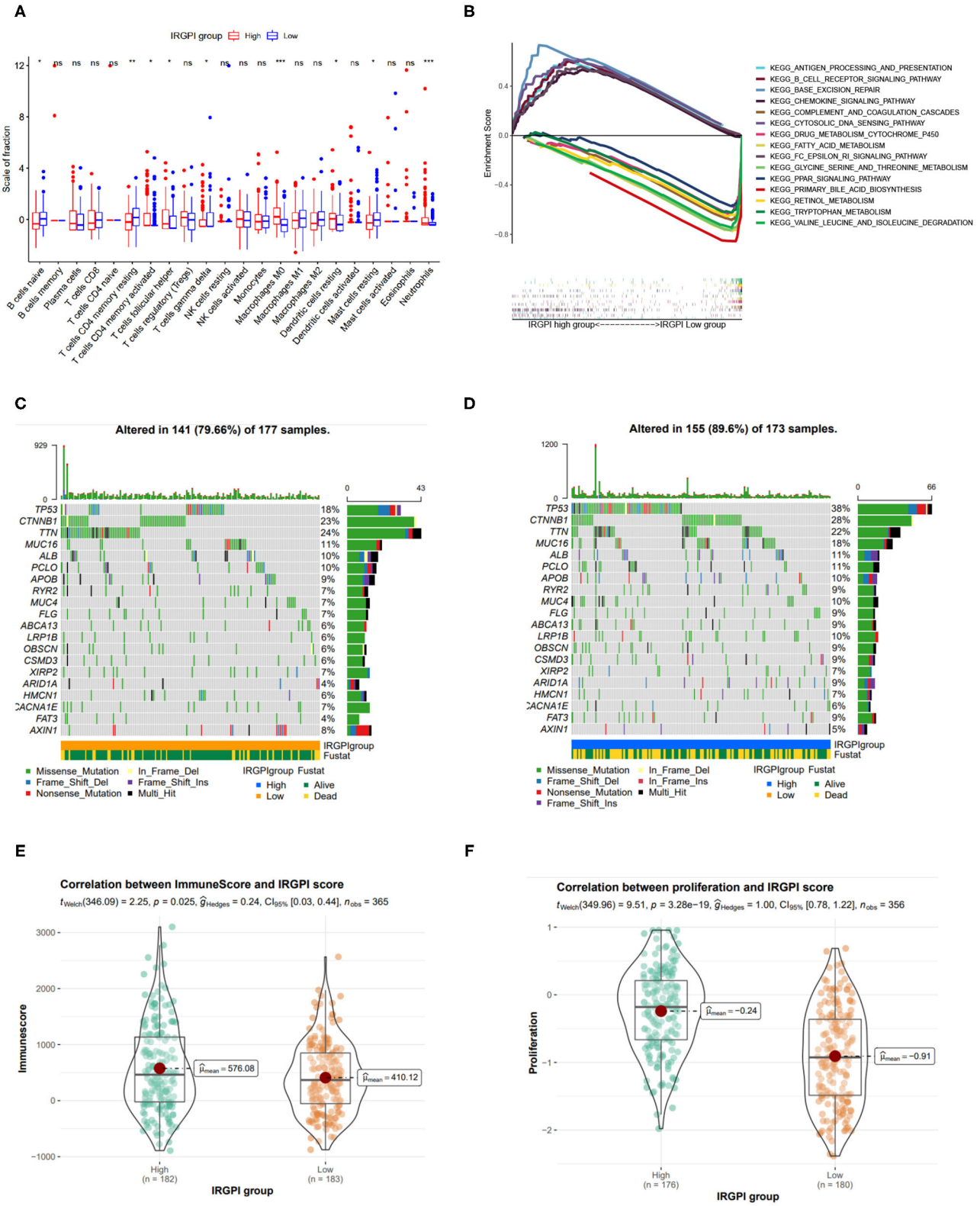


FIGURE 5 | Tumor immune environment, somatic mutation landscape, and immune scores between IRGPI low- and high-risk groups. **(A)** Comparison of tumor immune environment in two IRGPI groups. **(B)** Gene set enrichment analysis (GSEA) between two IRGPI groups. **(C,D)** Somatic mutation landscape between IRGPI low-risk and high-risk groups, respectively. **(E)** Difference of ESTIMATE immune score between two IGRPI groups. **(F)** Difference of proliferation score between two IGRPI groups.

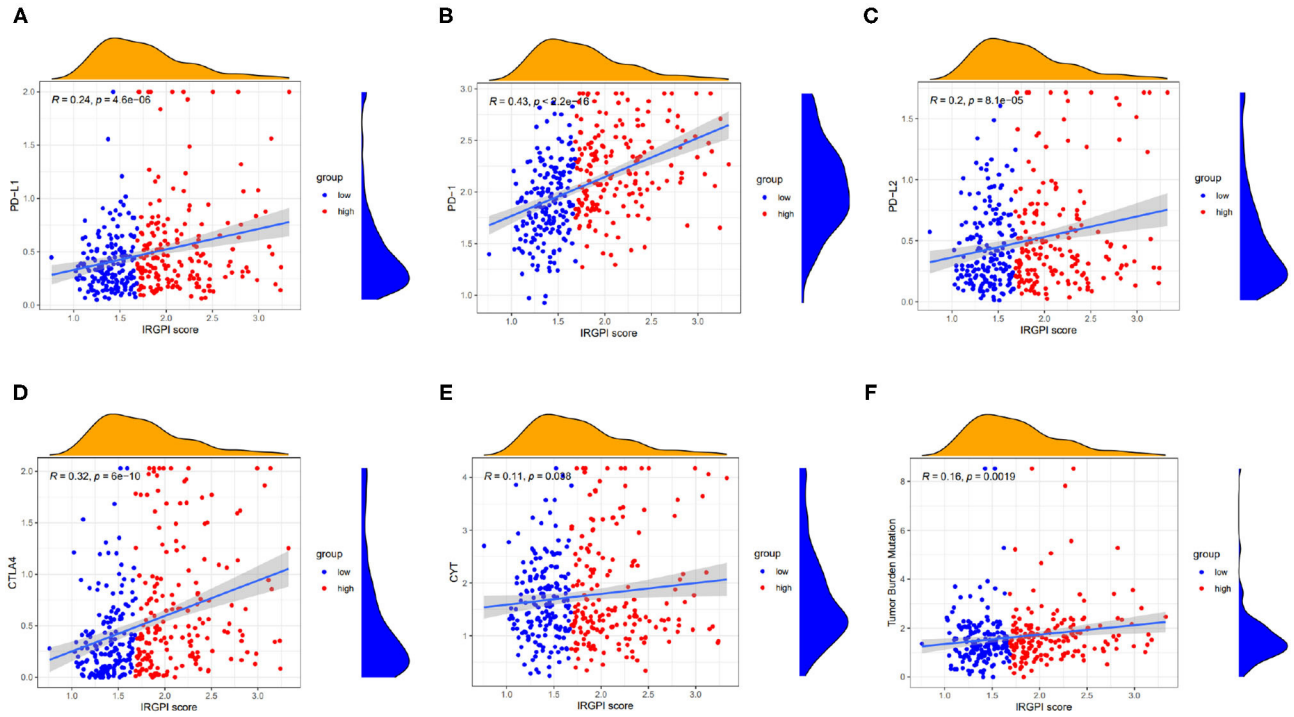


FIGURE 6 | Scatter plots coordinated by IRGPI risk score and other immune biomarkers, and **(A–F)** represents PD-L1, PD-1, PD-L2, CTLA-4, CYT, and TMB, respectively.

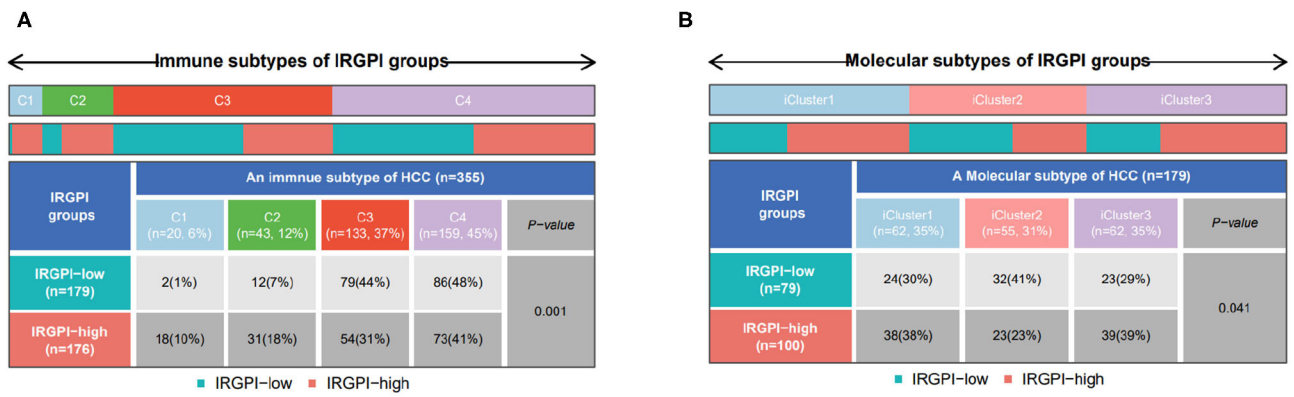


FIGURE 7 | (A) Patient proportions between IRGPI groups among TCGA immune subtypes, and **(B)** molecular subtypes.

3.8. IRGPI Is Highly Predictive of Benefit From ICI Therapy

It has been reported that some other indicators, such as TIDE and TIS, could predict patient response to ICI therapy. TIDE (Jiang et al., 2018) is a computational method to identify factors that underlie two mechanisms of tumor immune escape. Higher TIDE score means higher potential for immune evasion, which suggested that the patients were less likely to benefit from ICI therapy and associated with worse outcome.

TIDE score was used to estimate the potential clinical efficacy of ICI therapy in different IRGPI groups. The IRGPI high-risk group had higher TIDE scores than the IRGPI low-risk group ($p = 0.004$) (**Figure 8A**). Accordingly, we found that the IRGPI low-risk group had a higher microsatellite instability (MSI-H) score (**Figure 8B**), while the IRGPI high-risk group had a higher T cell exclusion score and lower T cell exclusion score in T cell dysfunction (**Figures 8C,D**). All these results verified that IRGPI high-risk patients would gain less benefit from ICI therapy than IRGPI-low patients.

We also tested the predictive performance of the IRGPI score on the efficacy of ICB treatment on another HCC cohort

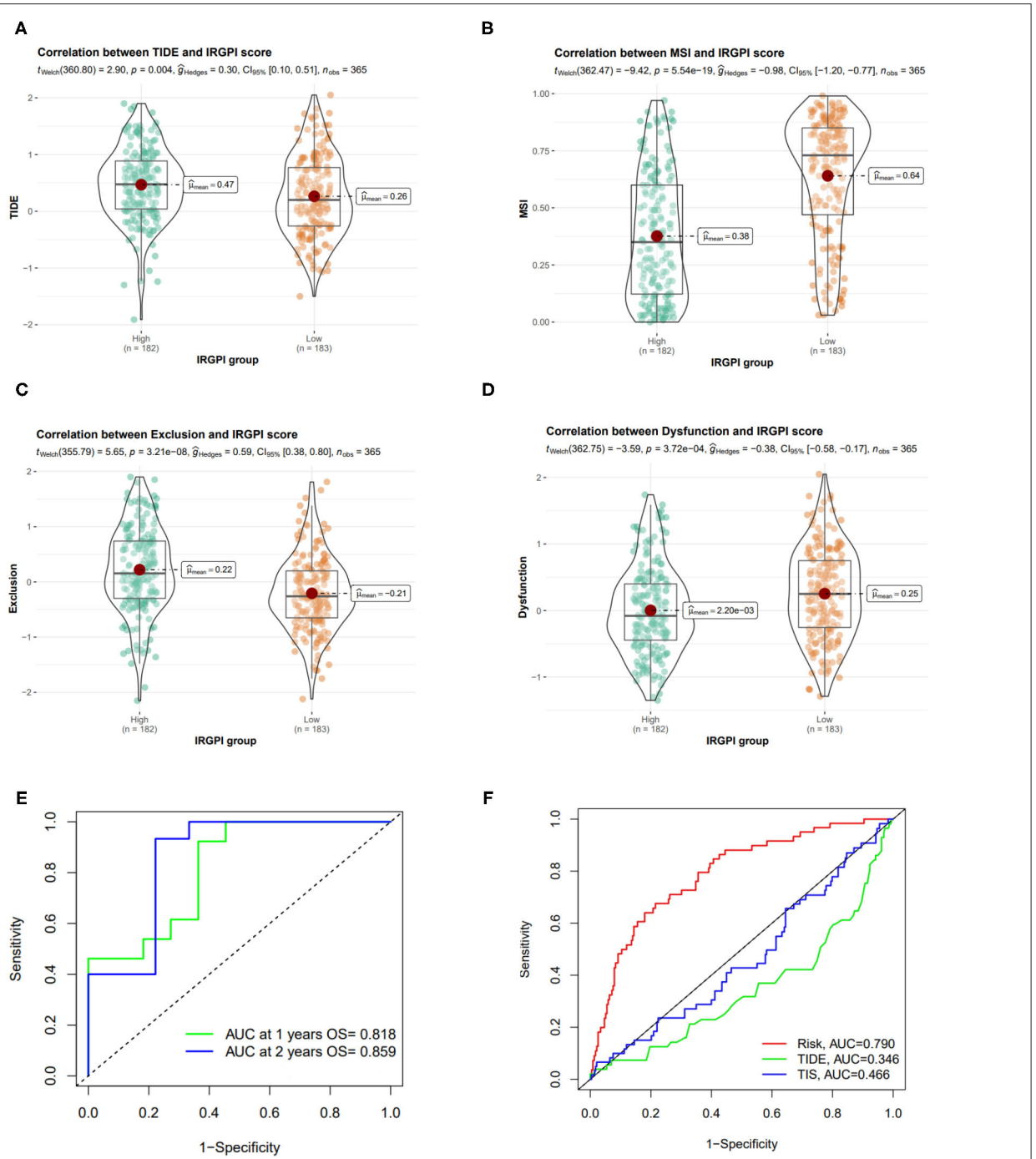


FIGURE 8 | Correlation of IRGPI risk score and other immune-related prognostic scores. **(A–D)** TIDE, MSI, T cell exclusion, and T cell dysfunction score in different IRGPI groups. **(E)** ROC curve and AUC values of IRGPI risk score in predicting the 1- and 2-year OS on GSE140901 cohort who received anti-PD-L1 therapy. **(F)** Performance comparison between of IRGPI risk, TIDE and TIS in predicting 1-year OS on GSE140901 cohort.

(GSE140901 dataset, $n = 24$). As showed in **Figure 8E**, the AUC of the ROC curve of the prognostic model reached 0.818 for 1-year OS and 0.859 for 2-year OS. Furthermore, we compare the predictive power of IRGPI with TIDE and TIS scores, and verified that the accuracy of IRGPI score was obviously higher than TIDE and TIS (**Figure 8F**). The results indicated that the IRGPI score might be a potential biomarker for predicting the immunotherapy response.

4. DISCUSSION AND CONCLUSION

Immune checkpoint inhibitor has been proven to be an effective treatment for HCC. Given that the overall response rate to ICI therapy remains limited, identifying patients who can benefit from ICI treatment is crucial. Since current widely used biomarkers, such as PD-L1 level, TMB, and MSI-H, have been proven to be not consistently reliable, we actually have no gold-standard biomarker for clinical ICI therapy yet. This highlights the need for an accurate prognostic biomarker for immunotherapy in HCC.

Based on the immune-related differentially expressed genes in HCC, we used WGCNA and regression analysis to identify 10 immune-related hub genes to establish a prognostic index. We computed the IRGPI score using weighted gene expression levels (see **Table 1**), and verified that it was an independent and effective prognostic factor. Specifically, the HCC patients with low IRGPI risk scores have improved prognosis, while those with high IRGPI risk scores have poor prognosis on TCGA cohort and two GEO cohorts. We attempted to summary the function of these genes so as to yield a mechanism explanation of the prognostic index. CARS1 is located in chromosome 11 p15.5, an important tumor-suppressor gene region. Cho et al. (2020) found that the special region of antigen-presenting cells was associated with CARS secreted by cancer cells to activate the immune response, thus stimulating a strong humoral and cellular immune response. SLC7A11 encodes a highly specific all-source of cysteine and glutamate. A few studies have shown that blocking SLC7A11 can inhibit HCC cells growth through the ROS autophagy pathway. Zhang et al. (2018) found that SLC7A11 expression profile was associated with the prognosis of liver cancer. CISD1 is an iron-containing outer mitochondrial membrane protein, and has been revealed that impaired CISD1 expression leads to tumor growth (e.g., breast and liver cancer) and has been considered a potential chemotherapeutic target (Salem et al., 2012). G6PD, a rate-limiting enzyme of the PPP, is upregulated in many cancers and contributes to tumor growth. It has been found that G6PD overexpression is significantly associated with HCC metastasis and poor prognosis of HCC (Lu et al., 2018). SLC1A5, also known as ASCT2, is one of the most studied proteins of plasma membrane transporter. Its high expression in hepatocellular carcinoma is associated with poor prognosis (Zhang et al., 2019). Consistent to these previous studies, we found these cancerogenic genes have positive weights (see **Table 1**) and thus contribute to high IRGPI risk score.

For in-depth understanding in terms of genetic alterations, we inspect the somatic mutation difference between IRGPI groups, and found that there was a great deal of difference in TP53 mutation. TP53 has more mutations in IRGPI-high samples than IRGPI-low samples (38 vs. 18%). TP53 is one of the most mutated genes in human cancers than any other gene, and linked with more aggressive disease and poorer patient outcomes in many cancers (Olivier et al., 2006), particularly in HCC (Hussain et al., 2007). Besides, there was a higher mutation rate occurred in MUC16 gene in the IRGPI high-risk group than IRGPI low-risk group (18 vs. 11%). It has been reported that high baseline MUC16 level is associated with poor prognosis in patients with HBV-related HCC (Qin et al., 2021). The two exemplar genes indicated that IRGPI high-risk patients bearing high TP53 and MUC16 mutations often have worse outcomes than IRGPI low-risk patients bearing low TP53 and MUC16 mutations, which is in agreement with the results of survival analysis.

TME also contributes to the difference of efficacy in immunotherapy. The composition of immune cells in tumor tissues was markedly different between two IRGPI groups. We observed that follicular helper T cells, neutrophils, resting dendritic cells, and M0 macrophages were more enriched in the IRGPI high-risk group, while the naive B cells, gamma delta T cells, resting memory CD4+ T cells were more enriched in the IRGPI low-risk group. Follicular helper T cells are perceived as a distinct CD4+ helper T-cell subset, which activates B-cell and products specific antibody responses, and acts as a basilic role in the progression of autoimmune disease. A substantial body of studies have revealed that follicular helper T cells suppress regulatory B cell development, meaning poor outcome in lung squamous cell carcinoma and gastric cancer (Zhang et al., 2017; Xu et al., 2020). In most tumors, M0 macrophages, a predominant subtype of macrophages, have been proven to be related to chronic inflammation and favor tumor growth and development of an invasive phenotype, and these cells have been associated with a negative association with prognosis (Le et al., 2021). Also, resting dendritic cell enriched in IRGPI high-risk group often leads to poor prognosis (Le et al., 2021). In contrast, gamma delta T cells are a distinct subset of T cells whose T cell receptors consist of γ chains and δ chains, regarded as a bridge between innate immunity and acquired immunity. Gamma delta T cells can not only directly kill a variety of tumor cells, but also exert indirect anti-tumor immune responses by facilitating the function of other immune cells, which suggests that gamma delta T cells may be a favorable prognostic factor (Ma et al., 2012).

For current biomarkers, such as PD-L1, TMB, and MSI-H, IRGPI risk score is significantly correlated to all of them. Other diagnostic indicator for ICI therapy, such as TIDE and TIS scores, have showed predictive performance in many solid tumors. TIDE is developed to identify factors underlying two mechanisms of tumor immune escape: the induction of T cell dysfunction in tumors with high infiltration of cytotoxic T lymphocytes (CTLs) and the prevention of T cell infiltration in tumors with low CTL levels (Jiang et al., 2018). Accordingly, IRGPI high-risk patients showed higher CTL infiltration and less T cell exclusion score (but not T cell dysfunction score), so their lower ICI response might be due to immune evasion *via* T cell exclusion. In contrast, the IRGPI low-risk patients had higher MSI scores, lower T cell exclusion score, which suggested these patients had lower levels of immune escape of HCC tumor cells.

In conclusion, our IRGPI risk score can characterize the tumor immune microenvironment, we believe it is a promising

immune-related prognostic index that can predict response to ICI treatment and overall survival of HCC patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YL and HL made a comprehensive conception and analyzed the data. YL and LZ made collection of the data and implemented the analysis in a statistical manner. XL and HL gave a help to upgrade the conception and make the framework of the manuscript complete. YL was responsible for writing the draft paper. HL was responsible for revision and polished the drafts of the paper. JL and HL reinforced the supervision of the study and supplied funding support. All authors read and commented on the manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (62072058; 82073339).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.760079/full#supplementary-material

REFERENCES

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., et al. (2017). Ifn-γ-related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* 127, 2930–2940. doi: 10.1172/JCI91190

Baehrecke, E. H., Dang, N., Babaria, K., and Shneiderman, B. (2004). Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics* 5:84. doi: 10.1186/1471-2105-5-84

Bezzecchi, E., Ronzio, M., Mantovani, R., and Dolfini, D. (2020). NF-Y overexpression in liver hepatocellular carcinoma (HCC). *Int. J. Mol. Sci.* 21:9157. doi: 10.3390/ijms21239157

Bhattacharya, S., Andorf, S., Gomes, L., Dunn, P., Schaefer, H., Pontius, J., et al. (2014). Immport: disseminating data to the public for the future of immunology. *Immunol. Res.* 58, 234–239. doi: 10.1007/s12026-014-8516-1

Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., et al. (2013). Innatedb: systems biology of innate immunity and beyond-recent updates and continuing curation. *Nucleic Acids Res.* 41, D1228–D1233. doi: 10.1093/nar/gks1147

Cheng, K., Zhao, Z., Wang, G., Wang, J., and Zhu, W. (2018). lncRNA GAS5 inhibits colorectal cancer cell proliferation via the miR-182-5p/FOXO3a axis. *Oncol. Rep.* 40, 2371–2380. doi: 10.3892/or.2018.6584

Cho, S., Kim, S., Lee, Y., Song, E., Kim, U., Kim, H., et al. (2020). Endogenous TLR2 ligand embedded in the catalytic region of human cysteinyl-tRNA synthetase 1. *J. Immunother. Cancer* 8:e000277. doi: 10.1136/jitc-2019-000277

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302. doi: 10.1093/nar/gkx1067

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44:e71. doi: 10.1093/nar/gkv1507

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Ghouri, Y. A., Mian, I., and Rowe, J. H. (2017). Review of hepatocellular carcinoma: Epidemiology, etiology, and carcinogenesis. *J. Carcinog.* 16:1. doi: 10.4103/jcar.JCar__16

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* 38, 675–678. doi: 10.1038/s41587-020-0546-8

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., et al. (2018). Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 46, D380–D386. doi: 10.1093/nar/gkx1013

Hastie, T., and Qian, J. (2014). *Glmnet Vignette* (retrieved June 9).

Hou, J., Zhang, H., Sun, B., and Karin, M. (2020). The immunobiology of hepatocellular carcinoma in humans and mice: basic concepts and therapeutic implications. *J. Hepatol.* 72, 167–182. doi: 10.1016/j.jhep.2019.08.014

Hussain, S., Schwank, J., Staib, F., Wang, X., and Harris, C. (2007). Tp53 mutations and hepatocellular carcinoma: insights into the etiology and pathogenesis of liver cancer. *Oncogene* 26, 2166–2176. doi: 10.1038/sj.onc.1210279

Jeggari, A., Marks, D. S., and Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* 28, 2062–2063. doi: 10.1093/bioinformatics/bts344

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24, 1550–1558. doi: 10.1038/s41591-018-0136-1

Khalil, D. N., Smith, E. L., Brentjens, R. J., and Wolchok, J. D. (2016). The future of cancer treatment: immunomodulation, cars and combination immunotherapy. *Nat. Rev. Clin. Oncol.* 13, 273–290. doi: 10.1038/nrclinonc.2016.25

Langfelder, P., and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Le, T., Su, S., and Shahriyari, L. (2021). Immune classification of osteosarcoma. *Math. Biosci. Eng.* 18:1879. doi: 10.3934/mbe.2021098

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20. doi: 10.1016/j.cell.2004.12.035

Liu, J., Yao, J., Li, X., Song, Y., Wang, X., Li, Y., et al. (2014). Pathogenic role of lncRNA-malat1 in endothelial cell dysfunction in diabetes mellitus. *Cell Death Dis.* 5:e1506. doi: 10.1038/cddis.2014.466

Liu, X., Niu, X., and Qiu, Z. (2020). A five-gene signature based on stromal/immune scores in the tumor microenvironment and its clinical implications for liver cancer. *DNA Cell Biol.* 39, 1621–1638. doi: 10.1089/dna.2020.5512

Liu, Y., Deguchi, Y., Tian, R., Wei, D., Wu, L., Chen, W., et al. (2019). Pleiotropic effects of PPARD accelerate colorectal tumorigenesis, progression, and invasion. *Cancer Res.* 79, 954–969. doi: 10.1158/0008-5472.CAN-18-1790

Llovet, J. M., Kelley, R. K., Villanueva, A., Singal, A. G., Pikarsky, E., Roayaie, S., et al. (2021). Hepatocellular carcinoma. *Nat. Rev. Dis. Primers.* 7:6. doi: 10.1038/s41572-020-00240-3

Lu, M., Lu, L., Dong, Q., Yu, G., Chen, J., Qin, L., et al. (2018). Elevated G6PD expression contributes to migration and invasion of hepatocellular carcinoma cells by inducing epithelial-mesenchymal transition. *Acta Biochim. Biophys. Sin.* 50, 370–380. doi: 10.1093/abbs/gmy009

Ma, C., Zhang, Q., Ye, J., Wang, F., Zhang, Y., Wevers, E., et al. (2012). Tumor-infiltrating $\gamma\delta$ t lymphocytes predict clinical outcome in human breast cancer. *J. Immunol.* 189, 5029–5036. doi: 10.4049/jimmunol.1201892

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*. 28, 1747–1756. doi: 10.1101/gr.239244.118

Mellman, I., Coukos, G., and Dranoff, G. (2011). Cancer immunotherapy comes of age. *Nature* 480, 480–489. doi: 10.1038/nature10673

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Nishino, M., Ramaiya, N. H., Hatabu, H., and Hodi, F. S. (2017). Monitoring immune-checkpoint blockade: response evaluation and biomarker development. *Nat. Rev. Clin. Oncol*. 14, 655–668. doi: 10.1038/nrclinonc.2017.88

Olivier, M., Langer, A., Carrieri, P., Bergh, J., Klaar, S., Eyfjord, J., et al. (2006). The clinical value of somatic tp53 gene mutations in 1,794 patients with breast cancer. *Clin. Cancer Res*. 12, 1157–1167. doi: 10.1158/1078-0432.CCR-05-1029

Puthanveetil, P., Chen, S., Feng, B., Gautam, A., and Chakrabarti, S. (2015). Long non-coding RNA malat 1 regulates hyperglycaemia induced inflammatory process in the endothelial cells. *J. Cell. Mol. Med*. 19, 1418–1425. doi: 10.1111/jcmm.12576

Qin, C., Gao, Y., Li, J., Huang, C., and He, S. (2021). Predictive effects of preoperative serum ca125 and AFP levels on post-hepatectomy survival in patients with hepatitis b-related hepatocellular carcinoma. *Oncol. Lett*. 21, 1–13. doi: 10.3892/ol.2021.12748

Raitoharju, E., Seppälä, I., Lyytikäinen, L.-P., Viikari, J., Ala-Korpela, M., Soininen, P., et al. (2016). Blood hsa-miR-122-5p and hsa-miR-885-5p levels associate with fatty liver and related lipoprotein metabolism-the young finns study. *Sci. Rep*. 6, 1–13. doi: 10.1038/srep38262

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). LIMMA powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*. 43:e47. doi: 10.1093/nar/gkv007

Salem, A. F., Whitaker-Menezes, D., Howell, A., Sotgia, F., and Lisanti, M. P. (2012). Mitochondrial biogenesis in epithelial cancer cells promotes breast cancer tumor growth and confers autophagy resistance. *Cell Cycle* 11, 4174–4180. doi: 10.4161/cc.22376

Sangro, B., Sarobe, P., Hervás-Stubbs, S., and Melero, I. (2021). Advances in immunotherapy for hepatocellular carcinoma. *Nat. Rev. Gastroenterol. Hepatol*. 18, 525–543. doi: 10.1038/s41575-021-00438-0

Shrestha, R., Prithviraj, P., Anaka, M., Bridle, K. R., Crawford, D. H., Dhungel, B., et al. (2018). Monitoring immune checkpoint regulators as predictive biomarkers in hepatocellular carcinoma. *Front. Oncol*. 8:269. doi: 10.3389/fonc.2018.00269

Sun, G., Sun, K., and Shen, C. (2021). Human nuclear receptors (NRS) genes have prognostic significance in hepatocellular carcinoma patients. *World J. Surg. Oncol*. 19, 1–12. doi: 10.1186/s12957-021-02246-x

Sun, M., Lv, S., and Zhong, J. (2020). Identification of lncRNA FAM99A gene as a prognostic biomarker of hepatocellular carcinoma. doi: 10.21203/rs.3.rs-16068/v1

Takahashi, H., Kawaguchi, T., Yan, L., Peng, X., Qi, Q., Morris, L. G., et al. (2020). Immune cytolytic activity for comprehensive understanding of immune landscape in hepatocellular carcinoma. *Cancers* 12:1221. doi: 10.3390/cancers12051221

Therneau, T. M., and Lumley, T. (2014). *Package-Survival*. Survival analysis Published on CRAN. Rochester, Minnesota.

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Yang, T.-H. O., et al. (2018). The immune landscape of cancer. *Immunity* 48, 812–830. doi: 10.1016/j.immuni.2018.03.023

Wang, Y.-W., Chen, X., Gao, J.-W., Zhang, H., Ma, R.-R., Gao, Z.-H., et al. (2015). High expression of cAMP responsive element binding protein 1 (CREB1) is associated with metastasis, tumor stage and poor outcome in gastric cancer. *Oncotarget* 6:10646. doi: 10.18632/oncotarget.3392

Wong, N., and Wang, X. (2015). miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*. 43, D146–D152. doi: 10.1093/nar/gku1104

Xu, F., Zhang, H., Chen, J., Lin, L., and Chen, Y. (2020). Immune signature of t follicular helper cells predicts clinical prognostic and therapeutic impact in lung squamous cell carcinoma. *Int. Immunopharmacol*. 81:105932. doi: 10.1016/j.intimp.2019.105932

Yang, X., Shi, J., Chen, X., Jiang, Y., and Zhao, H. (2020). Efficacy of cabozantinib and nivolumab in treating hepatocellular carcinoma with ret amplification, high tumor mutational burden, and PD-L1 expression. *Oncologist* 25:470. doi: 10.1634/theoncologist.2019-0563

Yang, Z., Tsuchiya, H., Zhang, Y., Hartnett, M. E., and Wang, L. (2013). MicroRNA-433 inhibits liver cancer cell migration by repressing the protein expression and function of camp response element-binding protein. *J. Biol. Chem*. 288, 28893–28899. doi: 10.1074/jbc.M113.502682

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun*. 4, 1–11. doi: 10.1038/ncomms3612

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an R package for comparing biological themes among gene clusters. *Omics* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, H., Yue, R., Zhao, P., Yu, X., Li, J., Ma, G., et al. (2017). Proinflammatory follicular helper t cells promote immunoglobulin g secretion, suppress regulatory b cell development, and correlate with worse clinical outcomes in gastric cancer. *Tumor Biol*. 39:1010428317705747. doi: 10.1177/1010428317705747

Zhang, L., Huang, Y., Ling, J., Zhuo, W., Yu, Z., Luo, Y., et al. (2018). Overexpression of SLC7A11: a novel oncogene and an indicator of unfavorable prognosis for liver carcinoma. *Future Oncol*. 14, 927–936. doi: 10.2217/fon-2017-0540

Zhang, P., Wang, Q., Lin, Z., Yang, P., Dou, K., and Zhang, R. (2019). Berberine inhibits growth of liver cancer cells by suppressing glutamine uptake. *OncoTargets Ther*. 12:11751. doi: 10.2147/OTT.S235667

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Li, Zou, Liu, Luo and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 04 November 2021
doi: 10.3389/fcell.2021.782473

Competitive Endogenous RNA Landscape in Epstein-Barr Virus Associated Nasopharyngeal Carcinoma

Xiandong Lin[1,2†], Steven Wang[3†], Keyu Lin[1], Jingfeng Zong[4], Qianlan Zheng[1], Ying Su[1] and Tao Huang[5]**

[1]Laboratory of Radiation Oncology and Radiobiology, Fujian Medical University Cancer Hospital and Fujian Cancer Hospital, Fuzhou, China, [2]Fujian Provincial Key Laboratory of Translational Cancer Medicine, Fuzhou, China, [3]Department of Biological Sciences, Columbia University, New York, NY, United States, [4]Department of Radiotherapy, Fujian Medical University Cancer Hospital and Fujian Cancer Hospital, Fuzhou, China, [5]Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Jie Wang,
Guangzhou Institutes of Biomedicine and Health (CAS), China
Yungang Xu,
Xi'an Jiaotong University, China

***Correspondence:**
Ying Su
zjsuying@126.com
Tao Huang
tohuangtao@126.com

†These authors share first authorship

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 24 September 2021
Accepted: 20 October 2021
Published: 04 November 2021

Citation:
Lin X, Wang S, Lin K, Zong J, Zheng Q, Su Y and Huang T (2021) Competitive Endogenous RNA Landscape in Epstein-Barr Virus Associated Nasopharyngeal Carcinoma. Front. Cell Dev. Biol. 9:782473. doi: 10.3389/fcell.2021.782473

Non-coding RNAs have been shown to play important regulatory roles, notably in cancer development. In this study, we investigated the role of microRNAs and circular RNAs in Nasopharyngeal Carcinoma (NPC) by constructing a circRNA-miRNA-mRNA co-expression network and performing differential expression analysis on mRNAs, miRNAs, and circRNAs. Specifically, the Epstein-Barr virus (EBV) infection has been found to be an important risk factor for NPC, and potential pathological differences may exist for EBV+ and EBV- subtypes of NPC. By comparing the expression profile of non-cancerous immortalized nasopharyngeal epithelial cell line and NPC cell lines, we identified differentially expressed coding and non-coding RNAs across three groups of comparison: cancer vs. non-cancer, EBV+ vs. EBV- NPC, and metastatic vs. non-metastatic NPC. We constructed a ceRNA network composed of mRNAs, miRNAs, and circRNAs, leveraging co-expression and miRNA target prediction tools. Within the network, we identified the regulatory ceRNAs of *CDKN1B, ZNF302, ZNF268,* and *RPGR*. These differentially expressed axis, along with other miRNA-circRNA pairs we identified through our analysis, helps elucidate the genetic and epigenetic changes central to NPC progression, and the differences between EBV+ and EBV- NPC.

Keywords: nasopharyngeal carcinoma (NCP), epstein-barr virus (EBV), circRNA, miRNA, network

INTRODUCTION

Nasopharyngeal carcinoma (NPC) is an Epstein-Barr virus (EBV) associated malignancy with a characteristic geographical distribution. Globally, NPC is a rare condition, with less than one case per 100,000 people per year. However, the occurrence of NPC is much more common among the populations in Southern China and Southeast Asia, with up to 25–50 cases per 100,000 people per year (Jain et al., 2016). Ethnic Chinese born in North America develop NPC less frequently compared to those in Southern China, implying that both genetic susceptibility and environmental factors contribute to the development of NPC (Buell, 1974). Ample evidence shown that infection of EBV is a risk factor for NPC (Tsao et al., 2017; Chan et al., 2018). For instance, EBV genome and gene products are detected in virtually all tumors in NPC-endemic areas. Increased levels of IgA

antibodies against EBV antigens, among other EBV-related biomarkers, have been used for early detection and screening for NPC in a few high-incidence areas (Shotelersuk et al., 2000; Li et al., 2018; Wu et al., 2018). Nevertheless, while many risk factors have been established, our understanding of the molecular regulatory mechanisms that lead to the development of NPC is limited.

The role of various non-coding RNAs in the regulation of many biological pathways and functions have been widely explored and verified in recent years. MicroRNAs (miRNAs) are non-coding RNAs around 22 nucleotides in length that play important regulatory roles, specifically by inhibiting gene expression through cleavage of mRNA or translational repression (Bartel, 2004). In addition to mRNAs, miRNAs interact with any target RNAs that contain complementary sites known as miRNA response elements (MREs). Since miRNAs can bind to multiple targets, the ceRNA hypothesis was proposed, stating that these target RNAs compete for a limited amount of miRNA (Salmena et al., 2011). In other words, the amount of ceRNAs can collectively impact the degree to which miRNAs regulate gene expression.

Circular RNAs (circRNAs) is a type of non-coding RNA structured as a covalently-bonded closed continuous loop, where the 5'-cap and 3'-poly-A tail are joined together. Studies have found circRNAs to play important regulatory roles in NPC growth and metastasis (Chen et al., 2019; Guo et al., 2019; Zhu et al., 2019; Yang et al., 2020). It is widely accepted that circRNAs inhibit target miRNA activity through a miRNA sponge mechanism, which in turn results in an upregulation of target gene expression (Panda, 2018; Zu et al., 2020). A study by Zhu et al. showed that highly expressed circ-ZNF609 absorbs microRNA-150-5p to upregulate Sp1 expression, which in turn promotes the proliferation and metastatic ability of NPC (Zhu et al., 2019). Another study by Yang et al. utilized a circRNA-miRNA-target gene network to reveal potential mechanism between circKITLG and miR-3198 (Yang et al., 2020). Furthermore, studies have shown the circRNA expression is tissue specific (Memczak et al., 2013; You et al., 2015). In short, the interactions between circRNAs and miRNAs have significant influence on key genes, and as a consequence affect the development and progression of cancer (Peng and Croce, 2016; Hirono et al., 2019; Hoey et al., 2019; Hong et al., 2019; Liu et al., 2020).

To analyze and illustrate this complicated collection of interactions, we constructed a circRNA-miRNA-mRNA network. We also performed differential expression (DE) analysis on three groups of comparison: cancerous vs. non-cancerous, EBV+ vs. EBV- cell lines, and metastatic vs. non-metastatic samples. Using these results, we identified axes of circRNA-miRNA-mRNA that are differentially expressed in each group of comparison. We identified 6 differentially expressed circRNA-miRNA-mRNA axes between EBV+ and EBV- cell lines, out of which we highlighted the *hsa_circ_0008129/miR-221-3p/CDKN1B* axis. Through this research, we identified several potential ceRNA axes that regulate NPC progression or EBV associated traits in NPC. These ceRNA pathways can help better understand the molecular landscape of NPC, and help guide therapeutic efforts.

RESULTS

Overview of Computational Approach

We obtained the mRNA, miRNA, and circRNA expression profile for four cancer cell lines and four patient samples. We then performed differential expression (DE) analysis and constructed RNA interaction networks to isolate circRNA-miRNA-mRNA axes of interest. Simultaneously, we performed functional enrichment of RNAs of interest to reveal relevant pathways in the pathology of NPC. We provided a graphical outline of the computational workflow in **Figure 1**. Lastly, we cross-referenced our analysis results with two publicly available NPC RNA-seq datasets, GSE143797 (Yang et al., 2020) and GSE118721 (Lin et al., 2018), to highlight intersecting DE-RNAs.

Differential Expression Analysis and Functional Enrichment Analysis

DE analysis on mRNA, miRNA, and circRNA was performed across four cell lines, falling under three groups of comparison: cancer (C666, CNE-2, SUNE-1) vs. non-cancer (NP69); EBV+ (C666) vs. EBV- (CNE-2, SUNE-1); and metastatic vs. non-metastatic using patient samples. In the cancer vs. non-cancer and EBV+ vs EBV- DE comparisons, where more than one set of DE analysis was performed, we focused on the intersecting differentially expressed mRNAs, miRNAs, circRNAs (DE-mRNAs, DE-miRNAs, DE-circRNAs). We plotted the top 10 overexpressed and under-expressed DE-RNA by log2 fold change. The summary of the DE analysis results across cell lines and samples for mRNA, miRNA, and circRNA was shown respectively in **Figure 2**; **Figure 3**; **Figure 4**. The entire DE analysis results can be found in **Supplementary Table S1**; **Supplementary Table S2**; **Supplementary Table S3** which contains all DE analysis results of cancer vs. non-cancer cell lines, EBV+ vs. EBV- cell lines, metastatic vs. non-metastatic patient samples, respectively.

We then performed functional enrichment analysis on the intersecting set of DE-mRNA across each of the three groups of analysis (cancer, EBV, metastatic). We searched the GO Biological Pathways (BP), Molecular Functions (MF), and Cellular Component (CC) database using an over-representation test in the clusterProfiler R package (Wu et al., 2021). We highlighted functionally relevant and statistically significant pathways (**Figures 2D–F**). Similarly, we performed functional enrichment analysis on the intersecting set of DE-miRNA under each category of comparison. We used the miEAA website (Kern et al., 2020), searching across the GO, KEGG, and miRWalk (Sticht et al., 2018) databases. We highlighted functionally relevant and statistically significant pathways in **Figures 3D–F**. We then performed miRNA target gene enrichment using the miRTarBase (Huang et al., 2020) database, under the miEAA. The target enrichment results can be found in **Supplementary Table S4**.

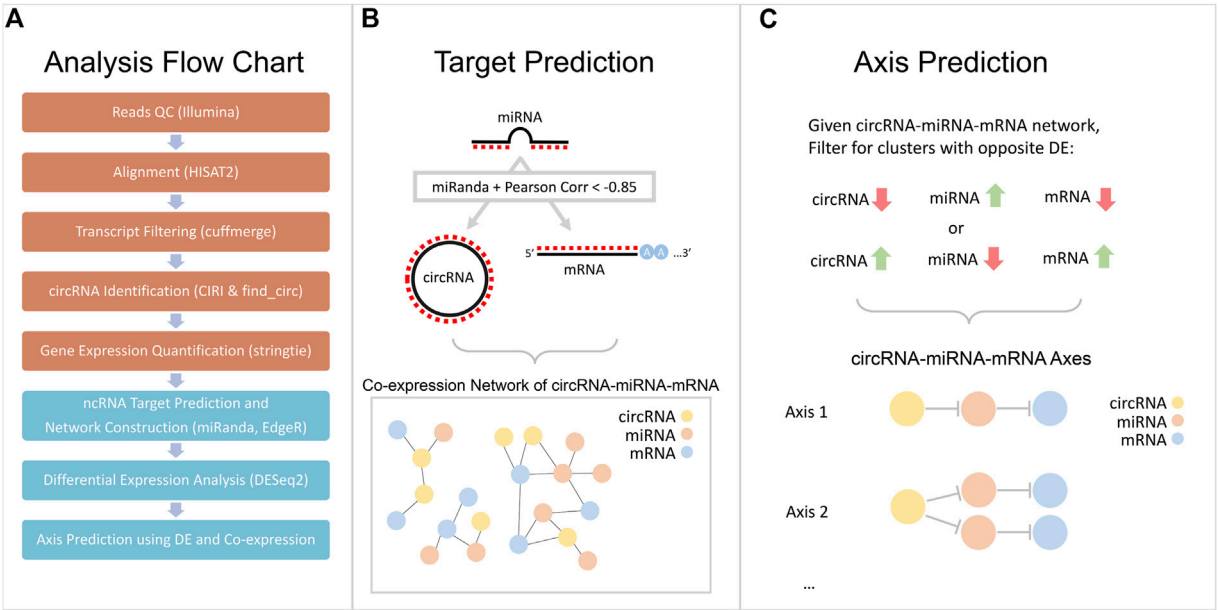


FIGURE 1 | Flowchart of computational workflow. **(A)** Four cell lines and two patient samples were sequenced and processed to obtain an expression quantification for all three types of RNAs: circRNA, miRNA, and mRNA. **(B)** A graphical illustration of the construction of a co-expression network using Pearson correlation and miRanda target prediction. **(C)** A graphical illustration of axis prediction method, utilizing DE results and co-expression network.

Enrichment of miRNA Target Genes and circRNA Source Genes

We performed target gene enrichment on the set of DE-miRNAs under each group of DE comparison. We utilized the miEAA (Kern et al., 2020) to perform Over-Representation Analysis (ORA), separately for the overexpressed and under-expressed DE-miRNAs to capture the directionality of mRNA regulation. **Supplementary Figure S1** highlighted interesting functional enrichment results of DE-miRNA, and the full enrichment result can be found in **Supplementary Table S4**. In theory, the target genes for the set of overexpressed DE-miRNA would be down-regulated, and vice versa. Studies have shown that circRNAs regulate the expression of its source gene, such as circSEP3 and circSMARCA5 (Conn et al., 2017; Xu et al., 2020). To evaluate circRNA's role as source gene regulators in NPC, we performed ORA using the KEGG database on the source genes of DE-circRNAs under each DE group (**Supplementary Table S5**).

Target Prediction and Network Construction

We constructed a circRNA-miRNA-mRNA co-expression network to investigate the role of ceRNA regulation in NPC. The Pearson correlation between all possible circRNA-miRNA and miRNA-mRNA pairs were calculated, and pairs with a correlation coefficient < -0.85 were considered to be significant. In efforts to reduce false positives, we used miRanda and its default parameters to determine whether the circRNA-miRNA and miRNA-mRNA pairs were valid targets, and removed pairs that were not deemed target pairs (**Supplementary Table S6**). Lastly, we filtered for miRNAs that were paired to at least one mRNA and circRNA to isolate complete ceRNA axes. The interaction network analysis yielded 428 miRNA-mRNA pairs, and 131 miRNA-circRNA pairs. We then extracted a subset of the network that contained only DE-miRNAs and their associated miRNAs and circRNAs, and visualized it using Cytoscape (**Figure 5**).

Differentially Expressed circRNA-miRNA-mRNA Axes

The previous network analysis allowed us to gain an overview of the ceRNA landscape in NPC. However, not all elements in the network are implicated in NPC, therefore we leveraged DE analysis results to reveal the functionally relevant axes. From the previously constructed co-expression network, we extracted axes that contain at least one DE-circRNA and one DE-miRNA, and with opposite directions of differential expression between circRNA-miRNA and mRNA-miRNA pairs. In this process, we implicitly removed circRNAs that did not function as miRNA sponges, but instead regulate gene expression through other plausible routes such as binding RNA polymerases (Zhang et al., 2013), inducing methylation (Chen N. et al., 2018), or alternative splicing (Conn et al., 2017). These circRNAs are false positives, and will not show opposite differential expression to its predicted target miRNAs. In efforts to reduce false negatives, we included high scoring (TargetScore≥97) miRNA-mRNA pairs from the miRDB database (Chen and Wang, 2020). The result across three groups of DE comparisons: cancer (C666, CNE-2, SUNE-1) vs. non-cancer (NP69), EBV+ (C666) vs. EBV- (CNE-2,

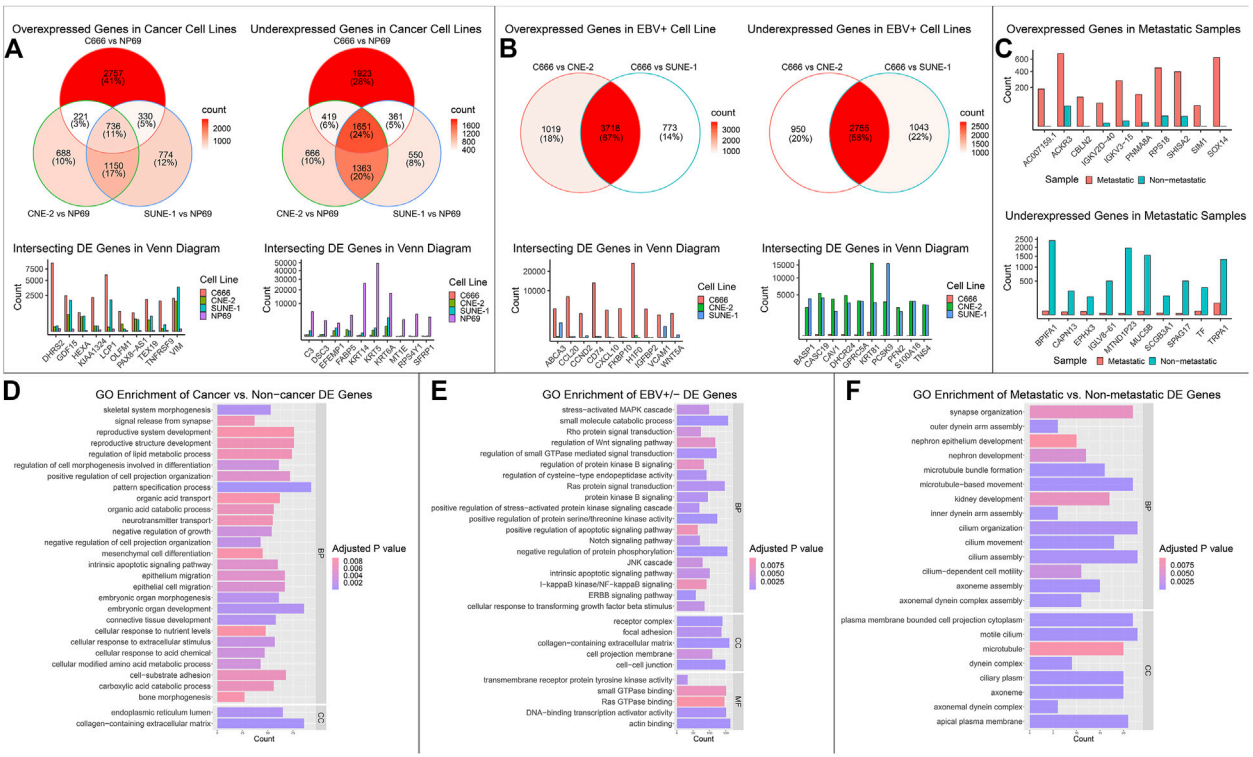


FIGURE 2 | Overview of mRNA landscape across comparisons. The differential expression (DE) results for mRNAs across the three groups of comparison: cancer vs. non-cancer, EBV + vs. EBV-, and metastatic vs non-metastatic, were integrated into one figure. **(A)** The cancer vs. non-cancer group consisted of three sets of DE analysis: NP69 (non-cancer) vs. C666, CNE-2, and SUNE-1 (cancer). The expression of top DE genes that intersected all three sets of DE analysis was plotted. **(B)** The EBV+ vs. EBV- group consisted of two sets of DE analysis: C666 (EBV+) vs. CNE-2, SUNE-1 (EBV-). The expression of intersecting DE genes was plotted. **(C)** Since there was only one set of DE analysis, the expression of top DE genes was plotted. **(D–F)** The intersecting DE genes in each comparison were functionally enriched using the GO database, and top enriched pathways were plotted. The x-axis value represented the gene count in the corresponding pathway, and the color of the bar represented the adjusted *p* value of the over representation test.

SUNE-1), and metastatic vs. non-metastatic patient samples, are shown in **Table 1**; **Table 2**; **Table 3**, respectively.

In **Table 1**, the differentially expressed circRNA-miRNA-mRNA axes in cancerous (C666, SUNE-1, CNE-2) vs. non-cancer (NP69) cell lines were listed. They must show significant differential expression between cancer and noncancerous cell lines (|log2 fold|≥ 1.0 and adjusted *p* value ≦ 0.05) and the miRNA must show negative correlation (correlation coefficient < −0.85) with predicted target by miRanda or miRDB. Adjusted *p* values and insignificant DE-mRNAs associations were omitted for visual clarity. The hsa_circ_103,862 is verified to bind miR-493-5p (Wang et al., 2020).

Similarly, in **Table 2**, the differentially expressed circRNA-miRNA -mRNA axis in EBV+ (C666) vs EBV- (CNE-2, SUNE-1) cell lines were listed. A positive value indicated higher expression in the C666 cell line and vice versa. They must show significant differential expression between EBV+ and EBV- cell lines (|log2 fold|≥ 1.0 and adjusted *p* value ≦ 0.05) and the miRNA must show negative correlation (correlation coefficient < −0.85) with predicted target by miRanda or miRDB. Insignificant DE-mRNAs associations and adjusted *p* values were omitted for visual clarity. CDKN1B, Ensemble ID ENSG00000111276, is verified to be a target of miR-221-3p *via* luciferase reporter assay (Fornari et al., 2008; Yin et al., 2019).

In **Table 3**, differentially expressed miRNAs in two metastatic vs. two non-metastatic patient samples were shown. A positive value indicated higher expression in the metastatic group and vice versa. Only miRNAs showing significant differential expression (| log2 fold| ≥ 1.0 and adjusted *p* value ≦ 0.05) were included.

From the list of circRNA-miRNA-mRNA axes, we identified the hsa_circ_0008129/miR-221-3p/CDKN1B axis, which was significantly differentially expressed in EBV+ (C666) vs. EBV- (CNE-2, SUNE-1) cell lines. We also highlighted the ceRNA axis behind *ZNF302, ZNF268, TAB3*, *SAR1B*, *BRMS1L, ZIC2*, and *MBNL2*.

Differential Expression Analysis in Public Data

To verify our computational findings and highlight promising DE-RNAs, we utilized two additional public datasets: GSE143797 for circRNA expression in NPC and normal tissue (Yang et al., 2020), GSE118721 for miRNA and mRNA expression in NPC and normal tissue (Lin et al., 2018). We performed a Two-way Student T test to test for differentially expressed RNAs in tumor

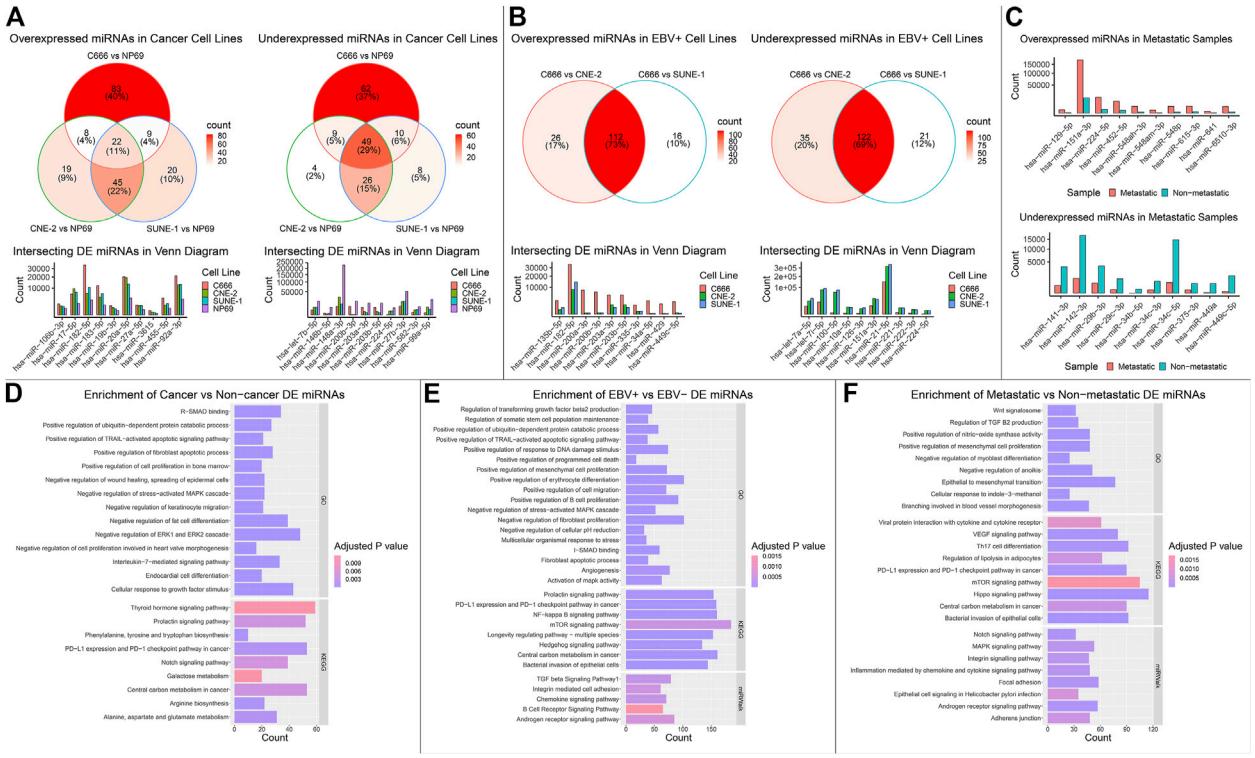


FIGURE 3 | Overview of miRNA landscape across comparisons. The differential expression (DE) results for miRNAs across the three groups of comparison were integrated into one figure. **(A–C)** The expression of top intersecting DE-miRNA in each comparison was plotted. **(D–F)** The intersecting DE-miRNAs in each comparison were functionally enriched using the miEAA website, surveying across GO, KEGG, and miRWalk databases. The top enriched pathways were plotted. The x-axis value represented the gene count in the corresponding pathway, and the color of the bar represented the adjusted *p* value of the over representation test.

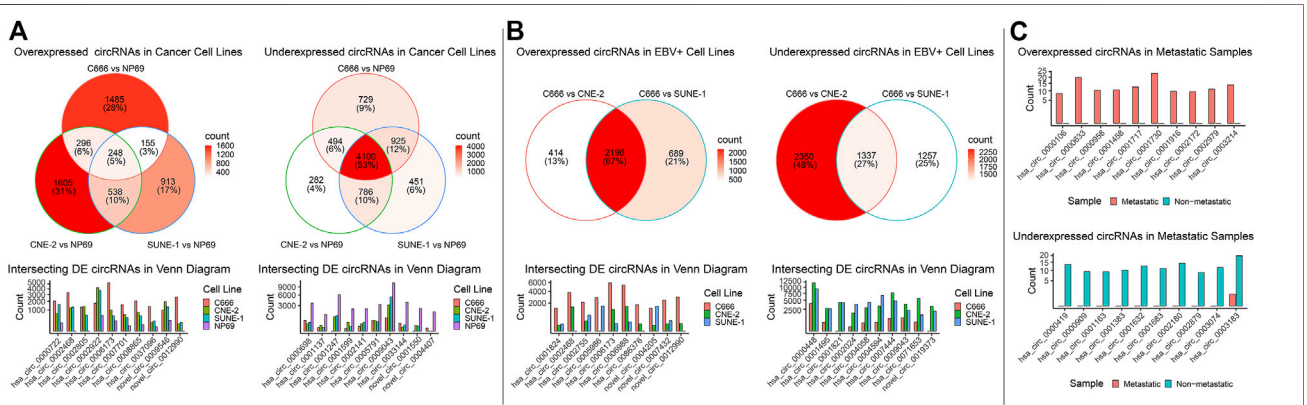


FIGURE 4 | Overview of circRNA landscape across comparisons. The differential expression (DE) results for circRNAs across the three groups of comparison were integrated into one figure. **(A, C)** The expression of top intersecting DE-circRNA in each comparison was plotted. Functional enrichment was not performed as there were no commonly recognized tools for circRNA enrichment.

verses normal tissue, and reported genes with adjusted *p* value ≦ 0.05 and whose direction of differentiation aligns with our data (**Supplementary Table S7**). We plotted the most highly differentially expressed mRNA, miRNA, and circRNA, and their respective fold changes in data from this study and prior ones (**Figure 6**).

DISCUSSION

The role of ceRNAs in inducing the epigenetics changes necessary for the development of tumors is a topic of great research interest. The ceRNA network method is an efficient method to capture the complexity of interactions between a diverse pool of ceRNAs, and

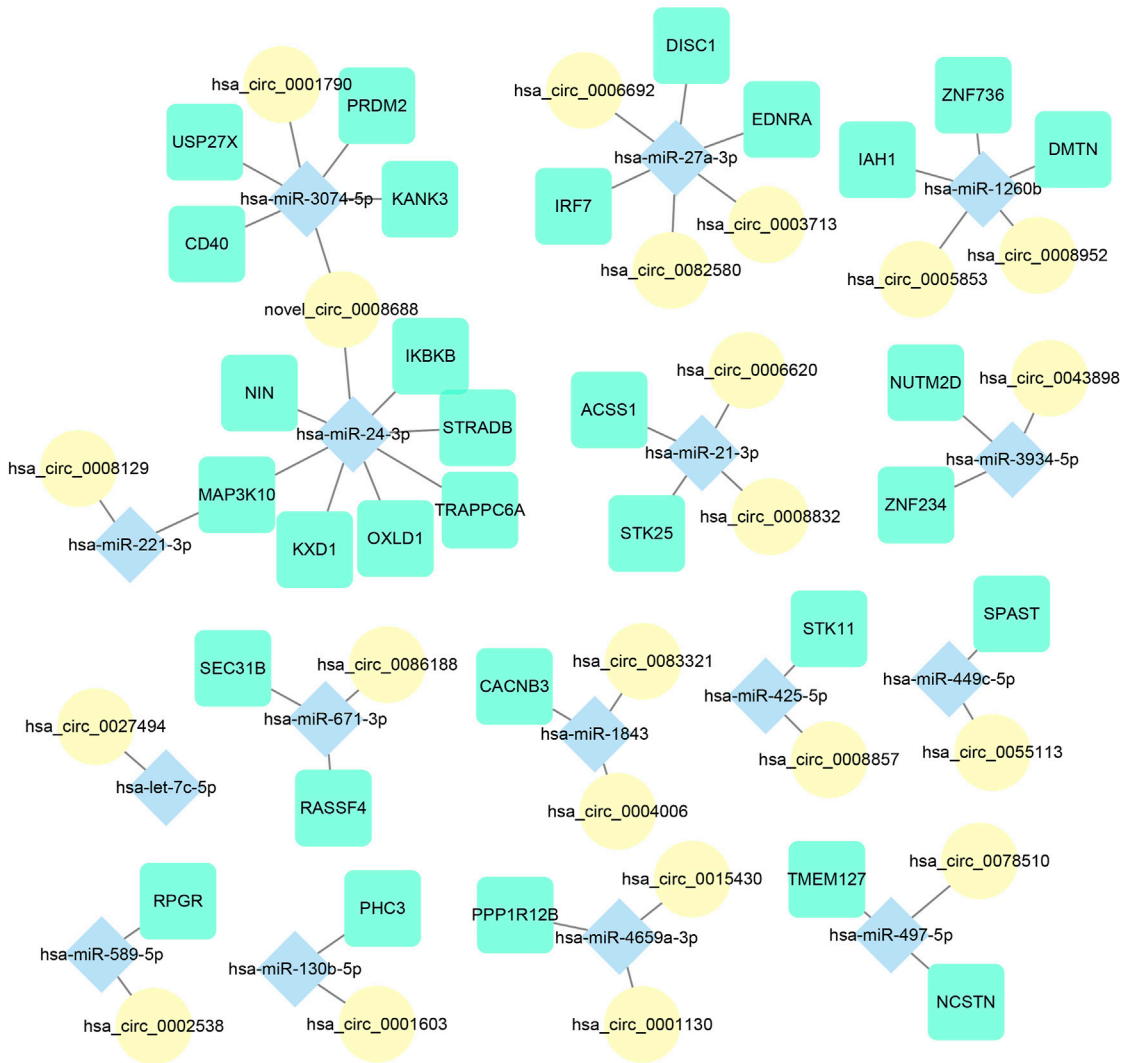


FIGURE 5 | Visualization of circRNA-miRNA-mRNA network. Edges represent RNA pairs that have Pearson correlation coefficient < −0.85 and verified target interaction calculated by miRanda. mRNAs, miRNAs, and circRNAs were illustrated using different shapes and colors.

different studies have demonstrated its effectiveness in discovering important epigenetic changes in cancer. Specifically, this method allows the emerging role of circRNAs as both miRNA sponges and direct transcription regulators to be integrated into the omics landscape of cancer.

In this study, we profiled the expression of mRNAs, miRNAs, and circRNAs in NPC cell lines and non-cancerous human immortalized nasopharyngeal epithelial cell line. We utilized the differential expression of mRNAs, miRNAs, circRNAs, and the construction of a co-expression network to identify differentially expressed circRNA-miRNA-mRNA axes in NPC. Many of the axes we found contain well-established oncogenic miRNAs, and potentially point to post-transcriptional regulatory events in NPC. We found convincing evidence for the existence of a *hsa_circ_0002538/miR-589-5p/RPGR* axis dysregulated in NPC, and a *hsa_circ_0008129/miR-221-3p/CDKN1B* axis, which is differentially regulated in EBV+ vs. EBV- cell lines. Again, the final collection of axes underwent a stringent set of filters, namely, the RNA components have to: negatively co-expressed with correlation coefficient < −0.85, exhibit significant DE in its respective group, and predicted to interact by the miRanda software or miRDB database. It is nevertheless important to keep in mind that cell culture often drift from its ancestral expression profiles, therefore our findings warrant further validation using ideally fresh tissue samples.

TABLE 1 | DE circRNA-miRNA-mRNA axis in cancer vs non-cancer comparison.

Axis	RNA type	RNA ID	C666 vs NP69 Log2	SUNE-1 vs NP69 Log2	CNE-2 vs NP69 Log2
1	circRNA	hsa_circ_0002538	+1.55	+1.25	Not significant
	miRNA	hsa-miR-589-5p	−3.02	−1.10	−1.45
	mRNA	RPGR	+2.65	+1.48	+1.44
2	miRNA	hsa-miR-27b-3p	−2.87	−1.27	−1.35
	mRNA	TAB3	+1.23	+1.53	+1.25
3	miRNA	hsa-miR-20a-5p	+2.35	+2.03	+2.62
	mRNA	SAR1B	−2.12	−1.68	−1.66
	mRNA	BRMS1L	−1.13	−2.10	−1.61
4	miRNA	hsa-miR-493-5p	−5.60	−4.61	−5.76
	mRNA	ZIC2	+7.55	+7.80	+6.70
	mRNA	MBNL2	+1.91	+1.16	+0.90

TABLE 2 | Differentially expressed circRNA-miRNA-mRNA axes in EBV+ vs EBV- cell lines.

Axis	RNA	RNA ID	C666 vs CNE-2 Log2FC	C666 vs SUNE-1 Log2FC
1	circRNA	hsa_circ_0027494	+9.23	+9.37
	miRNA	hsa-let-7c-5p	−2.60	−2.63
2	circRNA	hsa_circ_0005853	+10.27	+10.41
	circRNA	hsa_circ_0008952	+7.57	+1.69
	miRNA	hsa-miR-1260b	−4.06	−3.23
	mRNA	ZNF268	+3.06	+3.79
	mRNA	ZNF302	+10.07	+10.01
3	circRNA	hsa_circ_0008129	+8.89	+9.03
	miRNA	hsa-miR-221-3p	−3.86	−3.90
	mRNA	CDKN1B	+1.76	+1.17
	mRNA	RIMS3	+5.63	+3.57
4	circRNA	novel-circ-0008688	+3.21	+4.28
	circRNA	hsa_circ_0001790	+8.16	+8.29
	miRNA	hsa-miR-24-3p	−2.75	−2.60
	miRNA	hsa-miR-3074-5p	−2.76	−2.60
5	circRNA	hsa_circ_0055113	−9.22	−8.71
	miRNA	hsa-miR-449c-5p	+6.42	+7.28
6	circRNA	hsa_circ_0078510	−7.79	−7.60
	miRNA	hsa-miR-497-5p	+5.66	+5.11

TABLE 3 | Differentially expressed miRNAs in metastatic vs non-metastatic patient samples.

RNA type	RNA ID	Log2FC	Adjusted *p* value
miRNA	hsa-miR-130b-5p	+2.03	0.00605
miRNA	hsa-miR-449c-5p	−6.93	1.09E-19
miRNA	hsa-miR-589-5p	+2.45	0.00204
miRNA	hsa-miR-671-3p	+1.82	0.0348

Cancer vs. Noncancer Cell Lines: Differentially Expressed circRNA -miRNA-mRNA Axes

We have found evidence in Axis 1 (**Table 1**) for a potential hsa_circ_0002538↑miR-589-5p↓RPGR↑ axis. miR-589-5p is a well-established cancer-associated miRNA, found to inhibit *MAP3K8* in hepatocellular carcinoma (Zhang et al., 2016), regulate tumor growth in HCC by targeting *MIG-6* (Xu et al., 2018), and act as tumor suppressor in prostate cancer (Ji et al., 2019). Alongside the under-expressed miR-589-5p, we observe an overexpression of hsa_circ_0002538, predicted to bind to miR-589-5p. We postulate that dysregulation of hsa_circ_0002538 is the source of downstream abnormalities in expression. *RPGR* encodes the retinitis pigmentosa GTPase regulator, whose function in NPC and cancer is unknown. Given that other various GTPases are well known in their regulatory roles in cancer, it is plausible that RPGR is associated with NPC progression (Fernández-Medarde and Santos, 2011; Prieto-Dominguez et al., 2019; Boudhraa et al., 2020; Clayton and Ridley, 2020). It is also possible that *RPGR* is simply a passenger event in the axis.

Furthermore, we identified three miRNA-mRNA axes, for which no DE-circRNA component was found in our analysis. In Axis 2 (**Table 1**), miR-27b-3p↓TAB3↑, over-expression of

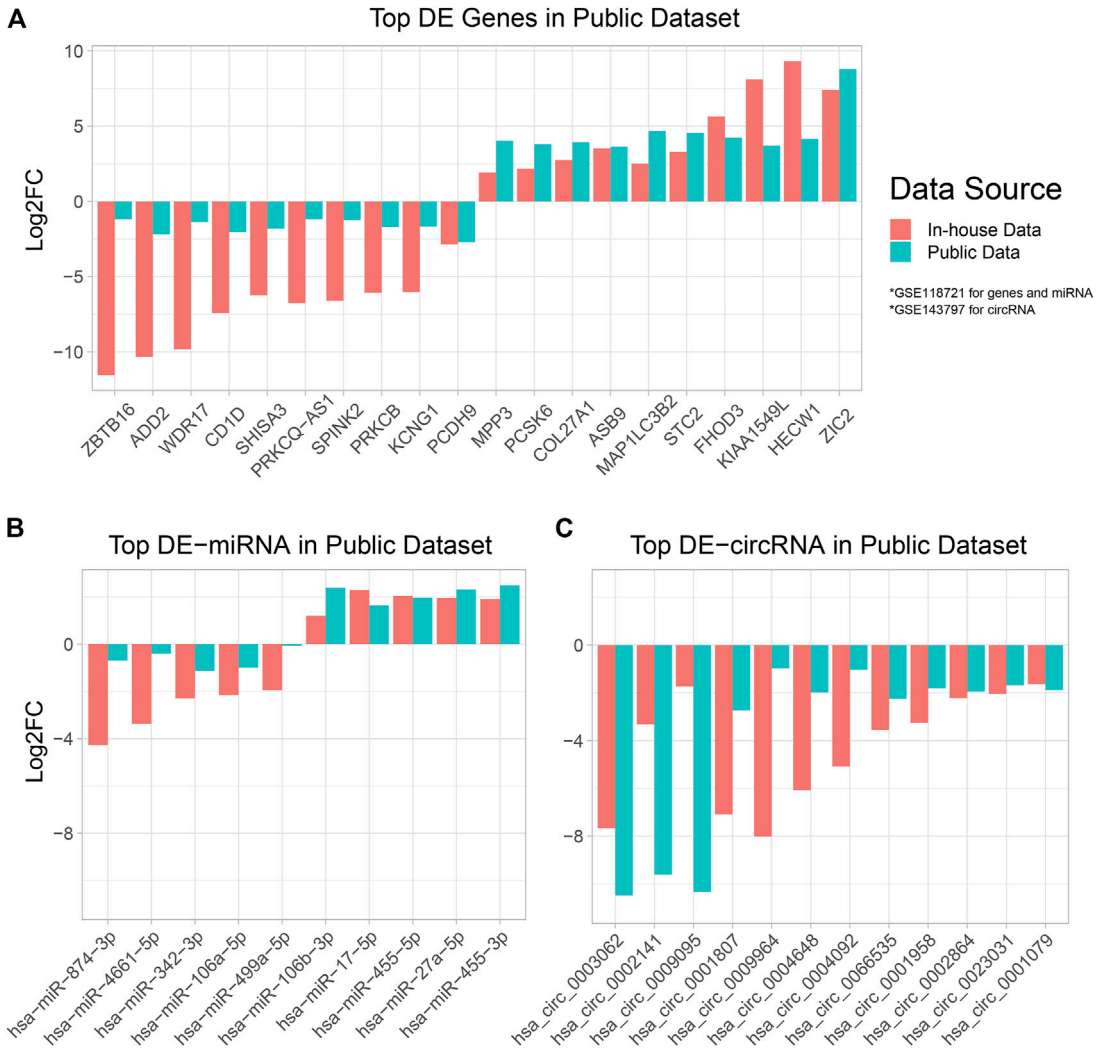


FIGURE 6 | Bar plot of DE-RNAs in cross-referenced public datasets. The differentially expressed mRNA, miRNA, and circRNAs across two public datasets GSE118721 (Lin et al., 2018) and GSE143797 (Yang et al., 2020) were established and plotted alongside in-house data. A positive Log2FC denotes overexpression in tumor tissue, and a negative Log2FC denotes under-expression in tumor tissue. **(A,B)** DE performed using data from this study and GSE118721 (Lin et al., 2018). **(C)** DE performed using data from this study and GSE143797 (Yang et al., 2020).

TAB3 has been found to promote tumor progression in NSCLC (Chen et al., 2016), colorectal cancer (Luo et al., 2017), and triple-negative breast cancer (Tao et al., 2016). Under-expression of miR-27b-3p has been shown to induce drug resistance in breast cancer (Chen D. et al., 2018). In Axis 3 (**Table 1**), miR-20a-5p↑ SAR1B↓ BRMS1L↓, reduced BRMS1L in breast cancer tissues was shown to be associated with metastasis and poor patient survival (Gong et al., 2014). Specifically, Gong et. Al. found that BRMS1L inhibited epithelial-mesenchymal transition, and thus inhibiting breast cancer metastasis. In Axis 4 (**Table 1**), miR-493-5p↓ ZIC2↑ MBNL2↑, MBNL2 is abnormally expressed in lung and breast cancer (Zhang J. et al., 2019), as well as hepatocellular carcinoma (Lee et al., 2016). Interestingly, despite overexpressed in these cancer types, both studies showed that MBNL2 suppresses tumor progression. ZIC2 was shown to be highly overexpressed in NPC both in the data used in this study, as well as in a study by Lv et al. (Lv et al., 2021) and Lin et al. (Lin et al., 2018).

EBV+ vs. EBV- Cell Lines: Differentially Expressed circRNA-miRNA -mRNA Axes

In comparison to few complete axes in the cancer vs non-cancer comparison, the EBV+ (C666) vs. EBV- (CNE-2, SUNE-1) comparisons revealed several axes of circRNA-miRNA exhibiting differential expression. Surveying through the DE results of target miRNA-circRNA pairs, we observed that the directions of differential expression are almost uniformly opposite, which confirms the theory that circRNAs serve as miRNA sponges.

In Axis 3 (**Table 2**) we observed hsa_circ_0008129↑miR-221-3p↓*CDKN1B*↑*RIMS3*↑ in EBV+ (C666) cell line. Prior studies

confirmed that miR-221-3p targets the cell cycle regulator *CDKN1B*, also known as p27, through luciferase reporter assay (Fornari et al., 2008; Yin et al., 2019). *CDKN1B* encodes the cyclin dependent kinase inhibitor 1B protein, belonging to the Cip/Kip protein family. CDKN1B functions as a cell cycle check point at G_1-S by repressing cyclin-dependent kinases that are necessary for progression from G_1 to S phase (Sherr and Roberts, 1999). Overall, *CDKN1B* is mainly dysregulated at the post-transcriptional level in human cancer, which supports our finding of a significantly differentially expressed ceRNA axis. Further, *CDKN1B* is described as haplo-insufficient tumor suppressor gene, where animals lacking one copy of *CDKN1B* already displayed tumor-prone phenotypes (Fero et al., 1998). RIMS3 is not reported to be associated with cancer to the best of our knowledge, and was likely a passenger event.

The role of the miRNA component of this axis, miR-221-3p, has been studied extensively across cancer types. A study by Wang et al. showed that miR-221-3p serve as potential prognostic predictors for hepatocellular carcinoma (Wang et al., 2019). Other studies have shown miR-221-3p to be involved in drug resistance in glioma cells (Milani et al., 2019) and breast cancer (Chen et al., 2020) by increasing antiapoptotic abilities. There have been no studies done on hsa_circ_0008129 to the best of our knowledge. Our results show evidence for a *hsa_circ_0008129/miR-221-3p/CDKN1B* axis, which potentially plays a role in explaining the pathological differences between EBV+ and EBV- NPC.

In Axis 2 (**Table 2**), we observed an overexpression of miR-1260b in all EBV- cell lines. MiR-1260b is a well-studied oncogenic miRNA, and has been found to regulate proliferation in non-small cell lung carcinoma (NSCLC) (Xu et al., 2015; Xia et al., 2019) and prostate cancer (Hirata et al., 2014). *ZNF268* is predicted to be a high affinity target of miR-1260b, with a target score of 100 (the maximum of target score) in the miRDB database (Chen and Wang, 2020), and is found to be overexpressed in cervical cancer (Wang et al., 2012) and ovarian carcinomas (Hu et al., 2013). Specifically, Wang et al. found that knockdown of *ZNF268* in cervical cancer cells caused cell cycle arrest at the G0/G1 phase (Wang et al., 2012). Similarly, *ZNF302* is predicted to be a miR-1260b target gene (score = 98), and high expression of *ZNF302* is associated with poor survival in Endometrial Carcinoma in TCGA-UCEC (**Supplementary Figure S2**). We found both *ZNF268* and *ZNF302* to be overexpressed in EBV + cell line C666, which may explain why there is higher malignancy in the EBV + subtype of NPC. Alongside miR-1260b and zinc finger proteins, we observed up-regulation of hsa_circ_005853 and hsa_circ_0008952 in the EBV+ subtype of NPC., which were predicted to bind miR-1260b. These circRNAs have not been studied to the best of our knowledge. The above results suggest that the Axis 2 (**Table 2**) plays a key role in the progression of NPC and warrants further studies.

In Axis 1, 4, 5, 6 (**Table 2**), we observed abnormal expression of various circRNA-miRNA pairs, some of which were previously found to be associated with cancer. In Axis 1 (**Table 2**), has_circ_0027494↑ let-7c-5p↓, let-7c-5p is found to be a tumor suppressor across many cancer types, including breast cancer (Fu et al., 2017), mucosal melanoma (Tang et al., 2019), acute erythroleukemia (Mortazavi and Sharifi, 2018), among others (Wagner et al., 2014; Liu et al., 2018; Chirshev et al., 2019). We postulate that hsa_circ_0027494 has inhibitory effects on let-7c-5p (Panda, 2018; Zhang P.-F. et al., 2019; Zhang X. et al., 2019), leading to downstream overexpression of let-7c-5p targeted genes, which could explain tumorigenic activities that stands differently in EBV + vs EBV- NPC. Similarly, for Axis 4, 5, 6 (**Table 2**), it is plausible that they regulate certain target genes that underlies traits of EBV+ NPC.

In sum, our method combining DE analysis and co-expression network analysis yielded 6 circRNA-miRNA pairs that show significant DE and co-expression. Out of these results, we were able to establish the *hsa_circ_0008129*↑ *miR-221-3p*↓ *CDKN1B*↑ axis, where overexpression of hsa_circ_0008129 led to the sponging of miR-221-3p and upregulation of *CDKN1B*.

Metastatic vs. Non-metastatic Patient Samples: Differentially Expressed mRNAs

No significant mRNA or circRNA differential expression was detected in the differential expression analysis of metastatic vs. non-metastatic samples. However, we did find 4 miRNAs that were differentially expressed in the metastatic samples than the non-metastatic samples (**Table 3**).

miR-130b-5p was overexpressed in the metastatic samples. This miRNA is likely not involved in pan-cancer metastatic mechanisms, as different cancer types exhibit different expression patterns with contradicting effects. Namely, miR-130b-5p has been shown to promote proliferation and migration in gastric cancer *via* targeting *RASAL1* (Chen H. et al., 2018), as well as in osteosarcoma *via* binding to *TIMP2* (Cheng et al., 2019). On the other hand, miR-130b-5p exhibits anti-tumor effects in pancreatic ductal adenocarcinoma (Fukuhisa et al., 2019) and prostate cancer (Chen et al., 2015). Little literature exists on the functions of miR-449c-5p (**Table 3**), yet given its significant downregulation in the metastatic group (Log2 = -6.9275, p = 1.0903E-19), we postulated that miR-449c-5p could target a metastasis-associated mRNA.

MATERIALS AND METHODS

Data Source

We used four cell lines, including three nasopharyngeal carcinoma (NPC) cell line: SUNE-1, C666 (EBV+), CNE-2, and one non-cancerous human immortalized nasopharyngeal epithelial cell line NP69. Furthermore, we collected samples from four patients diagnosed with NPC in Fujian Cancer Hospital. Two of the four NPC patients were grouped into the metastatic group, and the remaining two patients were grouped into the non-metastatic group.

Statistics

All adjusted p values were reported at false discovery rate of 0.05 unless otherwise specified. All measures of correlation referred to Pearson correlation unless otherwise specified.

Sequencing and Preprocessing

We used Hisat2 to align the RNA reads to the reference genome (Kim et al., 2019). We used StringTie (https://github.com/gpertea/stringtie) (Pertea et al., 2015) to perform RNA transcript assembly. After RNA transcript assembly, we used cuffmerge in the Cufflinks software (Trapnell et al., 2010) to filter the transcript, namely by removing transcripts smaller than 200 nt, or have less than 2 exons, or transcripts with unidentified directions. Then we referenced the remaining transcript assembly using cuffcompare to select for known circRNA, mRNA, and miRNAs. We used StringTie to profile the expression of different RNAs given the filtered transcript data. Information related to QC and mapping can be found in **Supplementary Table S8**.

Identification of circRNAs

We used two software to identify the circRNAs within the transcript assembly: find_circ (Memczak et al., 2013) and CIRI2 (Gao et al., 2015; Gao et al., 2017). Each software produced a set of identified circRNAs, and we took the intersecting set between both to minimize false positive rates. We profiled the expression of circRNAs using CIRI2. We obtained the expression of 15,036 circRNAs, out of which 7,029 were novel circRNAs found in this study. The complete list of all circRNAs identified in this study can be found in **Supplementary Table S9**.

Differential Expression Analysis

We performed DE analysis for miRNA, circRNA, and mRNA separately among different pairs of cell lines using DESeq2 (Love et al., 2014). Namely, we performed DE analysis on each category of comparison: cancer (C666, CNE-2, SUNE-1) vs. non-cancer (NP69); EBV+ (C666) vs. EBV- (CNE-2, SUNE-1); and an additional set of DE analysis using two metastatic vs. two non-metastatic patient samples. The threshold for qualifying as differential expression was set to be edgeR adjusted p value $\leqq 0.05$ and $|\log2FoldChange| \geq 1.0$.

Network Construction

To investigate the role of ceRNA regulation in NPC, we constructed a co-expression network representing the interactions between miRNA-circRNA pairs and miRNA-mRNA pairs. We first calculated Pearson correlation between all possible miRNA-circRNA and miRNA-mRNA pairs, and keep only pairs with correlation coefficient < -0.85. The specific cutoff was chosen to capture the biological mechanism where circRNA inhibits miRNA and miRNA inhibits mRNA. Further, we used the miRanda software (Betel et al., 2010) to predict miRNA-circRNA and miRNA-mRNA target interactions, and we kept only pairs that were calculated to be valid targets by the default parameters of miRanda. Due to the limited number of cell lines for co-expression calculation, and potential flaws in the miRanda software, we further curate the list of miRNA-mRNA pairs by adding those with target score≥97 in the miRDB database (Chen and Wang, 2020). The target score are assigned by the target prediction algorithm in miRDB and the higher the target score is, the more confidence the prediction has (Chen and Wang, 2020). It ranged from 50 to 100. The target scores of miRNA-mRNA pairs in **Table 1**; **Table 2** were given in **Supplementary Table S10**.

Identifying Significant circRNA-miRNA-mRNA Axes

In order to identify functionally relevant axes of regulation between circRNA-miRNA-mRNA, we applied additional filters specifying negative correlations. Namely, we selected circRNA-miRNA-mRNA axes in the network satisfying: 1) contained at least 2 differentially expressed components; 2) circRNA-miRNA and miRNA-mRNA pairs within the axis exhibited opposite differential expression. The motivation behind such criteria was that given a functionally relevant gene with abnormal expression, the upstream circRNAs and miRNAs, if indeed responsible for the abnormal expression, should have opposing directions of DE. Namely, miRNAs inhibit expression of target gene using Dicer and Drosha protein complexes, and circRNAs inhibit target miRNA through the miRNA sponge mechanism. Therefore, within a relevant axis, miRNAs should be negatively correlated to circRNAs and mRNAs.

DE of Public Datasets

GSE143797 profiled the expression of four NPC tissue and matched healthy tissue, identifying 93 upregulated circRNAs and 77 downregulated circRNAs (Yang et al., 2020). GSE118721 profiled the mRNA and miRNA expression of six EBV-positive NPC biopsy specimens and normal nasopharyngeal samples (Lin et al., 2018). We performed a Two-way Student T test to identify differentially expressed coding and non-coding RNAs both datasets, and calculated Log2 Fold Changes.

Kaplan-Meier Plots of Genes of Interest

The GEPIA portal (Tang et al., 2017) (gepia.cancer-pku.cn) was used to plot Kaplan-Meier curves for genes of interest. Appropriate TCGA datasets are used, and the cutoff for high and low expression is the top and bottom 25th percentile. Overall survival was plotted alongside Hazard Ratio and the 95% confidence interval.

CONCLUSION

Our study compared the expression profile between NPC vs. non-cancerous human immortalized nasopharyngeal epithelial cell lines, EBV+ vs. EBV- NPC cell lines, and metastatic vs. non-metastatic patient samples. Specifically, we constructed a circRNA-miRNA-mRNA co-expression network, and performed differential expression analysis. This allowed us to filter for highly correlated non-coding RNA axis that show significant differential expression, suggesting that they are both biologically linked and significant in the genetic landscape of NPC. We found multiple circRNA-miRNA-mRNA axes relevant in both NPC and more specifically in EBV+ NPC. We highlighted the regulatory ceRNA axes behind key differentially expressed genes including *CDKN1B, ZNF302, ZNF268, TAB3*, *SAR1B*, *BRMS1L, ZIC2*, and *MBNL2*. We believe our in-silico findings illustrated the regulatory role that ceRNAs play in NPC, and these results should be further studied using experimental techniques.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Fujian Medical University Cancer Hospital and Fujian Cancer Hospital (Fuzhou, China). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XL, YS, TH designed the experiments, KL, JZ, QZ performed the experiments, SW did the data analysis, XL, SW, and TH wrote the manuscript.

FUNDING

This study was supported by the Natural Science Foundation of Fujian Province (Nos. 2019J01196, 2020J011109), Medical Innovation Program of Fujian Province (No. 2019-CX-5) and National Natural Science Foundation Project of China (No. 81972717), Fujian Provincial Clinical Research Center for Cancer Radiotherapy and Immunotherapy(No. 2020Y2012), Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200, XDA26040304), National Key R&D Program of China (2018YFC0910403, 2017YFC1201200) and Shanghai Municipal Science and Technology Major Project (2017SHZDZX01).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.782473/full#supplementary-material

REFERENCES

Bartel, D. P. (2004). MicroRNAs. *Cell* 116 (2), 281–297. doi:10.1016/S0092-8674(04)00045-5

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive Modeling of microRNA Targets Predicts Functional Non-Conserved and Non-Canonical Sites. *Genome Biol.* 11 (8), R90. doi:10.1186/gb-2010-11-8-r90

Boudhraa, Z., Carmona, E., Provencher, D., and Mes-Masson, A.-M. (2020). Ran GTPase: A Key Player in Tumor Progression and Metastasis. *Front. Cel Dev. Biol.* 8, 345. doi:10.3389/fcell.2020.00345

Buell, P. (1974). The Effect of Migration on the Risk of Nasopharyngeal Cancer Among Chinese. *Cancer Res.* 34 (5), 1189–1191.

Chan, A. T. C., Hui, E. P., Ngan, R. K. C., Tung, S. Y., Cheng, A. C. K., Ng, W. T., et al. (2018). Analysis of Plasma Epstein-Barr Virus DNA in Nasopharyngeal Cancer after Chemoradiation to Identify High-Risk Patients for Adjuvant Chemotherapy: A Randomized Controlled Trial. *J. Clin. Oncol.* 36 (31), 3091–3100. doi:10.1200/JCO.2018.77.7847

Chen, Y., and Wang, X. (2020). miRDB: an Online Database for Prediction of Functional microRNA Targets. *Nucleic Acids Res.* 48 (D1), D127–D131. doi:10.1093/nar/gkz757

Chen, Q., Zhao, X., Zhang, H., Yuan, H., Zhu, M., Sun, Q., et al. (2015). MiR-130b Suppresses Prostate Cancer Metastasis through Down-Regulation of MMP2. *Mol. Carcinog.* 54 (11), 1292–1300. doi:10.1002/mc.22204

Chen, J., Gu, J., Feng, J., Liu, Y., Xue, Q., Ni, T., et al. (2016). TAB3 Overexpression Promotes Cell Proliferation in Non-Small Cell Lung Cancer and Mediates Chemoresistance to CDDP in A549 Cells via the NF-Kb Pathway. *Tumor Biol.* 37 (3), 3851–3861. doi:10.1007/s13277-015-3896-y

Chen, D., Si, W., Shen, J., Du, C., Lou, W., Bao, C., et al. (2018a). miR-27b-3p Inhibits Proliferation and Potentially Reverses Multi-Chemoresistance by Targeting CBLB/GRB2 in Breast Cancer Cells. *Cell Death Dis* 9 (2), 188. doi:10.1038/s41419-017-0211-4

Chen, H., Yang, Y., Wang, J., Shen, D., Zhao, J., and Yu, Q. (2018b). miR-130b-5p Promotes Proliferation, Migration and Invasion of Gastric Cancer Cells via Targeting RASAL1. *Oncol. Lett.* 15 (5), 6361–6367. doi:10.3892/ol.2018.8174

Chen, N., Zhao, G., Yan, X., Lv, Z., Yin, H., Zhang, S., et al. (2018c). A Novel FLI1 Exonic Circular RNA Promotes Metastasis in Breast Cancer by Coordinately Regulating TET1 and DNMT1. *Genome Biol.* 19 (1), 218. doi:10.1186/s13059-018-1594-y

Chen, L., Zhou, H., and Guan, Z. (2019). CircRNA_000543 Knockdown Sensitizes Nasopharyngeal Carcinoma to Irradiation by Targeting miR-9/platelet-Derived Growth Factor Receptor B Axis. *Biochem. Biophysical Res. Commun.* 512 (4), 786–792. doi:10.1016/j.bbrc.2019.03.126

Chen, Z., Pan, T., Jiang, D., Jin, L., Geng, Y., Feng, X., et al. (2020). The lncRNA-GAS5/miR-221-3p/DKK2 Axis Modulates ABCB1-Mediated Adriamycin Resistance of Breast Cancer via the Wnt/β-Catenin Signaling Pathway. *Mol. Ther. - Nucleic Acids* 19, 1434–1448. doi:10.1016/j.omtn.2020.01.030

Cheng, Z. H., Luo, C., and Guo, Z. L. (2019). MicroRNA-130b-5p Accelerates the Migration and Invasion of Osteosarcoma via Binding to TIMP2. *Eur. Rev. Med. Pharmacol. Sci.* 23 (21), 9267–9276. doi:10.26355/eurrev_201911_19419

Chirshev, E., Oberg, K. C., Ioffe, Y. J., and Unternaehrer, J. J. (2019). Let-7as Biomarker, Prognostic Indicator, and Therapy for Precision Medicine in Cancer. *Clin. Translational Med.* 8 (1), 24. doi:10.1186/s40169-019-0240-y

Clayton, N. S., and Ridley, A. J. (2020). Targeting Rho GTPase Signaling Networks in Cancer. *Front. Cel Dev. Biol.* 8, 222. doi:10.3389/fcell.2020.00222

Conn, V. M., Hugouvieux, V., Nayak, A., Conos, S. A., Capovilla, G., Cildir, G., et al. (2017). A circRNA from SEPALLATA3 Regulates Splicing of its Cognate mRNA through R-Loop Formation. *Nat. Plants* 3 (5), 17053. doi:10.1038/nplants.2017.53

Fernández-Medarde, A., and Santos, E. (2011). Ras in Cancer and Developmental Diseases. *Genes Cancer* 2 (3), 344–358. doi:10.1177/1947601911411084

Fero, M. L., Randel, E., Gurley, K. E., Roberts, J. M., and Kemp, C. J. (1998). The Murine Gene p27Kip1 Is Haplo-Insufficient for Tumour Suppression. *Nature* 396 (6707), 177–180. doi:10.1038/24179

Fornari, F., Gramantieri, L., Ferracin, M., Veronese, A., Sabbioni, S., Calin, G. A., et al. (2008). MiR-221 Controls CDKN1C/p57 and CDKN1B/p27 Expression in Human Hepatocellular Carcinoma. *Oncogene* 27 (43), 5651–5661. doi:10.1038/onc.2008.178

Fu, X., Mao, X., Wang, Y., Ding, X., and Li, Y. (2017). Let-7c-5p Inhibits Cell Proliferation and Induces Cell Apoptosis by Targeting ERCC6 in Breast Cancer. *Oncol. Rep.* 38 (3), 1851–1856. doi:10.3892/or.2017.5839

Fukuhisa, H., Seki, N., Idichi, T., Kurahara, H., Yamada, Y., Toda, H., et al. (2019). Gene Regulation by Antitumor miR-130b-5p in Pancreatic Ductal Adenocarcinoma: the Clinical Significance of Oncogenic EPS8. *J. Hum. Genet.* 64 (6), 521–534. doi:10.1038/s10038-019-0584-6

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: An Efficient and Unbiased Algorithm for De Novo Circular RNA Identification. *Genome Biol.* 16 (1), 4. doi:10.1186/s13059-014-0571-3

Gao, Y., Zhang, J., and Zhao, F. (2017). Circular RNA Identification Based on Multiple Seed Matching. *Brief. Bioinform.* 19 (5), 803–810. doi:10.1093/bib/bbx014

Gong, C., Qu, S., Lv, X.-B., Liu, B., Tan, W., Nie, Y., et al. (2014). BRMS1L Suppresses Breast Cancer Metastasis by Inducing Epigenetic Silence of FZD10. *Nat. Commun.* 5 (1), 5406. doi:10.1038/ncomms6406

Guo, Y., Yang, J., Huang, Q., Hsueh, C., Zheng, J., Wu, C., et al. (2019). Circular RNAs and Their Roles in Head and Neck Cancers. *Mol. Cancer* 18 (1), 44. doi:10.1186/s12943-019-1003-5

Hirata, H., Hinoda, Y., Shahryari, V., Deng, G., Tanaka, Y., Tabatabai, Z. L., et al. (2014). Genistein Downregulates Onco-miR-1260b and Upregulates sFRP1 and Smad4 via Demethylation and Histone Modification in Prostate Cancer Cells. *Br. J. Cancer* 110 (6), 1645–1654. doi:10.1038/bjc.2014.48

Hirono, T., Jingushi, K., Nagata, T., Sato, M., Minami, K., Aoki, M., et al. (2019). MicroRNA-130b Functions as an oncomiRNA in Non-small Cell Lung Cancer by Targeting Tissue Inhibitor of Metalloproteinase-2. *Sci. Rep.* 9 (1), 6956. doi:10.1038/s41598-019-43355-8

Hoey, C., Ahmed, M., Fotouhi Ghiam, A., Vesprini, D., Huang, X., Commisso, K., et al. (2019). Circulating miRNAs as Non-invasive Biomarkers to Predict Aggressive Prostate Cancer after Radical Prostatectomy. *J. Transl Med.* 17 (1), 173. doi:10.1186/s12967-019-1920-5

Hong, B. S., Ryu, H. S., Kim, N., Kim, J., Lee, E., Moon, H., et al. (2019). Tumor Suppressor miRNA-204-5p Regulates Growth, Metastasis, and Immune Microenvironment Remodeling in Breast Cancer. *Cancer Res.* 79 (7), 1520–1534. doi:10.1158/0008-5472.CAN-18-0891

Hu, L., Wang, W., Cai, J., Luo, J., Huang, Y., Xiong, S., et al. (2013). Aberrant Expression of ZNF268 Alters the Growth and Migration of Ovarian Cancer Cells. *Oncol. Lett.* 6 (1), 49–54. doi:10.3892/ol.2013.1318

Huang, H.-Y., Lin, Y.-C. -D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2020). miRTarBase 2020: Updates to the Experimentally Validated microRNA-Target Interaction Database. *Nucleic Acids Res.* 48 (D1), D148–D154. doi:10.1093/nar/gkz896

Jain, A., Chia, W. K., and Toh, H. C. (20162016). Immunotherapy for Nasopharyngeal Cancer—A Review. *Chin. Clin. Oncol.* 5 (2), 10. doi:10.21037/cco.2016.03.08

Ji, L., Jiang, X., Mao, F., Tang, Z., and Zhong, B. (2019). miR-589-5p Is Downregulated in Prostate Cancer and Regulates Tumor Cell Viability and Metastasis by Targeting CCL-5. *Mol. Med. Rep.* 20 (2), 1373–1382. doi:10.3892/mmr.2019.10334

Kern, F., Fehlmann, T., Solomon, J., Schwed, L., Grammes, N., Backes, C., et al. (2020). miEAA 2.0: Integrating Multi-Species microRNA Enrichment Analysis and Workflow Management Systems. *Nucleic Acids Res.* 48 (W1), W521–W528. doi:10.1093/nar/gkaa309

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype. *Nat. Biotechnol.* 37 (8), 907–915. doi:10.1038/s41587-019-0201-4

Lee, Y.-H., Jhuang, Y.-L., Chen, Y.-L., Jeng, Y.-M., and Yuan, R.-H. (2016). Paradoxical Overexpression of MBNL2 in Hepatocellular Carcinoma Inhibits Tumor Growth and Invasion. *Oncotarget* 7 (40), 65589–65601. doi:10.18632/oncotarget.11577

Li, Y. Q., Khin, N. S., and Chua, M. L. K. (2018). The Evolution of Epstein-Barr Virus Detection in Nasopharyngeal Carcinoma. *Cancer Biol. Med.* 15 (1), 1–5. doi:10.20892/j.issn.2095-3941.2017.0176

Lin, C., Zong, J., Lin, W., Wang, M., Xu, Y., Zhou, R., et al. (2018). EBV-miR-BART8-3p Induces Epithelial-Mesenchymal Transition and Promotes Metastasis of Nasopharyngeal Carcinoma Cells through Activating NF-Kb and Erk1/2 Pathways. *J. Exp. Clin. Cancer Res.* 37 (1), 283. doi:10.1186/s13046-018-0953-6

Liu, G.-X., Ma, S., Li, Y., Yu, Y., Zhou, Y.-X., Lu, Y.-D., et al. (2018). Hsa-let-7c Controls the Committed Differentiation of IGF-1-Treated Mesenchymal Stem Cells Derived from Dental Pulps by Targeting IGF-1R via the MAPK Pathways. *Exp. Mol. Med.* 50 (4), 1–14. doi:10.1038/s12276-018-0048-7

Liu, Q., Zhang, W., Wu, Z., Liu, H., Hu, H., Shi, H., et al. (2020). Construction of a Circular RNA-microRNA-MessengerRNA Regulatory Network in Stomach Adenocarcinoma. *J. Cel Biochem* 121 (2), 1317–1331. doi:10.1002/jcb.29368

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Luo, C., Yuan, R., Chen, L., Zhou, W., Shen, W., Qiu, Y., et al. (2017). TAB3 Upregulates Survivin Expression to Promote Colorectal Cancer Invasion and Metastasis by Binding to the TAK1-TRAF6 Complex. *Oncotarget* 8 (63), 106565–106576. doi:10.18632/oncotarget.22497

Lv, B., Li, F., Liu, X., and Lin, L. (2021). The Tumor-Suppressive Role of microRNA-873 in Nasopharyngeal Carcinoma Correlates with Downregulation of ZIC2 and Inhibition of AKT Signaling Pathway. *Cancer Gene Ther.* 28 (1), 74–88. doi:10.1038/s41417-020-0185-8

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs Are a Large Class of Animal RNAs with Regulatory Potency. *Nature* 495 (7441), 333–338. doi:10.1038/nature11928

Milani, R., Brognara, E., Fabbri, E., Manicardi, A., Corradini, R., Finotti, A., et al. (2019). Targeting miR-155-5p and miR-221-3p by Peptide Nucleic Acids Induces Caspase-3 Activation and Apoptosis in Temozolomide-Resistant T98G Glioma Cells. *Int. J. Oncol.* 55 (1), 59–68. doi:10.3892/ijo.2019.4810

Mortazavi, D., and Sharifi, M. (2018). Antiproliferative Effect of Upregulation of Hsa-Let-7c-5p in Human Acute Erythroleukemia Cells. *Cytotechnology* 70 (6), 1509–1518. doi:10.1007/s10616-018-0241-5

Panda, A. C. (2018). Circular RNAs Act as miRNA Sponges. *Adv. Exp. Med. Biol.* 1087, 67–79. doi:10.1007/978-981-13-1426-1_6

Peng, Y., and Croce, C. M. (2016). The Role of MicroRNAs in Human Cancer. *Sig Transduct Target. Ther.* 1 (1), 15004. doi:10.1038/sigtrans.2015.4

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads. *Nat. Biotechnol.* 33 (3), 290–295. doi:10.1038/nbt.3122

Prieto-Dominguez, N., Parnell, C., and Teng, Y. (2019). Drugging the Small GTPase Pathways in Cancer Treatment: Promises and Challenges. *Cells* 8 (3), 255. doi:10.3390/cells8030255

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* 146 (3), 353–358. doi:10.1016/j.cell.2011.07.014

Sherr, C. J., and Roberts, J. M. (1999). CDK Inhibitors: Positive and Negative Regulators of G1-phase Progression. *Genes Develop.* 13 (12), 1501–1512. doi:10.1101/gad.13.12.1501

Shotelersuk, K., Khorprasert, C., Sakdikul, S., Pornthanakasem, W., Voravud, N., and Mutirangura, A. (2000). Epstein-Barr Virus DNA in Serum/Plasma as a Tumor Marker for Nasopharyngeal Cancer. *Clin. Cancer Res.* 6 (3), 1046–1051.

Sticht, C., De La Torre, C., Parveen, A., and Gretz, N. (2018). miRWalk: An Online Resource for Prediction of microRNA Binding Sites. *PLOS ONE* 13 (10), e0206239. doi:10.1371/journal.pone.0206239

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: A Web Server for Cancer and normal Gene Expression Profiling and Interactive Analyses. *Nucleic Acids Res.* 45 (W1), W98–w102. doi:10.1093/nar/gkx247

Tang, H., Ma, M., Dai, J., Cui, C., Si, L., Sheng, X., et al. (2019). miR-let-7b and miR-Let-7c Suppress Tumourigenesis of Human Mucosal Melanoma and Enhance the Sensitivity to Chemotherapy. *J. Exp. Clin. Cancer Res.* 38 (1), 212. doi:10.1186/s13046-019-1190-3

Tao, T., He, Z., Shao, Z., and Lu, H. (2016). TAB3 O-GlcNAcylation Promotes Metastasis of Triple Negative Breast Cancer. *Oncotarget* 7 (16), 22807–22818. doi:10.18632/oncotarget.8182

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation. *Nat. Biotechnol.* 28 (5), 511–515. doi:10.1038/nbt.1621

Tsao, S. W., Tsang, C. M., and Lo, K. W. (2017). Epstein-Barr Virus Infection and Nasopharyngeal Carcinoma. *Phil. Trans. R. Soc. B* 372 (1732), 20160270. doi:10.1098/rstb.2016.0270

Wagner, S., Ngezahayo, A., Murua Escobar, H., and Nolte, I. (2014). Role of miRNALet-7and its Major Targets in Prostate Cancer. *Biomed. Res. Int.* 2014, 376326. doi:10.1155/2014/376326

Wang, W., Guo, M., Hu, L., Cai, J., Zeng, Y., Luo, J., et al. (2012). The Zinc Finger Protein ZNF268 Is Overexpressed in Human Cervical Cancer and Contributes to Tumorigenesis via Enhancing NF-Kb Signaling. *J. Biol. Chem.* 287 (51), 42856–42866. doi:10.1074/jbc.M112.399923

Wang, X., Liao, X., Huang, K., Zeng, X., Liu, Z., Zhou, X., et al. (2019). Clustered microRNAs Hsa-miR-221-3p/hsa-miR-222-3p and Their Targeted Genes Might Be Prognostic Predictors for Hepatocellular Carcinoma. *J. Cancer* 10 (11), 2520–2533. doi:10.7150/jca.29207

Wang, X., Wu, T., Wang, P., Yang, L., Li, Q., Wang, J., et al. (2020). Circular RNA 103862 Promotes Proliferation and Invasion of Laryngeal Squamous Cell Carcinoma Cells through the miR-493-5p/GOLM1 Axis. *Front. Oncol.* 10, 1064. doi:10.3389/fonc.2020.01064

Wu, L., Li, C., and Pan, L. (2018). Nasopharyngeal Carcinoma: A Review of Current Updates. *Exp. Ther. Med.* 15 (4), 3687–3692. doi:10.3892/etm.2018.5878

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *The Innovation* 2 (3), 100141. doi:10.1016/j.xinn.2021.100141

Xia, Y., Wei, K., Yang, F.-M., Hu, L.-Q., Pan, C.-F., Pan, X.-L., et al. (2019). miR-1260b, Mediated by YY1, Activates KIT Signaling by Targeting SOCS6 to Regulate Cell Proliferation and Apoptosis in NSCLC. *Cel Death Dis* 10 (2), 112. doi:10.1038/s41419-019-1390-y

Xu, L., Li, L., Li, J., Li, H., Shen, Q., Ping, J., et al. (2015). Overexpression of miR-1260b in Non-Small Cell Lung Cancer Is Associated with Lymph Node Metastasis. *Aging Dis.* 6 (6), 478–485. doi:10.14336/ad.2015.0620

Xu, M., Wang, Y., He, H. T., and Yang, Q. (2018). MiR-589-5p Is a Potential Prognostic Marker of Hepatocellular Carcinoma and Regulates Tumor Cell Growth by Targeting MIG-6. *Neoplasma* 65 (5), 753–761. doi:10.4149/neo_2018_171125N762

Xu, X., Zhang, J., Tian, Y., Gao, Y., Dong, X., Chen, W., et al. (2020). CircRNA Inhibits DNA Damage Repair by Interacting with Host Gene. *Mol. Cancer* 19 (1), 128. doi:10.1186/s12943-020-01246-x

Yang, J., Gong, Y., Jiang, Q., Liu, L., Li, S., Zhou, Q., et al. (2020). Circular RNA Expression Profiles in Nasopharyngeal Carcinoma by Sequence Analysis. *Front. Oncol.* 10, 601. doi:10.3389/fonc.2020.00601

Yin, G., Zhang, B., and Li, J. (2019). miR-221-3p Promotes the Cell Growth of Non-Small Cell Lung Cancer by Targeting P-27. *Mol. Med. Rep.* 20 (1), 604–612. doi:10.3892/mmr.2019.10291

You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., et al. (2015). Neural Circular RNAs Are Derived from Synaptic Genes and Regulated by Development and Plasticity. *Nat. Neurosci.* 18 (4), 603–610. doi:10.1038/nn.3975

Zhang, Y., Zhang, X.-O., Chen, T., Xiang, J.-F., Yin, Q.-F., Xing, Y.-H., et al. (2013). Circular Intronic Long Noncoding RNAs. *Mol. Cel* 51 (6), 792–806. doi:10.1016/j.molcel.2013.08.017

Zhang, X., Jiang, P., Shuai, L., Chen, K., Li, Z., Zhang, Y., et al. (2016). miR-589-5p Inhibits MAP3K8 and Suppresses CD90+ Cancer Stem Cells in Hepatocellular Carcinoma. *J. Exp. Clin. Cancer Res.* 35 (1), 176. doi:10.1186/s13046-016-0452-6

Zhang, J., Zheng, Z., Wu, M., Zhang, L., Wang, J., Fu, W., et al. (2019a). The Natural Compound Neobractatin Inhibits Tumor Metastasis by Upregulating the RNA-Binding-Protein MBNL2. *Cel Death Dis* 10 (8), 554. doi:10.1038/s41419-019-1789-5

Zhang, P.-F., Wei, C.-Y., Huang, X.-Y., Peng, R., Yang, X., Lu, J.-C., et al. (2019b). Circular RNA circTRIM33-12 Acts as the Sponge of MicroRNA-191 to Suppress Hepatocellular Carcinoma Progression. *Mol. Cancer* 18 (1), 105. doi:10.1186/s12943-019-1031-1

Zhang, X., Wang, S., Wang, H., Cao, J., Huang, X., Chen, Z., et al. (2019c). Circular RNA circNRIP1 Acts as a microRNA-149-5p Sponge to Promote Gastric Cancer Progression via the AKT1/mTOR Pathway. *Mol. Cancer* 18 (1), 20. doi:10.1186/s12943-018-0935-5

Zhu, L., Liu, Y., Yang, Y., Mao, X. M., and Yin, Z. D. (2019). CircRNA ZNF609 Promotes Growth and Metastasis of Nasopharyngeal Carcinoma by Competing with microRNA-150-5p. *Eur. Rev. Med. Pharmacol. Sci.* 23 (7), 2817–2826. doi:10.26355/eurrev_201904_17558

Zu, F., Han, H., Sheng, W., Sun, J., Zang, H., Liang, Y., et al. (2020). Identification of a Competing Endogenous RNA axis Related to Gastric Cancer. *aging* 12, 20540–20560. doi:10.18632/aging.103926

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, orclaim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Lin, Wang, Lin, Zong, Zheng, Su and Huang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 12 November 2021
doi: 10.3389/fcell.2021.773101

Therapeutic Response-Based Reclassification of Multiple Tumor Subtypes Reveals Intrinsic Molecular Concordance of Therapy Across Histologically Disparate Cancers

Yue Xu[1†], Jie Zheng[1†], Zhaoqing Cai[1], Wang Li[2], Jens Köhler[3], Yao Dai[1], Xiaojie Cheng[1], Tao Wu[1], Fan Zhang[1,2] and Haiyun Wang[1]**

[1]School of Life Sciences and Technology, Tongji University, Shanghai, China, [2]Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China, [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, United States

OPEN ACCESS

Edited by:
Chuan-Xing Li,
Karolinska Institutet (KI), Sweden

Reviewed by:
Yanhong Zhou,
Central South University, China
Bairong Shen,
Sichuan University, China

***Correspondence:**
Fan Zhang
fzhang@tongji.edu.cn
Haiyun Wang
wanghaiyun@tongji.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 09 September 2021
Accepted: 22 October 2021
Published: 12 November 2021

Citation:
Xu Y, Zheng J, Cai Z, Li W, Köhler J, Dai Y, Cheng X, Wu T, Zhang F and Wang H (2021) Therapeutic Response-Based Reclassification of Multiple Tumor Subtypes Reveals Intrinsic Molecular Concordance of Therapy Across Histologically Disparate Cancers. Front. Cell Dev. Biol. 9:773101. doi: 10.3389/fcell.2021.773101

Cancers that are histologically defined as the same type of cancer often need a distinct therapy based on underlying heterogeneity; likewise, histologically disparate cancers can require similar treatment approaches due to intrinsic similarities. A comprehensive analysis integrated with drug response data and molecular alterations, particularly to reveal therapeutic concordance mechanisms across histologically disparate tumor subtypes, has not yet been fully exploited. In this study, we integrated pharmacological, genomic, and transcriptomic profiling data provided from the Cancer Genome Project (CGP) in a systematic in silico investigation of the pharmacological subtypes of cancers and the intrinsic concordance of molecular mechanisms leading to similar therapeutic responses across histologically disparate tumor subtypes. We further developed a novel approach to redefine cell-to-cell similarity and drug-to-drug similarity from the therapeutic concordance, providing a new point of view to study cancer heterogeneity. This study demonstrates how pharmacological and omics data can be used to systematically classify cancers in terms of response to various compounds and provides us with a purely therapy-oriented perspective to view tumor classifications independent of histology subtypes. The knowledge of pharmacological subtypes of 367 drugs are available via our website (http://www.hywanglab.cn/dtdb/), providing the resources for precision medicine in the perspective of therapeutic response-based re-classification of tumor.

Keywords: tumor classification, OMICS data, pharmacological subtypes, precision medicine, CGP dataset

INTRODUCTION

Traditional tumor classification based on histopathologic diagnosis and the TNM staging system offers highly practical guidance for surgical resection, regional radiotherapy or chemotherapies. However, due to complex intertumor and intratumor heterogeneity, cancer patients with such histological diagnoses often suffer from receiving effective drug treatment and therapeutic resistance (Bedard et al., 2013; Arozarena and Wellbrock, 2019; Ding et al., 2019). One of the key mechanisms behind this significant challenge is the fact that cancers from the same tissue of origin often present

quite different mechanisms of oncogenesis at the molecular level (Bedard et al., 2013; Arozarena and Wellbrock, 2019). Decades of studies have focused on finding molecular subtypes within histopathologically defined tumor types by analyzing large-scale genomic, transcriptomic, proteomic, and epigenomic alterations (Ko et al., 2020). As a result, a variety of molecular signatures have been identified to distinguish intrinsic molecular subtypes associated with patient survival, prognosis and response to different therapeutic modalities (Gay et al., 2021). For example, BRAF mutation in melanoma (Menzer et al., 2019); EGFR-mutant lung adenocarcinomas (Marcoux et al., 2019); luminal A, luminal B, HER2-enriched, basal-like and normal-like subtypes in breast cancer (Prat et al., 2012; Rueda et al., 2019; Waks and Winer, 2019); and four prominent genetic subtypes in Diffuse large B-cell lymphoma (DLBCL), termed MCD based on the co-occurrence of MYD88L265P and CD79B mutations, BN2 based on BCL6 fusions and NOTCH2 mutations, N1 based on NOTCH1 mutations, and EZB based on EZH2 mutations and BCL2 translocations are several well-established molecular subtypes with explicit clinical significance (Schmitz et al., 2018).

With these advances in challenging cancer heterogeneity by identifying subtypes within the same tissues, emerging studies are uncovering facts in the other direction that cancers across disparate tissues of origin can explicitly share common molecular mechanism of oncogenesis (Hoadley et al., 2014; Lu et al., 2019; Petralia et al., 2020). For example, one study based on a large-scale genomic analysis revealed that lung squamous, head and neck, and a subset of bladder cancers shared highly concordant signatures typified by TP53 alterations, TP63 amplifications, and high expression of immune and proliferation pathway genes, implicating that those different cancers perhaps require similar treatment approaches (Hoadley et al., 2014). Another study based on large-scale genomic data found that TP53 and KRAS were mutually exclusive in COAD, READ, and LUAD, but significantly coexisted in PAAD. These observations reveal the feasibility of considering the same treatment strategy in different tumor types (Drilon et al., 2018). Similarly, HER2-targeted therapy may be applied to other cancer types analogous to breast cancer because ERBB2/HER2, which can be amplified in breast cancer, is also mutated and/or amplified in subsets of glioblastoma and gastric, serous endometrial, bladder and lung cancers (Cancer Genome Atlas Research et al., 2013). In recent years, the FDA granted approval to larotrectinib, which had marked and durable antitumor activity in a variety of patients with RTK fusion-positive cancer, regardless of age or tumor tissue type (Drilon et al., 2018). Such examples explicitly illuminate a distinct avenue for reclassifying multiple tumor types independent of histopathologic diagnosis, highlighting that treatment approaches specifically discovered in one disease can be applied to another due to their intrinsic concordance of molecular patterns.

Currently, emerging efforts, such as pan-cancer analysis projects, are being conducted to comprehensively define commonalities and differences across cancer types and tissues of origin (Cancer Genome Atlas Research et al., 2013). Nonetheless, integrative analysis, particularly towards revealing intrinsic therapeutic concordance across histologically disparate tumor subtypes, has not yet been fully exploited. Currently, a large-scale pharmacogenomics study, the Cancer Genome Project (CGP), provides high-throughput genomic, expression information and pharmacological profiling of anticancer drugs across hundreds of cell lines that represent explicit molecular subtypes of histologically defined tumors (Garnett et al., 2012). Hence, we integrated drug response information to pharmacological reclassify tumor subtypes regardless of their tissues of origin. Tumor cells in the same class present similar drug responses, and those in different classes show varied drug responses. Furthermore, by integrating genomic alteration and expression information, we unravelled the intrinsic concordant molecular mechanism associated with the common drug response across histologically disparate cancers. Importantly, this research provides us with a purely therapy-oriented perspective to re-examine tumor classifications independent of histology subtypes.

MATERIALS AND METHODS

Data from a large-scale pharmacogenomics study, the Cancer Genome Project (CGP), was accessible from its website: http://www.cancerrxgene.org. Gene expression, mutation and drug sensitivity data were downloaded. The CGP dataset includes 987 cell lines, genome-wide analysis of mutations, copy number variations and expression profiling, as well as the presence of commonly rearranged cancer genes, and 367 pharmacological profiles (dataset version 2020) (Garnett et al., 2012). For pharmacological assays, the dose range of drugs varied greatly for the different drugs. Generally, about 96% drugs used the minimal tested concentration below 0.07820 uM and the maximal tested concentrated below 20 uM. In these data, the natural logarithm of the Half maximal inhibitory concentration (IC50) value represents the drug sensitivity value. IC50 is the half maximal inhibitory concentration of an anticancer drug, and a lower value means higher sensitivity. Cell lines cannot be classified according to histological subtypes provided by TCGA, labelled as "UNCLASSIFIED", were excluded from our study.

Generating the Pharmacological Subtypes Tree

The cell lines, IC50 of which are smaller than maximal tested concentrations, are defined as sensitive, otherwise resistant. Our method iteratively splits the cancer cells into two groups in a way that gains the best separation of drug sensitivity between two groups until reaching two terminal conditions: The p value of the difference of two groups' drug sensitivity values is smaller than 0.05 or all cancer cells in a node are sensitive or resistant.

Suppose the set of all cells is S_{total}, and each cell has its drug sensitivity value. The procedure to grow a pharmacological subtype tree is as follows:

1) Define the root node. Set S as the set of cancer cells in this node, and assign S_{total} to S.

2) Define the root node as the current node.
3) N denotes the number of cells in the current node. If N is smaller than 6 or all cells in such node are sensitive or resistant, finish the procedure of growing the tree. Otherwise, a heuristic algorithm splits S into two child nodes, the left node and the right node, which achieves the greatest drug sensitivity difference.
4) If no way can be found to split the current node ($p < 0.05$), finish the procedure of growing the tree. Otherwise, the current node is split into two child nodes.
5) Define the left node and right node. Suppose the set of cells in the left node is S_l and the set in the right node is S_r. S_l and S_r are the subsets of S.
6) Define the left node as the current node. Set S as the set of cells in this node, and assign S_l to S. Repeat steps 3–5.
7) Define the right node as the current node. Set S as the set of cells in this node, and assign S_r to S. Repeat steps 3–5.

Algorithm for Splitting the Node

We used a heuristic algorithm to search for a reasonable number of divisions to split S into S_l and S_r. Suppose the number of the cells set is ns. We first sorted the cells by IC50 ascending and divided the cells with higher drug sensitivity into the S_l group and the cells with lower drug sensitivity into the S_r group. Since each node at least 6 cells, a total of $ns - 5$ divisions were considered.

For each division, the Mann-Whitney U test was applied to calculate the p value of drug sensitivity between S_l and S_r. The division with the smallest p value was selected as the optimized one to split the cancer cells in the current node.

The Cell Similarity and the Drug Similarity

The similarity of cells is calculated based on their response to the different drugs. For each drug k, a pharmacological subtypes tree S^k is generated, consisting of n subtypes:

$$S^k = \left(S_1^k, S_2^k, \ldots, S_n^k\right) = \left(S_t^k\right)_{t=1\to n}$$

Based on this tree, a matrix that defines a similarity of cancer cells, regarding whether they are in the same subtype and their drug response to drug k, is calculated. We assume there are m cell lines tested for drug k. The cell lines are labelled as sensitive, resistant, or NA (not available). A sensitive cell line has its IC50 value smaller than maximum tested concentration; a resistant cell line has its IC50 value greater than maximal tested concentration. For two given cell lines i and j, their similarity matrix (D_{ij}^k) is defined as:

$$D_{ij}^k = \begin{cases} 1, & if\ (i,j) \in S_t^k\ and\ both\ are\ labelled\ as\ sensitive\ ones; \\ -1, & if\ (i,j) \in S_t^k\ and\ both\ are\ labelled\ as\ resistant\ ones; \\ 0, & otherwise. \end{cases}$$

If two cell lines i and j are in the same subtype and both are sensitive to drug k, their score is 1, and −1 for both cell lines with a resistant response, while if they are not in the same subtype or show opposite response to the drug, their score is 0. A $m \times m$ similarity matrix of cell lines is then generated.

Then, the similarity of cell lines i and j can be further calculated as:

$$Cell_{sim} = \frac{\sum_{k=1}^{q}\left|D_{ij}^k\right|}{q}, \quad i \neq j$$

Wherein i and j denote the cell lines tested for q drugs.

Furthermore, for two given drugs a and b, their similarity is calculated as the summarized similarity across the same cell lines which are tested for both drugs.

$$Drug_{sim} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} D_{ij}^a \cdot D_{ij}^b}{\sum_{i=1}^{m}\sum_{j=1}^{m} D_{ij}^a \cdot D_{ij}^a + \sum_{i=1}^{m}\sum_{j=1}^{m} D_{ij}^b \cdot D_{ij}^b - \sum_{i=1}^{m}\sum_{j=1}^{m} D_{ij}^a \cdot D_{ij}^b}, \quad i \neq j$$

Wherein i and j denote the cell lines tested for both drugs, and m is the total number of the same cell lines.

Connecting Pharmacological Subtypes With Genomic Alterations

The maximum concentration of the drug was used to distinguish between the resistant pharmacological subtypes and sensitive pharmacological subtypes. The chi-squared test was used to calculate the connection between genomic alterations, including mutation and translocation, and pharmacological subtypes. Genomic alterations were determined with p values corrected with the Benjamini-Hochberg method for controlling the false discovery rate (Benjamini and Hochberg, 1995). Here we used the corrected p value <0.1 as the threshold. We defined two types of connections: sensitive and resistant. If pharmacological subtypes with higher drug sensitivity had more frequent alteration occurrences, we defined this connection as sensitive. If the opposite was true, we defined the connection as resistant.

The connection between mRNA expression and pharmacological subtypes was determined with p values from the student t test corrected with the Benjamini-Hochberg method. If pharmacological subtypes with higher drug sensitivity had higher expression, we defined this connection as sensitive. If the opposite was true, we defined the connection as resistant. Here we used the corrected p value <0.1 as the threshold. To obtain more confident connections, the Pearson correlation coefficient (PCC) was used to measure the correlation between drug sensitivity (IC50) and gene expression in the sensitive subtypes, and only genes with PCC greater than 0.3 were considered. The individual subtypes with sensitive and resistant cells mixed were excluded in the above analyses.

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) (Subramanian et al., 2005) was employed to determine the 333 gene sets from KEGG, enriched by a pre-ranked list of all genes, which were sorted by the statistical significance of differential expression defined by DESeq2 analysis (Love et al., 2014). Gene sets with FDR <0.05 were statistically significant.

Statistical Analysis

The Fisher's exact test was used to respectively determine whether there is a significant association between pathways and clusters of drugs

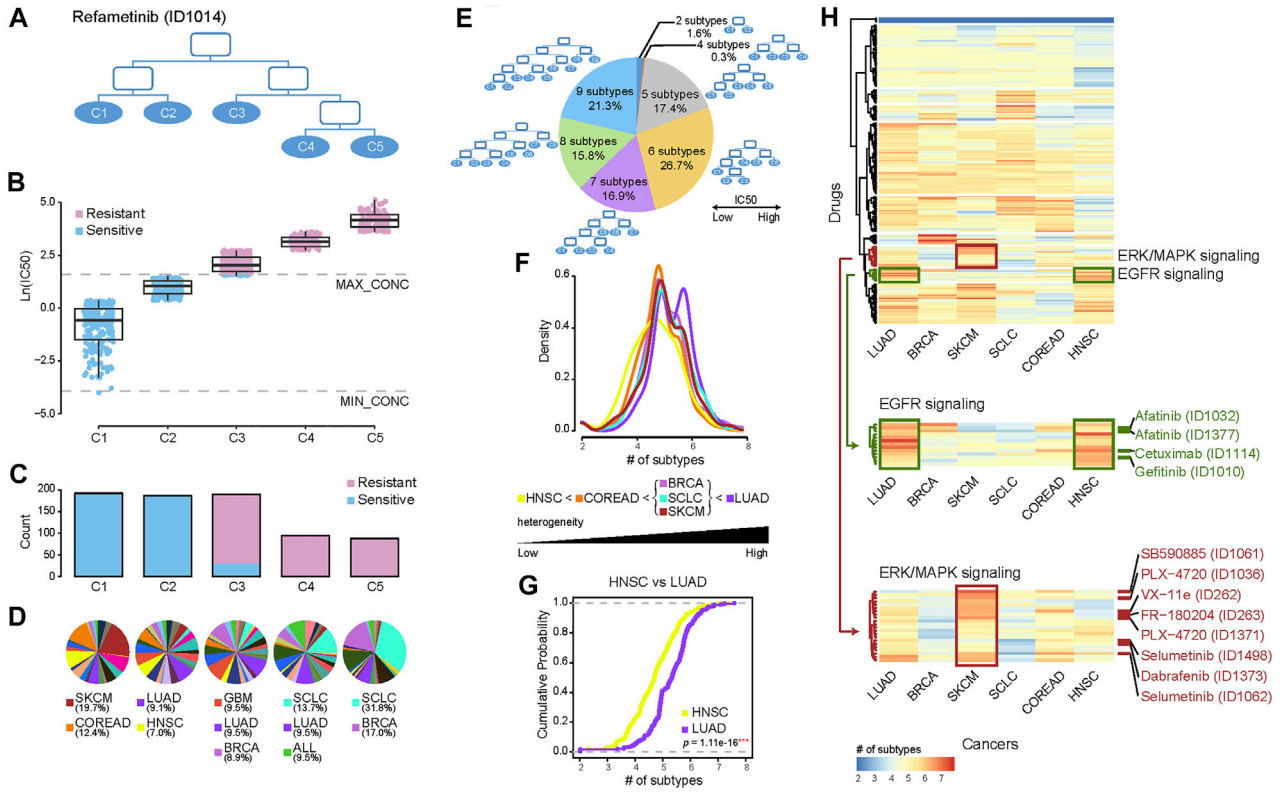


FIGURE 1 | Re-classifying cancers to generate pharmacological subtypes and drug sensitivity profiles. **(A)** Pharmacological subtype tree of the drug Refametinib. Cancer cells were classified into five pharmacological subtypes based on their drug sensitivity to MEK inhibitor Refametinib. **(B)** The boxplot shows the drug sensitivities of five subtypes (leaf nodes): C1, C2, C3, C4, and C5, indicating a gradually increasing degree of resistance to the drug Refametinib. Blue points refer to cell lines sensitive to the drug, while red points resistant. **(C)** The barplot shows the number of cell lines in each subtype. **(D)** The pie chart shows the distribution of histological subtypes across pharmacological subtypes. **(E)** The pie chart illustrates the numerical proportion of different degrees of the furthest divisible pharmacological subtypes. Cancer cells could be maximally classified into up to 9 subtypes for different drugs. Among them, 7–9 subtypes accounted for more than half of drugs. **(F)** The density of the number of pharmacological subtypes indicates the varied treatment heterogeneity across six histological cancers, in which the average number of cell lines among all drugs are over 35. **(G)** The empirical cumulative distribution function visualizes the density of drug subtypes in LUAD vs. HNSC. *p* value was calculated by using the Kolmogorov-Smirnov test (K-S test). **(H)** Hierarchical clustering of the number of pharmacological subtypes across six cancers. Each row in the heatmap represents a drug, and each cell in the heatmap represents the number of subtypes of the cancer in the drug. The red color indicates more subtypes, and the blue color less subtypes. There are more pharmacological subtypes of EGFR inhibitors (green drugs) in LUAD and HNSC, and more subtypes of inhibitors targeting in ERK/MAPK signalling (red drugs) in SKCM.

determined by the number of drug pharmacological subtypes across 6 cancer types or not, whether any difference of phenotypes (cancers/pathways) in clusters of cell lines/drugs derived from the similarity matrix was significant, and whether cancer censor genes are enriched in a list of genes with the most connections to pharmacological subtypes. A hypothesis test based on hypergeometric distribution was used to determine whether a histological cancer is enriched in the most sensitive pharmacological subtype of a drug. The p value <0.05 was regarded as statistically significant. Kolmogorov-Smirnov test (K-S test) was employed to compare cumulative distribution function (CDF) of the number of pharmacological subtypes across the drugs between two histological cancers.

RESULTS

We analysed 367 drugs in the CGP dataset and established their pharmacological subtype trees. The leaf node in each tree represents a pharmacological subtype of cancers, which is composed of cells from the histological disparate tumor subtypes. We further employed the mRNA expression profile and mutation/fusion profile to connect the molecular alterations with pharmacological subtypes, delineating a novel perspective to re-classify the tumor according to therapeutic response dependent of the intrinsic concordant molecular mechanism, by regardless of tissues of origin. The cell similarity and the drug similarity were also redefined based on the pharmacological subtypes.

Pharmacological Subtypes of Cancers

We built a pharmacological tree for each drug based on the divisibility of the drug sensitivity of cancer cells. Taking the tree of the MEK1/2 inhibitor Refametinib as an example (**Figure 1A**), all cells in the root node were first divided into left and right child nodes with relatively high and low sensitivity, respectively, and further cells in these two nodes were capable of being divided into

five final subgroups, C1, C2, C3, C4, and C5. Each subgroup had varied drug sensitivity and reached maximum divisibility, thus representing distinct pharmacological subtypes. C1 denotes the most sensitive and C5 the most resistant subtype to the drug Refametinib. Except for C3 where both sensitive and resistant cells were mixed, the other subgroups contained the homologous sensitive or resistant cells (**Figure 1B**). There were much more sensitive cells in C1 and C2 than resistant cells in C4 and C5 (**Figure 1C**). We then investigated how histological subtypes were distributed in the pharmacological subtypes (**Figure 1D**). C1 was composed of 25 histological subtypes, with top 2 cancers, SKCM (19.7%) and COREAD (12.4%). C5 was composed of 22 histological subtypes, with leading cancers including SCLC (31.8%) and BRCA (17.0%).

Our analysis established 367 pharmacological trees. The number of pharmacological subtypes with distinct drug sensitivities varied from 2 to 9 across 367 trees (**Figure 1E**). This suggested that heterogeneity of treatment effects across histological subtypes existed widely, and one pharmacological subtype presenting concordant drug response indeed comprised histological disparate cells to some extent. A high degree of treatment heterogeneity, meaning more than seven pharmacological subtypes, was observed for 54.0% (16.9% of the drugs had 7; 15.8% of the drugs had 8; and 21.3% of the drugs had 9 pharmacological subtypes) of the drugs. These drugs included Dabrafenib (BRAF inhibitor), Selumetinib (MEK1/2 inhibitor), Erlotinib (EGFR inhibitor), and Alectinib (ALK inhibitor). 44.1% of the drugs had 5 or 6 subtypes, showing a moderate degree of treatment heterogeneity, while 1.9% of the drugs had subtypes lower than 4, showing a low degree of heterogeneity related to drug therapy (**Figure 1E**).

We also particularly examined how pharmacological subtypes were constituted within the same histological tumor subtypes, including LUAD, BRCA, SKCM, SCLC, COREAD, and HNSC, in which the average number of cell lines per drug was more than 35 (**Figures 1F–H**). We built approximately 310 pharmacological trees for each type of the above cancers. The density distribution of the number of pharmacological subtypes indicated the varied treatment heterogeneity across 6 histological subtypes (**Figure 1F**). The shape of the distribution curve in LUAD was characterized by double kurtosis and significantly positively biased, and oppositely the curve in HNSC was negatively biased, suggesting the highest treatment heterogeneity in LUAD and lowest treatment heterogeneity in HNSC. Furthermore, K-S test was applied to evaluate if two different histological subtypes had the same level of treatment heterogeneity by comparing their cumulative distributions. The results showed the cumulative distribution curve of LUAD is significantly different from that of HNSC ($p = 1.11e\text{-}16$, K-S test) (**Figure 1G**). The pairwise comparisons between the curves of any two histological subtypes showed BRCA, SCLC, and SKCM had the statistically similar distributions ($p > 0.05$, K-S test), and other cancers had not ($p < 0.05$, K-S test) (**Supplementary Figure S1**).

To observe the varied treatment heterogeneity of the different histological subtypes in detail, the number of pharmacological subtypes of drugs across 6 histological cancers were shown in the heatmap (**Figure 1H**). Six histological tumors were arranged in columns according to their treatment heterogeneity, from high to low. Interestingly we observed that EGFR inhibitors (Afatinib, Cetuximab, and Gefitinib) targeting in EGFR signalling pathway (green drugs in **Figure 1H**) were clustered together ($p < 0.001$, Fisher's exact test). And in LUAD and HNSC there were more pharmacological subtypes of EGFR inhibitors than other drugs, suggesting high treatment heterogeneity for EGFR inhibitors in LUAD and HNSC. In addition, drugs targeting ERK/MAPK signalling (red drugs in **Figure 1H**) were clustered together ($p < 0.001$, Fisher's exact test). There was high treatment heterogeneity for these drugs in SKCM.

Genomic Alterations Associated With Pharmacological Subtypes

To identify the molecular alterations associated with varied drug sensitivity, we examined whether the genomic alterations, including mutations and translocations, significantly changed across the pharmacological subtypes. The statistical significance was determined by p values calculated with the Chi-squared test and corrected with the Benjamini-Hochberg method for controlling the false discovery rate. If the pharmacological subtypes with higher drug sensitivity had more frequent mutation or translocation events, we defined their association as a sensitive connection. If the opposite was true, we defined their association as resistant. Consequently, genes with specific genomic alterations were associated with the respective drugs. Since one gene could have multiple connected drugs and one drug could have multiple connected genes, these gene-drug connections ultimately composed a network that allowed us to investigate the contribution of either a drug or gene to the holistic understanding of how altered genes connect to drug responses (**Figure 2A**). The hub genes with the most connections to the drugs, including known cancer genes KRAS, RB1, BRAF, TP53, NOTCH2, and other genes LTB, TNFRSF9, DERA et al. were highlighted in the network.

We first investigated the distribution of genes on their connected drugs (**Figure 2B**). Our analysis identified 624 genes whose genomic alterations were connected with at least two drugs' pharmacological subtypes. Most of the genes (70.67%) had fewer than 3 connections with drugs, but 40 (6.41%) genes whose genomic alterations were associated with more than 5 drugs. By ranking genes by their number of connections to drugs, we found that the top ranked genes were enriched in cancer genes (**Figure 2C**; **Supplementary Figure S2**), a catalogue of genes with mutations or fusions that are causally implicated in cancer provided by the COSMIC database (Tate et al., 2019). The top 10 genes connected to more than 12 drugs, 30% of which were remarkably annotated as cancer genes (**Figure 2C**). These cancerous percentages for the top 10 (p = 3.62e-03, Fisher's exact test), top 20 (p = 1.27e-07, Fisher's exact test), and top 30 genes (p = 1.62e-09, Fisher's exact test) were significantly higher than the background when using genome-wide 21,972 genes as a reference (**Figure 2C**). Among the top ranked oncogenes, ABL1_BCR ranked first with connections to 28

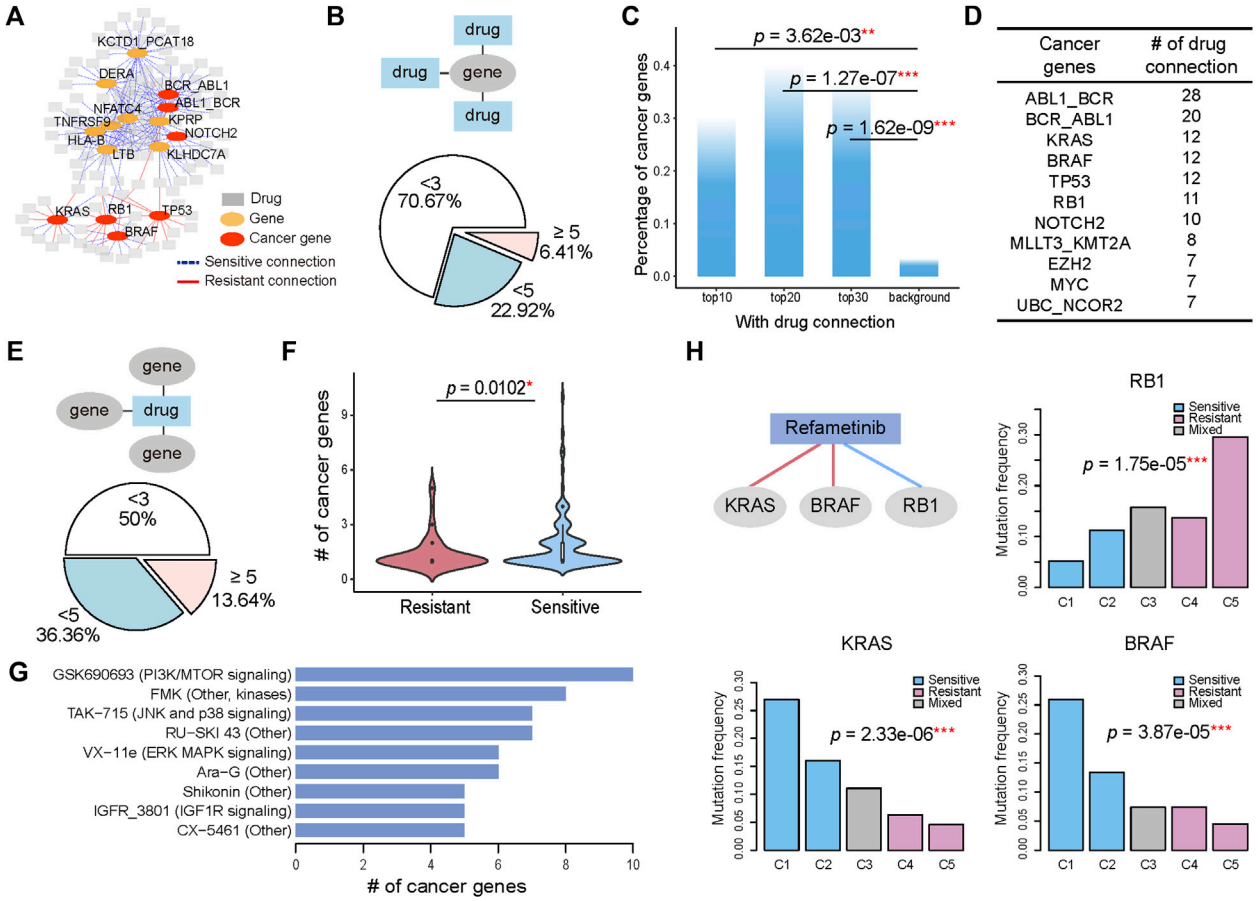


FIGURE 2 | Genomic alterations determine the various drug sensitivities across pharmacological subtypes. **(A)** A gene with its genomic alteration significantly changed across the pharmacological subtypes of a drug is linked with the drug, generating a gene-drug network. A sensitive connection means that subtypes with high drug sensitivity have a more frequent occurrence of genomic alterations. Conversely, those that are less frequent are defined as resistant connections. The 15 hub genes with most connections to drugs are shown. **(B)** 6.41% genes have more than 5 connected drugs. In addition, 70.67% of genes have fewer than 3 connected drugs. **(C)** The percentage of cancer genes in the top 10, 20, and 30 genes, and the genome-wide 21,972 genes are referred to as background. **(D)** The top cancer censor genes defined by COSMIC and the corresponding number of drug connections. **(E)** When considering the drugs connected to the genes, 13.64% of drugs have more than 5 connected cancer genes. In addition, 50% of drugs have fewer than 3 connected cancer genes. **(F)** The number of genes with sensitive connections is significantly greater than the number of genes with resistant connections. **(G)** Drugs with cancer gene connections greater than 5 were ranked. **(H)** BRAF mutation, RB1 mutation and KRAS mutation are associated with the five pharmacological subtypes of Refametinib. Histograms show the mutation ratio of the three genes in the five pharmacological subtypes of Refametinib.

drugs, followed by BCR_ABL1 and KRAS with 20 and 12 drugs (**Figure 2D**), implying that the highly drug-connected genes that are not currently identified as cancer genes could actually be causally implicated in cancer therapy, such as KCTD1_PCAT18 fusion, TNFRSF9, LTB, and NFATC4, which showed connections to over 18 drugs and ranked as the top five genes (**Supplementary Figure S2**).

We then investigated the distribution of drugs on their connected genes. In this analysis, we focused only on cancer censor genes. For the majority of drugs (86.36%), the number of connected genes ranged from 2 to 5 (**Figure 2E**). Among the pharmacological subtypes, genes whose alteration was sensitively connected with drug resistance were significantly more abundant than those that were connected in a resistant manner ($p = 0.0102$, student t test). (**Figure 2F**). We ranked the drugs with more than 5 genes' connection with their pharmacological subtypes (**Figure 2G**). Drugs with gene connections greater than 5 included GSK690693, affecting the PI3K/MTOR signaling pathway, TAK−715, affecting the JNK and p38 signaling pathway, VX−11e, affecting the ERK MAPK signaling pathway, and IGFR_3801, affecting the IGF1R signaling pathway.

Next, we observed how the mutation rate of genes associated with the variability of drug sensitivity changed among drug-sensitive and drug-resistant subtypes. Taking the MEK1/2 inhibitor Refametinib as an example, three genomic alterations were found to connect to the drug and be associated with pharmacological subtypes. As shown in **Figure 2H**, BRAF mutations and KRAS mutations occurred significantly more frequently (over 25%) in the sensitive subtype C1 and then

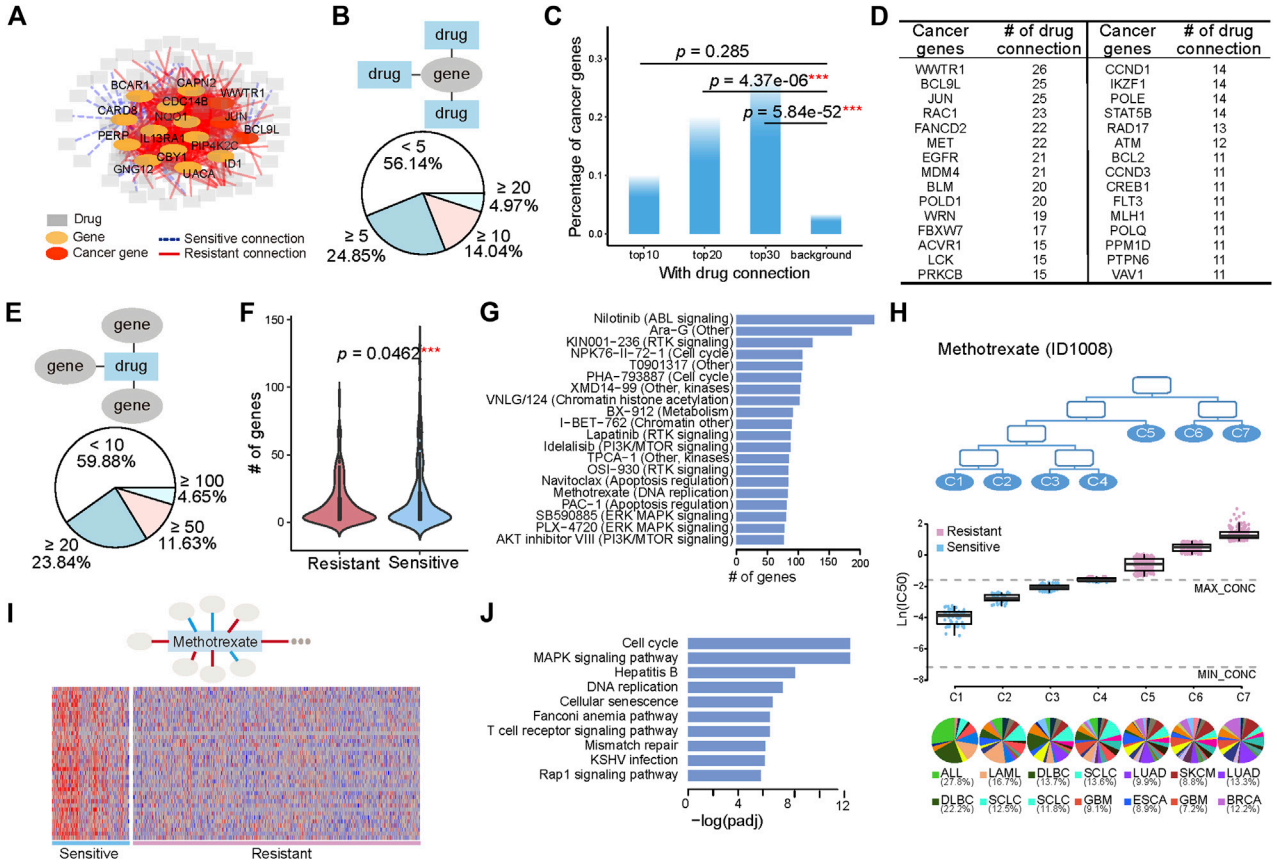


FIGURE 3 | Tissue-specific gene expression determines the various drug sensitivities across pharmacological subtypes. **(A)** A gene with its expression significantly changed across the pharmacological subtypes of a drug is linked with this drug, generating a gene-drug network. A sensitive connection means that subtypes with high drug sensitivity have higher gene expression. Conversely, those with low expression are defined as resistant connections. The 15 hub genes with the most connections to drugs are shown. **(B)** Pie chart illustrates the numerical proportion of the number of connected drugs with genes whose expression is associated with the pharmacological subtypes. A total of 4.97% of genes have more than 20 connected drugs. In addition, 56.14% of genes have less than 5 connected drugs. **(C)** The percentage of cancer genes in the top 10, 20, and 30 genes, and the genome-wide 21,972 genes are referred to as background. **(D)** The top cancer censor genes defined by COSMIC and the corresponding number of drug connections. **(E)** Distribution of the genes functioning in the core cancer pathways whose expression is connected with pharmacological subtypes. A total of 4.65% of drugs have more than 100 connected genes. **(F)** The number of genes with sensitive connections is significantly greater than the number of genes with resistant connections. **(G)** Top 20 drugs with the most connections to genes. **(H)** Seven subtypes, C1, C2, C3, C4, C5, C6, and C7, have varied drug sensitivity to methotrexate, with C1 being the most sensitive to the drug and C7 being the most resistant to the drug. The boxplot below shows the drug sensitivities of the three subtypes (leaf nodes) C1, C2, C3, C4, C5, C6, and C7, indicating a gradually increasing degree of resistance to the drug. **(I)** A total of 38 genes are highly expressed in sensitive group (C1, C2, and C3) and lowly expressed in resistant group (C5, C6, and C7). **(J)** Pathways enriched by the 38 highly expressed genes in sensitive group (C1, C2, and C3).

decreased gradually in the subtypes C2, C3, C4, and C5. Here one subtype C3 with sensitive and resistant cells mixed was excluded in the analysis. Conversely, RB1 mutation, occurred more frequently in the resistant subtypes. Therefore, C1 group, consisting mainly of SKCM (19.7%) and COREAD (12.4%) (**Figure 1D**), was significantly sensitive to Refametinib due to its high occurrence of BRAF and KRAS mutations and low occurrence of RB1 mutations (**Figure 2H**). Other types of MEK inhibitors, including PD0325901, Trametinib, Selumetinib, and CI-1040, further confirmed that BRAF, KRAS, and RB1 mutations were robustly connected to pharmacological subtypes, contributing to the drug sensitivity of MEK inhibitors (**Supplementary Figure S3**). Our analysis revealed that the varied distribution of genomic alterations across pharmacological subtypes could lead to their differences in response to anticancer therapies.

Expression Alterations Associated With Pharmacological Subtypes

We also associated mRNA expression with the pharmacological subtypes. The connection between gene expression and pharmacological subtypes was determined by p values calculated with the student t test and corrected with the Benjamini-Hochberg method for controlling the false discovery rate. If the subtypes with higher drug sensitivity had higher gene expression, we defined this connection as sensitive. If the opposite was true, we defined the connection as resistant. The

connections between genes and drugs constituted a network (**Figure 3A**). To narrow down the genes whose expression levels were directly associated with drug sensitivity, we only kept genes which act in core cancer pathways (**Supplementary Table S1**), and whose Pearson correlation between drug sensitivity and gene expression were greater than 0.3 in the sensitive subtypes. The hub genes with the most connections to the drugs, including known cancer genes WWTR1, JUN, BCL9L and other genes NQO1, CAPN2, and PERP et al., were highlighted in the network.

We first investigated the distribution of genes on their connected drugs (**Figure 3B**). Our analysis identified 684 genes whose expression was connected to at least two drugs' pharmacological subtype. Among them, 4.97%, 14.04%, and 24.85% of genes respectively had connections with greater than 20, 10, and 5 drugs. The top 30 genes connected to more than 20 drugs, 25% of which were annotated as cancer genes (**Figure 3C**; **Supplementary Figure S4**). These cancerous percentages for the top 20 (p = 4.37e-06, Fisher's exact test), and top 30 genes (p = 5.84e-52, Fisher's exact test) were significantly higher than the background. The top genes with their expression alterations associated with subtypes were different from those with genomic alterations associated with subtypes (**Figures 2D**, **3D**). Among them, WWTR1, BCL9L, and JUN ranked the top three with connections to 26 drugs and 25 drugs (**Figure 3D**).

We then investigated the distribution of drugs on their connected genes. The number of genes whose expression was associated with the subtypes varied from 0 to 223 across the drugs, with 4.65% drugs had more than 100 gene connections (**Figure 3E**). Comparing with genomic alterations, there were more genes whose expression was connected to the drugs. Moreover, genes whose expression was sensitively connected with drug resistance across pharmacological subtypes were significantly more abundant than genes that were resistant connected (p = 0.0462, student t test) (**Figure 3F**). The top 20 drugs with the most connections to genes included Nilotinib, affecting the ABL signaling, PHA−793887 and NPK76−II−72−1, affecting cell cycle, Lapatinib and KIN001−236, affecting RTK signaling pathway, Navitoclax, affecting Apoptosis regulation, and some compounds such as Ara−G, XMD14−99 et al., affecting multiple or unknown pathways (**Figure 3G**).

Methotrexate, as an example, is a chemotherapy that specifically acts during DNA and RNA synthesis, and cancer cells were classified into seven subtypes (**Figure 3H**) in terms of drug sensitivity to methotrexate. The first subtype, C1, was most sensitive to the drug, and the seventh subtype, C7, was the most resistant to the drug (**Figure 3H**). C1 included 12 histological subtypes, with the top 2 shown in the bottom of **Figure 3H**, ALL and DLBC. C2 and C3 were composed of 18 and 19 histological subtypes, respectively, with SCLC both included in the top 2 (**Figure 3H**). A total of 38 genes with high expression in sensitive group (C1, C2, and C3) and low expression in resistant group (C5, C6, and C7), which had sensitive connections to pharmacological subtypes of methotrexate, were selected (**Figure 3I**). Pathway enrichment analysis showed that these genes were enriched in pathways including cell cycle and DNA replication et al. (**Figure 3J**), suggesting that the ectopic activation of cell cycle led to the sensitive response of ALL, DLBC, LAML, and SLCL to Methotrexate.

Redefining the Similarity of Cells Based on Pharmacological Subtypes

Pharmacological subtypes provide us a new point of view to redefine the similarity of cells purely from therapeutic concordance. Thus, we calculated the similarity of pairwise cells based on pharmacological subtypes of drugs. If there are quite a few cases that two cells are in the same sensitive or resistant pharmacological subtypes of drugs, these cells gain a high similarity. Otherwise, they gain a low similarity. The similarity of cells defined a hierarchical cluster as shown in **Figure 4A**. We further applied the Fisher's exact test to observe if the cells in the same clusters were from the same histological subtypes (**Figures 4B–D**). The points with black border in the figure represent statistical significance ($p < 0.05$), with the smaller p values shown in red.

The results showed that the cancer cell lines were respectively divided into 2, 4, and 15 clusters at the different hierarchical cluster levels (**Figure 4A**). Two clusters at the high level were reflected by two distinct patterns of similarity (**Figure 4A**). One cluster (Cluster H1) was significantly correlated with solid tumors including BRCA, COREAD, LUAD, MESO, OV, and PAAD, and the other (Cluster H2) hematological tumors including ALL, DLBC, LAML, LCML, and MM (**Figure 4B**). Here, the point with black border represents statistical significance ($p < 0.05$, Fisher's exact test) of the association between the cluster and cancer type. So in a holistic view, the different types of hematological tumors shared a similar drug sensitivity profile, which was quite different from the profile of solid tumors. However, for a given cancer type (the right panel of **Figure 4B**), such as ACC, BLCA, CESC, CLL, or ESCA et al., its cells were unbiasedly distributed in two clusters ($p > 0.05$, Fisher's exact test), showing the treatment heterogeneity for the cells from the same tissue origin. At the median hierarchical levels, 4 clusters were identified (**Figure 4C**). Among them, Cluster M3 was specifically associated with haematological tumors ($p < 0.001$, Fisher's exact test) including ALL, DLBC, LAML, LCML, and MM. The treatment heterogeneity was still observed in cancers.

When cell lines were divided into 15 clusters at the lower hierarchical level (**Figure 4D**), we observed that the cells from ACC, CLL, STAD, or UCEC were unbiasedly distributed in the different clusters, still showing their treatment heterogeneity. Interestingly, the cells from some cancers such as HNSC, KIRC, and SCLC were enriched in more than 2 clusters with distinctly different drug response. For instance, HNSC cells were enriched in Clusters L12, L4 and L5; KIRC cells were enriched in Clusters L15 and L4; SCLC cells were enriched in Clusters L13, L3, and L7. Interestingly, the SCLC cells in Cluster L13 shared a more similar profile with hematological tumors including DLBC, LAML, and LCML. To investigate the underlying mechanism of such treatment heterogeneity in SCLC, we applied DESeq2 analysis to call the differentially expressed genes between

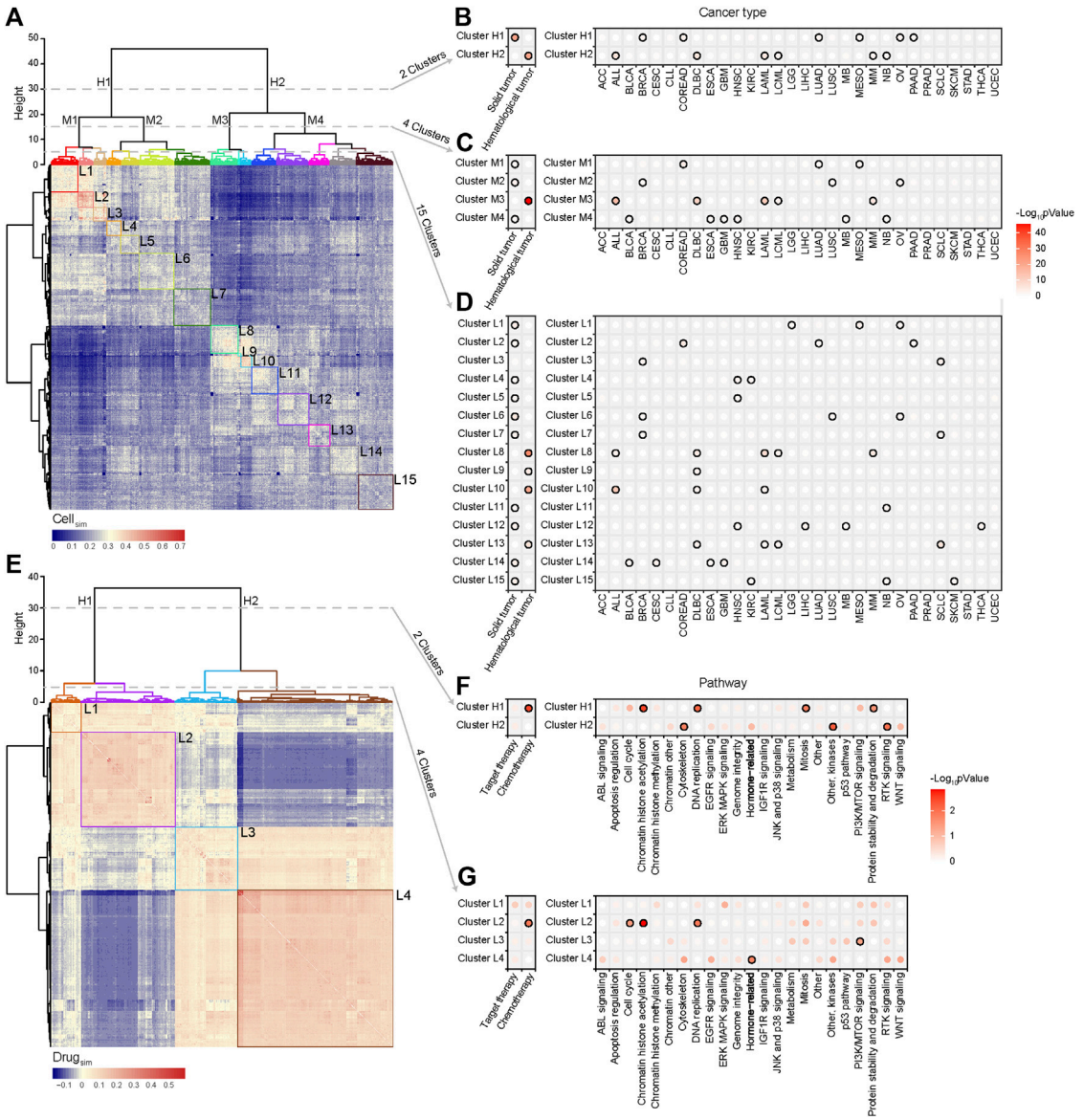


FIGURE 4 | Cell similarity and drug similarity based on pharmacological subtypes. **(A)** Hierarchical clustering of a cell similarity matrix which was generated based on drug response of cells to drugs, with the red color indicating high similarity, and the blue color indicating low similarity between two cell lines. Clusters were formed by partitioning of the dendrogram at a high, median and low height respectively, which indicates the distance between the clusters. **(B)** Two clusters (H1, H2) were formed by cutting the dendrogram at a height of 30. Points with black border represent statistical significance ($p < 0.05$) determined by Fisher's exact test. Almost all histological subtypes of blood cancer were grouped together. **(C)** Four clusters (M1, M2, M3 and M4) were formed by cutting the dendrogram at a height of 15. **(D)** Fifteen clusters (L1-15) were formed by cutting the dendrogram at a height of 5. Cancers with the same tissue of origin and organ did not cluster together. **(E)** Hierarchical clustering of the drug similarity profiles with the red color indicating a high similarity and the blue color indicating a low similarity. Clusters were formed by partitioning of the dendrogram at a high and low height respectively. **(F)** Two clusters (H1, H2) were formed by cutting the dendrogram at a height of 30. Points with black border represents statistical significance ($p < 0.05$) determined by Fisher's exact test. Chemotherapy related drugs were grouped together. **(G)** Four clusters (L1, L2, L3, and L4) were formed by cutting the dendrogram at a height of 5. Cytoskeleton, EGFR signalling, Hormone-related and WNT signalling pathways cluster together.

Cluster L13 and Cluster L3, and further identified the differential functions via gene set enrichment analysis (GSEA) (Subramanian et al., 2005). The results showed that the functions like DNA replication, RNA transport et al. were significantly upregulated in Cluster L13 cells, while in Cluster L3 cells the functions like immune response, antigen processing and presentation et al. were mainly upregulated (**Supplementary Figure S5A**). Moreover, MYC, as an oncogenic transcription factor of cell growth and proliferation via enhancing the cell cycle regulated genes (Oster et al., 2002), was significantly upregulated in Cluster L13 than in L3, with a log2-fold change of 4.28 (**Supplementary Figure S5B**). These suggested that SCLC cell lines in Cluster L13 were higher Myc-dependent than those in Cluster L3, characterized with increased cell proliferation and evasion of immune response.

Redefining the Similarity of Drugs Based on Pharmacological Subtypes

The similarity of pairwise drugs was also defined as the summarized similarity of drug response across the same cells which were tested for both drugs. For two drugs, if there are quite a few of cell pairs which are in the sensitive (or resistant) pharmacological subtypes of one drug are also in the sensitive (or resistant) subtypes of the other drug, they gain a high similarity. Otherwise, they gain a low similarity. We then generated a drug similarity profile across 367 drugs, and similarly, a hierarchical clustering algorithm was used to group drugs based on drug similarity (**Figure 4E**). The results showed that at the high level, drugs were divided into two clusters, demonstrating two different similarity patterns. The first cluster (Cluster H1) mostly included chemotherapy drugs such as those effecting in DNA replication and mitosis, and the second cluster was mixed with target and chemotherapy drugs, specifically enriched by the drugs on RTK signaling, other kinases, and cytoskeleton (**Figure 4F**). When the drugs were divided into 4 clusters (**Figures 4E,G**), we observed that the drugs enriched in L2 included chemotherapy drugs effecting in cell cycle, chromatin histone accelylation, and DNA replication, the drugs enriched in L3 were functioning in PI3K/mTOR signaling pathway, and the drugs enriched in L4 were hormone related ones. Moreover, L4, accounting for a massive part of drugs, also included the drugs targeting the different types of signaling pathways such as ABL, EGFR, Jak, and P38, RTK, and WNT signalling. These drugs did not reach the significant level but provided the obvious association.

An Access to Pharmacological Subtypes of Drugs

To enable knowledge sharing and reuse, we developed a database (http://www.hywanglab.cn/dtdb/, **Supplementary Figure S6**), which provided the comprehensive knowledge of pharmacological subtypes of drugs in this study, for supporting the further exploration of therapeutic response-based classification of tumors. The web interface of the database is user-friendly and allows users to search, browse and download data. A record of the database includes the following information: the pharmacological subtypes of a drug, the drug sensitivity values of each subtype, the distribution of histological subtypes in each pharmacological subtype, genomic and expression alterations connected to subtypes. In addition, the users can acquire the other information of a drug, a gene, or a specific cancer type in terms of our analysis. For example, given a drug, the other drugs with the most similar drug response across cancer cells are provided. Similarly, given a cancer cell line, the other cell lines with the most similar drug response across drugs are provided. Given a histological cancer type, such as SCLC, the pharmacological subtypes with the distinct treatment response were depicted, as well as the underlying mechanism of such treatment heterogeneity. The resources are expected to support the researches of precision medicine from a purely therapy-oriented perspective.

DISCUSSION

In this study, we integrated drug response information to pharmacologically reclassify tumor subtypes, as well as identify intrinsic concordant molecular mechanisms. Unlike the many studies that have aimed to unravel cancer heterogeneity by defining subtypes within the same tissues, our study aimed to systematically uncover that cancers across disparate tissues of origin can belong to one pharmacological subtype, benefiting from similar anticancer therapies, because they share a common molecular mechanism of oncogenesis. Besides, we also developed the new measures to redefine cell similarity and drug similarity from the therapeutic concordance, which provided a new point of view to study cancer heterogeneity. The similarity of cells further depicted that cells from the different origin of tissue could share the similar responses of drugs; likewise, that cells from the same origin of tissue could have distinct drug responses, thus indicating the new subtypes.

Our analysis identified, for instance, that although SKCM and COREAD are histologically different, both belonged to one pharmacological subtype, the MEK1/2 inhibitor C1 subtype, and were similarly sensitive to MEK1/2 inhibitors (**Figures 1A–D**). By connecting genomic alterations with pharmacological subtypes, mutations in the oncogene BRAF or KRAS were found to be overwhelmingly more frequent in the C1 group (**Figure 2H**). BRAF and KRAS are the kinases upstream of MEK1/2 that transmits the signals down through MEK1/2 without other major signalling branches (McCubrey et al., 2007; Simanshu et al., 2017). Our results illustrated that MEK inhibitors were able to effectively block aberrantly activated signals from BRAF or KRAS mutations. Further investigation by cBioportal (https://www.cbioportal.org/) of TCGA patients also revealed that approximately 60% of THCA (thyroid carcinomas) and 50% of SKCM (melanomas) harboured BRAF mutations, and 65% of PAAD (pancreas) and 40% COREAD (colorectal carcinomas) possessed KRAS mutations, a rate that was strikingly higher than that of other cancers (**Supplementary Figures S7A,B**). Given the high mutant frequency of BRAF in THCA and SKCM, and KRAS in PAAD and COREAD, such 4 cancer types were significantly enriched in C1 pharmacological subtype of MEK1/2 inhibitor Refametinib (THCA, p = 1.90e-2; SKCM, p = 1.23e-14; PAAD, p = 3.01e-4; COREAD, p = 3.84e-5; Hypergeometric Distribution Test) (**Supplementary Table S2**). In contrast, mutations in the oncogenes RB1 was found to be less frequent in the C1 group than in other groups. RB1 is a negative regulator of the cell cycle, with its active hypophosphorylated form binding the transcription factor E2F1 (Chen et al., 2009). RB1 mutations lead to the ectopic activation of the cell cycle, which cannot be controlled using MEK inhibitors. This may be because RB1, located downstream of MEK1/2, gain ectopic activation independent of upstream stimulation. In another instance, therapeutic concordance was identified across histologically disparate blood cancers in regard to methotrexate, a chemotherapy drug specially acting during DNA and RNA synthesis (**Figure 3H**). This is well known that Methotrexate is used in haematological malignancies but less so in solid cancers (Koźmiński et al., 2020). Functional analysis of the sensitively

correlated genes showed that haematological tumors were ectopically activated in cell cycle than solid tumors (**Figures 3I,J**). The molecular basis of such extraordinary activation of the cell cycle in haematological tumors could explain their good response to chemotherapy drugs, which work mainly by inhibiting mitosis and cell division. This demonstrates pharmacological subtypes benefiting from similar anticancer therapies can be due to common molecular mechanisms related to tissue-specific gene expression and pathways.

We also investigated the treatment heterogeneity for six types of histological cancers, including LUAD, BRCA, SKCM, SCLC, COREAD, and HNSC. High treatment heterogeneity in LUAD was observed compared with other types of cancers, suggesting that anticancer therapies for LUAD should be more complicated than those for other types of cancers (**Figures 1F–H**). Moreover, the increased number of subtypes of EGFR inhibitors in LUAD, and of drugs targeting in ERK/MAPK signalling in SKCM, indicated there was high treatment heterogeneity when EGFR inhibitors were applied into the therapies of LUAD patients and when ERK/MAPK signalling targeting drugs were applied into SKCM patients (**Figure 1H**). We also observed the low treatment heterogeneity for HNSC (**Figures 1F,G**). However, our findings were only based on the current information provided by the CGP dataset. The low treatment heterogeneity could be due to the fact that active drugs in this histological cancer are potentially still missing. Drug development in future may increase the treatment heterogeneity as well.

Our analysis showed that genes whose genomic/expression alterations frequently connect to anticancer drugs had a higher likelihood of being cancer genes. At genomic level, the top 10 genes were connected to more than 10 drugs, 30% of which were annotated as cancer censor genes. At expression level, each of genes connected to more drugs and the top genes had higher cancerous percentages. This indicated that some genes currently not identified as cancer genes could be causally implicated in cancer. In our analysis, KCTD1_PCAT18, TNFRSF9, and LTB, as the top three most connected genes, were worth for further investigation. KCTD1 (potassium channel tetramerization domain containing 1) was reported to regulate the Wnt/β-catenin pathway (Li et al., 2014; Hu et al., 2020), and PCAT18 (prostate cancer-associated transcript 18, lncRNA), was found to promote the progression of colorectal or gastric cancer through miR-759 or miR-135b (Hu et al., 2020; Zhang et al., 2020), thus KCTD1_PCAT18 fusion may resulting in the development of cancer. TNFRSF9 (tumor necrosis factor receptor superfamily member 9), also known as 4-1BB and CD137, is an immune co-stimulatory receptor. TNFRSF9 is expressed on activated immune cells including natural killer (NK) cells, effector T cells and antigen presenting cells, among them dendritic cells, macrophages, and B cells (Fröhlich et al., 2020). LTB (Lymphotoxin Beta) is a type II membrane protein of the TNF family. Giuseppina et al. found that a decrease in lymphotoxin-β production by tumor cells was associated with a loss of follicular dendritic cell phenotype and diffuse growth of follicular lymphomas (Pepe et al., 2018). Recently, the gene was found to be associated with immune infiltration of breast and endometrial cancer tumors (Ding et al., 2020; Terkelsen et al., 2020). The analysis using the TCGA dataset showed that TNFRSF9 was altered in approximately 5% of cases of Cholangiocarcinoma, adrenocortical carcinoma (ACC) and lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), with deep deletion being the dominant alteration (**Supplementary Figure S8A**). In addition, more than 14% of DLBC patients had altered LTB genes, with mutation and deletion being the major alterations (**Supplementary Figure S8B**). Therefore, the importance of the roles of these genes in cancers could be underappreciated.

Our analysis also showed that some drugs connected to many known cancer genes (**Figure 2G**). This suggested that these drugs are unspecific or the genes are important for many different pathways. So these drugs may not be ideal from a therapeutic perspective. Moreover, genes whose alterations were sensitively connected with drug resistance were significantly more frequent than those resistant connected (**Figures 2F**, **3F**). A sensitive connection meant that genomic alterations or high expression indicated increased drug sensitivity. In this way, genes with a sensitive connection could be indicators of therapeutic efficiency for connected drugs. Genes with resistant connections could be potential biomarkers for developing new therapies to overcome the resistance of connected drugs. Our analysis suggests that most of the genes whose expressions are altered during cancer initiation and development (driver genes or not) are located in ectopically activated pathways that could be controlled by anticancer drugs. Only a proportion of altered genes are outside of the cellular pathways targeted by anticancer drugs, and their alterations may maintain the growth signals of cancer cells when anticancer drugs are used. Therefore, these genes may be potential therapeutic biomarkers for overcoming drug resistance to the connected drugs.

Based on pharmacological subtypes, the cell similarity and the drug similarity could be re-defined (**Figure 4**). The cells from the same origin of tissue can be dispersed in the different pharmacological subtypes, showing the completely different response to the same drug, because of their intrinsically molecular heterogeneity. For instance, HNSC, KIRC, and SLCL cells were separated into more than two clusters with distinct drug response (**Figure 4D**). Specifically, the SCLC cells with more malignant signatures, in particular, higher expression levels of MYC, upregulated pathways involving DNA replication and cell cycle and downregulated pathways on immune response, were far away from the other SCLC cells and close to hematological tumors, instead. Interestingly, by connecting to expression alterations, our analysis further confirmed that SCLC cells more like hematological tumors had the characteristic of the ectopic activation of cell cycle, which led to their sensitive to a chemotherapy Methotrexate (**Figures 3H–J**). Our pharmacological analysis has recognized a new subtype of SCLC cells, with the high expression of an oncogenic transcription factor MYC as a marker, can benefit from chemotherapy than other SCLC subtypes. This finding was consistent with the latest studies (Bian et al., 2020; Ireland et al., 2020). Therefore, the subtypes identified by re-classifying the same origin of tissue using pharmacological data are worth to be further investigated. In addition, our analysis of the drug similarity unraveled that the drugs were categorized into two major groups, wherein the chemotherapeutic drugs tended to share the similar pattern (**Figures 4E,F**).

There are several limitations to this study. We measured the similarity of cells/drugs on the scale of all drugs/cells, so the similarity at the local space might be underestimated. Moreover, there is lack of experimental validation for our findings. To overcome these limitations, further proof of concept studies that reclassifying tumors based on therapeutic responses independent of histology subtypes is warranted.

In summary, this study demonstrates how pharmacogenomic data can be used to systematically classify cancers in terms of response to various compounds and provides us with a purely therapy-oriented perspective to view tumor classifications independent of histology subtypes. Moreover, the knowledge of pharmacological subtypes of 367 drugs are available via our website, providing the resources for precision medicine in the perspective of therapeutic response-based re-classification of tumors.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.cancerrxgene.org/downloads/bulk_download.

AUTHOR CONTRIBUTIONS

HW conceived the hypothesis. YX, JZ, and HW designed and performed the data analysis. ZC, WL, YD, XC, and TW collected and preprocessed the data. HW, YX, JZ, JK, and FZ wrote and revised the manuscript.

FUNDING

This work was supported by grants from the National Key Research and Development Program (2017YFC0908500 to HW) and the National Natural Science Foundation of China (31771469 and 31571363 to HW; 81772465 to FZ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.773101/full#supplementary-material

REFERENCES

Arozarena, I., and Wellbrock, C. (2019). Phenotype Plasticity as Enabler of Melanoma Progression and Therapy Resistance. *Nat. Rev. Cancer* 19, 377–391. doi:10.1038/s41568-019-0154-4.3

Bedard, P. L., Hansen, A. R., Ratain, M. J., and Siu, L. L. (2013). Tumour Heterogeneity in the Clinic. *Nature* 501, 355–364. doi:10.1038/nature12627.1

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodological)* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.18

Bian, X., Wang, X., Zhang, Q., Ma, L., Cao, G., Xu, A., et al. (2020). The MYC Paralog-PARP1 Axis as a Potential Therapeutic Target in MYC Paralog-Activated Small Cell Lung Cancer. *Front. Oncol.* 10, 565820. doi:10.3389/fonc.2020.565820.34

Cancer Genome Atlas Research, N.Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45, 1113–1120. doi:10.1038/ng.2764.16

Chen, H.-Z., Tsai, S.-Y., and Leone, G. (2009). Emerging Roles of E2Fs in Cancer: an Exit from Cell Cycle Control. *Nat. Rev. Cancer* 9, 785–797. doi:10.1038/nrc2696.25

Ding, H., Fan, G.-L., Yi, Y.-X., Zhang, W., Xiong, X.-X., and Mahgoub, O. K. (2020). Prognostic Implications of Immune-Related Genes' (IRGs) Signature Models in Cervical Cancer and Endometrial Cancer. *Front. Genet.* 11, 725. doi:10.3389/fgene.2020.00725.33

Ding, X., He, M., Chan, A. W. H., Song, Q. X., Sze, S. C., Chen, H., et al. (2019). Genomic and Epigenomic Features of Primary and Recurrent Hepatocellular Carcinomas. *Gastroenterology* 157, 1630–1645. doi:10.1053/j.gastro.2019.09.005.2

Drilon, A., Laetsch, T. W., Kummar, S., DuBois, S. G., Lassen, U. N., Demetri, G. D., et al. (2018). Efficacy of Larotrectinib inTRKFusion-Positive Cancers in Adults and Children. *N. Engl. J. Med.* 378, 731–739. doi:10.1056/NEJMoa1714448.15

Fröhlich, A., Loick, S., Bawden, E. G., Fietz, S., Dietrich, J., Diekmann, E., et al. (2020). Comprehensive Analysis of Tumor Necrosis Factor Receptor TNFRSF9 (4-1BB) DNA Methylation with Regard to Molecular and Clinicopathological Features, Immune Infiltrates, and Response Prediction to Immunotherapy in Melanoma. *EBioMedicine* 52, 102647. doi:10.1016/j.ebiom.2020.102647.30

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., et al. (2012). Systematic Identification of Genomic Markers of Drug Sensitivity in Cancer Cells. *Nature* 483, 570–575. doi:10.1038/nature11005.17

Gay, C. M., Stewart, C. A., Park, E. M., Diao, L., Groves, S. M., Heeke, S., et al. (2021). Patterns of Transcription Factor Programs and Immune Pathway Activation Define Four Major Subtypes of SCLC with Distinct Therapeutic Vulnerabilities. *Cancer Cell* 39, 346–360. doi:10.1016/j.ccell.2020.12.014.5

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., et al. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell* 158, 929–944. doi:10.1016/j.cell.2014.06.049.14

Hu, L., Chen, L., Yang, L., Ye, Z., Huang, W., Li, X., et al. (2020). KCTD1 Mutants in Scalp-ear-nipple S-yndrome and AP-2α P59A in Char S-yndrome R-eciprocally A-brogate T-heir I-nteractions, but C-an R-egulate Wnt/β-catenin S-ignaling. *Mol. Med. Rep.* 22, 3895–3903. doi:10.3892/mmr.2020.11457.27

Ireland, A. S., Micinski, A. M., Kastner, D. W., Guo, B., Wait, S. J., Spainhower, K. B., et al. (2020). MYC Drives Temporal Evolution of Small Cell Lung Cancer Subtypes by Reprogramming Neuroendocrine Fate. *Cancer Cell* 38, 60–78. doi:10.1016/j.ccell.2020.05.001.35

Ko, T. K., Javed, A., Lee, K. L., Pathiraja, T. N., Liu, X., Malik, S., et al. (2020). An Integrative Model of Pathway Convergence in Genetically Heterogeneous Blast Crisis Chronic Myeloid Leukemia. *Blood* 135, 2337–2353. doi:10.1182/blood.2020004834.4

Koźmiński, P., Halik, P. K., Chesori, R., and Gniazdowska, E. (2020). Overview of Dual-Acting Drug Methotrexate in Different Neurological Diseases, Autoimmune Pathologies and Cancers. *Int. J. Mol. Sci.* 21. doi:10.3390/ijms21103483.26

Li, X., Chen, C., Wang, F., Huang, W., Liang, Z., Xiao, Y., et al. (2014). KCTD1 Suppresses Canonical Wnt Signaling Pathway by Enhancing β-catenin Degradation. *PloS one* 9, e94343. doi:10.1371/journal.pone.0094343.28

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.20

Lu, H. M., Li, S., Black, M. H., Lee, S., Hoiness, R., Wu, S., et al. (2019). Association of Breast and Ovarian Cancers with Predisposition Genes Identified by Large-Scale Sequencing. *Jama Oncol.* 5, 51–57. doi:10.1001/jamaoncol.2018.2956

Marcoux, N., Gettinger, S. N., O'Kane, G., Arbour, K. C., Neal, J. W., Husain, H., et al. (2019). EGFR-mutant Adenocarcinomas that Transform to Small-Cell Lung Cancer and Other Neuroendocrine Carcinomas: Clinical Outcomes. *Jco.* 37, 278, 285. doi:10.1200/Jco.18.01585.7

McCubrey, J. A., Steelman, L. S., Chappell, W. H., Abrams, S. L., Wong, E. W. T., Chang, F., et al. (2007). Roles of the Raf/MEK/ERK Pathway in Cell Growth, Malignant Transformation and Drug Resistance. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1773, 1263–1284. doi:10.1016/j.bbamcr.2006.10.001.23

Menzer, C., Menzies, A. M., Carlino, M. S., Reijers, I., Groen, E. J., Eigentler, T., et al. (2019). Targeted Therapy in Advanced Melanoma with Rare BRAF Mutations. *Jco* 37, 3142–3151. doi:10.1200/Jco.19.00489.6

Oster, S. K., Ho, C. S. W., Soucie, E. L., and Penn, L. Z. (2002). The Myc Oncogene: Omplex. *Adv. Cancer Res.* 84, 81–154. doi:10.1016/s0065-230x(02)84004-0.22

Pepe, G., Di Napoli, A., Cippitelli, C., Scarpino, S., Pilozzi, E., and Ruco, L. (2018). Reduced Lymphotoxin-Beta Production by Tumour Cells Is Associated with Loss of Follicular Dendritic Cell Phenotype and Diffuse Growth in Follicular Lymphoma. *J. Path: Clin. Res.* 4, 124–134. doi:10.1002/cjp2.97.31

Petralia, F., Tignor, N., Reva, B., Koptyra, M., Chowdhury, S., Rykunov, D., et al. (2020). Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. *Cell* 183, 1962–1985. doi:10.1016/j.cell.2020.10.044.13

Prat, A., Parker, J. S., Fan, C., and Perou, C. M. (2012). PAM50 Assay and the Three-Gene Model for Identifying the Major and Clinically Relevant Molecular Subtypes of Breast Cancer. *Breast Cancer Res. Treat.* 135, 301–306. doi:10.1007/s10549-012-2143-0.8

Rueda, O. M., Sammut, S.-J., Seoane, J. A., Chin, S.-F., Caswell-Jin, J. L., Callari, M., et al. (2019). Dynamics of Breast-Cancer Relapse Reveal Late-Recurring ER-Positive Genomic Subgroups. *Nature*. 567, 399, 404. doi:10.1038/s41586-019-1007-8.10

Schmitz, R., Wright, G. W., Huang, D. W., Johnson, C. A., Phelan, J. D., Wang, J. Q., et al. (2018). Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 378, 1396–1407. doi:10.1056/NEJMoa1801445.11

Simanshu, D. K., Nissley, D. V., and McCormick, F. (2017). RAS Proteins and Their Regulators in Human Disease. *Cell* 170, 17–33. doi:10.1016/j.cell.2017.06.009.24

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102.19

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 47, D941–d947. doi:10.1093/nar/gky1015.21

Terkelsen, T., Russo, F., Gromov, P., Haakensen, V. D., Brunak, S., Gromova, I., et al. (2020). Secreted Breast Tumor Interstitial Fluid microRNAs and Their Target Genes Are Associated with Triple-Negative Breast Cancer, Tumor Grade, and Immune Infiltration. *Breast Cancer Res.* 22, 73. doi:10.1186/s13058-020-01295-6.32

Waks, A. G., and Winer, E. P. (2019). Breast Cancer Treatment. *Jama* 321, 288–300. doi:10.1001/jama.2018.19323.9

Zhang, X.-Z., Mao, H.-L., Zhang, S.-J., Sun, L., Zhang, W.-J., Chen, Q.-Z., et al. (2020). lncRNA PCAT18 Inhibits Proliferation, Migration and Invasion of Gastric Cancer Cells through miR-135b Suppression to Promote CLDN11 Expression. *Life Sci.* 249, 117478. doi:10.1016/j.lfs.2020.117478.29

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Xu, Zheng, Cai, Li, Köhler, Dai, Cheng, Wu, Zhang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSARY

ACC Adrenocortical carcinoma

ALL Acute lymphoblastic leukemia

BLCA Bladder urothelial carcinoma

BRCA Breast invasive carcinoma

CESC Cervical squamous cell carcinoma and endocervical adenocarcinoma

CLL Chronic lymphocytic leukemia

COREAD Colon adenocarcinoma and rectum adenocarcinoma

DLBC Lymphoid neoplasm diffuse large B-cell Lymphoma

ESCA Esophageal carcinoma

GBM Glioblastoma multiforme

HNSC Head and neck squamous cell carcinoma

KIRC Kidney renal clear cell carcinoma

LAML Acute myeloid leukemia

LCML Chronic myelogenous leukemia

LGG Brain lower grade glioma

LIHC Liver hepatocellular carcinoma

LUAD Lung adenocarcinoma

LUSC Lung squamous cell carcinoma

MB Medulloblastoma

MESO Mlistdefinitionesothelioma

MM Multiple myeloma

NB Neuroblastoma

OV Ovarian serous cystadenocarcinoma

PAAD Pancreatic adenocarcinoma

PRAD Prostate adenocarcinoma

SCLC Small cell lung cancer

SKCM Skin cutaneous melanoma

STAD Stomach adenocarcinoma

THCA Thyroid carcinoma

UCEC Uterine corpus endometrial carcinoma

CGP The Cancer Genome Project

IC50 Half maximal inhibitory concentration

TCGA The Cancer Genome Atlas Program

FDA U.S. Food and Drug Administration

RTK Receptor tyrosine kinase

MEK Mitogen-activated protein kinase kinase.

frontiers in Cell and Developmental Biology

REVIEW
published: 25 November 2021
doi: 10.3389/fcell.2021.795883

Genomic Variation Prediction: A Summary From Different Views

Xiuchun Lin *

College of Information and Electrical Engineering, China Agricultural University, Beijing, China

Structural variations in the genome are closely related to human health and the occurrence and development of various diseases. To understand the mechanisms of diseases, find pathogenic targets, and carry out personalized precision medicine, it is critical to detect such variations. The rapid development of high-throughput sequencing technologies has accelerated the accumulation of large amounts of genomic mutation data, including synonymous mutations. Identifying pathogenic synonymous mutations that play important roles in the occurrence and development of diseases from all the available mutation data is of great importance. In this paper, machine learning theories and methods are reviewed, efficient and accurate pathogenic synonymous mutation prediction methods are developed, and a standardized three-level variant analysis framework is constructed. In addition, multiple variation tolerance prediction models are studied and integrated, and new ideas for structural variation detection based on deep information mining are explored.

Keywords: genome, variation, machine learning, genomic mutation, prediction

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Fei Guo,
Tianjin University, China
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Xiuchun Lin
linxc124@163.com
1241562853@qq.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *15 October 2021*
Accepted: *11 November 2021*
Published: *25 November 2021*

Citation:
Lin X (2021) Genomic Variation Prediction: A Summary From Different Views. Front. Cell Dev. Biol. 9:795883. doi: 10.3389/fcell.2021.795883

INTRODUCTION

The decreasing cost of genome sequencing has resulted in a large amount of sequence information and variation data becoming available. According to the number of mutated bases, genomic variations have been classified as: 1) single-nucleotide variations (formerly single-nucleotide polymorphisms); 2) very short insertions and deletions, usually less than 50 bp; and 3) structural variations, usually longer than 50 bp (Genome structural variati, 2011). Gene mutations are known to be closely related to the occurrence and development of diseases (Hunt et al., 2014; Alkan et al., 2011; Yin et al., 2020; Fang et al., 2019; Li et al., 2020a; He et al., 2020; Zhang et al., 2020a; Zhang et al., 2016; Hu et al., 2021; Hu et al., 1990; Hu et al., 2020). High-throughput sequencing technologies have allowed the mutations in the genomes of patients with particular diseases to be determined systematically, quickly and accurately, including common but less studied synonymous mutations in the coding regions of genomes (Meyerson et al., 2010; Li et al., 2017; Cheng et al., 2018; Zhou et al., 2019). Synonymous mutations are single-nucleotide mutations that occur in the coding regions of genes but do not change the amino acid sequence of the protein due to the degeneracy of the genetic code. Because they do not change the coded amino acids, synonymous mutations were once mistakenly thought to have no biological function (Hong et al., 2020; Tang et al., 2020; Cheng et al., 2021a). However, later systematic studies have shown that synonymous mutations are involved in a variety of biological processes and play important roles in the occurrence and development of diseases (Li et al., 2020b; Yang et al., 2021a). Whole genome sequencing using reversible terminator chemistry can generate accurate nucleotide sequences of billions of bases at low cost (Bentley et al., 2008), which greatly improves the data obtained in sequencing projects. Structural variations in genomes are closely related to the occurrence and development of many diseases that affect human health. Therefore, the detection of structural mutations is essential to

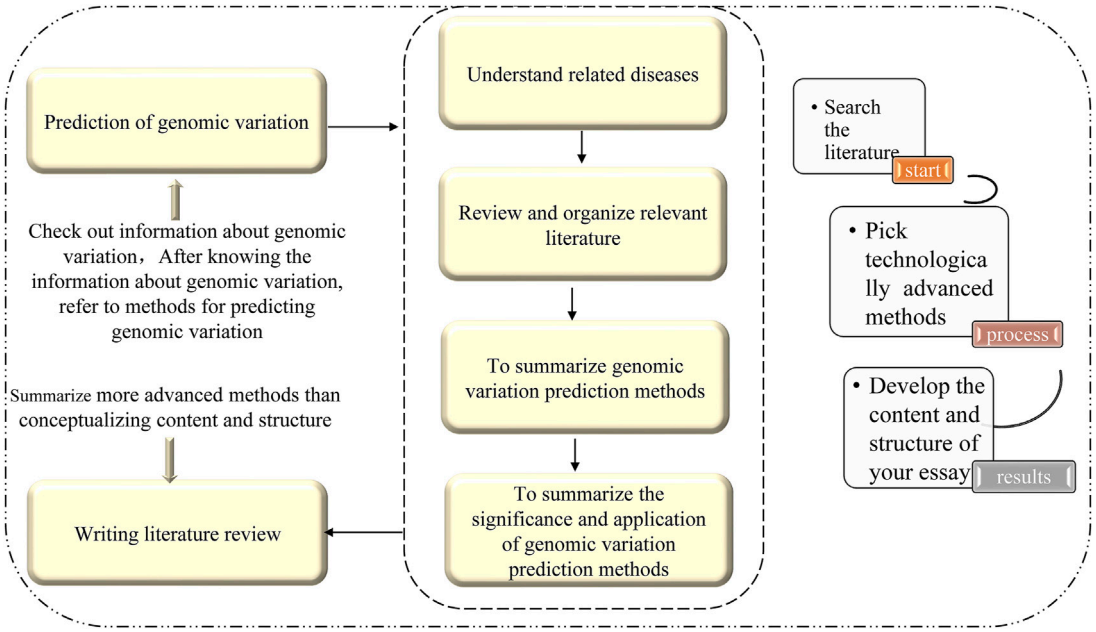


FIGURE 1 | Idea map of the paper.

understand the mechanisms of diseases, find pathogenic targets, and carry out personalized precision medicine (Guo et al., 2011; Liu et al., 2021a; Yu et al., 2021a; Yu et al., 2021b). However, detecting structural variations can be difficult, and therefore methods that can accurately and quickly predict genomic variations are urgently required.

In this paper, an efficient and accurate method for predicting disease-causing synonymous mutations based on existing research and using machine learning theories and methods is reviewed. This method not only incorporates the current understanding of the pathogenic mechanism of synonymous mutations, but also provides a theoretical basis for the diagnosis and treatment of diseases, drug development, and the development of memory precision medicine. There are three steps to this method: 1) defining a standardized variation analysis framework based entirely on genome sequencing data; 2) using computational methods to construct a variation tolerance prediction model as the classification basis, and two high-performance variation mechanisms (influence of protein solubility and metabolic stability) prediction models; and 3) developing software tools and combining multiple models to provide general variation analysis services, as well as medical research and precision medical services. For structural variation detection based on deep mining of information, the following key technologies and methods are established: 1) extract sequence features related to structural variations from different sides and establish a comprehensive representation of variation; 2) use the sequence to comparison text information and the corresponding variation feature map to generate images and amplify imbalanced images of small samples; and 3) using the powerful feature representation capabilities of deep learning, automatically extract global features, hidden features, and associated features to complete variation detection. This approach will be an effective way to improve the accuracy of genome structural variation detection, and will also help to promote the development of new structural variation detection technologies. The outline of this study is shown in **Figure 1**.

GENOMIC VARIATION PREDICTION METHODS

Variations in the human genome are related to human evolution and disease risk (Jiang and Liu, 2016; Jiang et al., 2017; Liu et al., 2017; Liu et al., 2018a; Liu et al., 2018b; Liu et al., 2018c; Liu et al., 2019; Wu et al., 2019; Xu et al., 2019; Deng et al., 2021). Moreover, with the systematic in-depth studies of single-nucleotide mutations, especially those that have special genetic variation patterns such as synonymous mutations, the understanding of the composition of the human genome, genetic differences between individuals, and the pathogenic mechanisms of diseases has greatly improved. The genetic variation prediction method reviewed in this paper will help to make such studies more convenient and economical by identifying variations of interest that can be targeted (**Table 1**).

Pathogenic Synonymous Mutations

In recent years, the interest and attention of researchers in the analysis and prediction of pathogenic synonymous mutations have increased. Published methods include SilVA (Buske et al., 2013), DDIG-SN (Livingstone et al., 2017), regSNPs-splicing (Zhang et al., 2017a), Syntool (Zhang et al., 2017b) and TraP (Gelfman et al., 2017). sHowever, the available prediction methods still have certain defects that need improvement.

TABLE 1 | Summary of genomic variation prediction methods.

Type	Methods	Algorithm
Pathogenic synonymous mutations	SilVA (Buske et al., 2013)	Random forest
	DDIG-SN(Livingstone et al., 2017)	Support vector machine
	regSNPs-splicing (Zhang et al., 2017a)	Random forest
	Syntool (Zhang et al., 2017b)	—
	TraP (Gelfman et al., 2017)	Random forest
Genome sequencing	CADD (Kircher et al., 2014)	Support vector machine
	MutationTaster2 (Cooper, 2014)	Naive Bayes
	Mut-Pred (Li et al., 2009)	Random forest
	PolyPhen-2 (Adzhubei et al., 2010)	Naive Bayes
	PON-P2 (Niroula et al., 2015)	Random forest
	VEST (Carter et al., 2013)	Random forest
Deep mining of structural variation information	DeepBind(Alipanahi et al., 2015)	deep learning
	DeepVariant (Angermueller et al., 2017)	deep neural networks
	DeepCpG (Poplin et al., 2018)	deep neural networks

Among them, the use of machine learning methods to predict pathogenic synonymous mutations is still in the preliminary stage. The main problems that remain to be solved include: 1) positive sample data is scarce and standard negative sample data is lacking (Zhang et al., 2020b); 2) feature representation ability is weak and not easy to promote (Buske et al., 2013; Wei et al., 2018; Xiong et al., 2018; Jin et al., 2019; Shen et al., 2019; Su et al., 2019; Wei et al., 2019; Yang et al., 2020a; Zhang et al., 2020c; Peng et al., 2020; Su et al., 2020; Teng et al., 2020; Chu et al., 2021a; Cheng et al., 2021b; Chu et al., 2021b; Jin et al., 2021; Su et al., 2021); and 3) the prediction performances of existing methods need to be improved, and the results of different methods have a low degree of coincidence (Cheng et al., 2019). The methods reviewed in this article aim to solve these problems.

The aim of this project is to develop an efficient and accurate method for predicting disease-causing synonymous mutations. The main steps are as follows: 1) establish a data set using a variety of different methods and data sources; 2) analyze in detail the biological characteristic attributes related to pathogenic synonymous mutations; 3) design machine learning methods to predict mutations; and 4) develop a public service platform and corresponding software system to predict disease-related mutations. The following aspects were included in the method. 1) A pathogenic synonymous mutation database and benchmark data set are constructed. Then, pathogenic synonymous mutation data reported in the literature are collected to supplement the database, and the two types of data are integrated to improve the database. 2) A feature representation method of the pathogenic mechanism of the synonymous mutation is established. Synonymous mutations can occur in various processes of gene expression. In this paper, the pathogenic principle of synonymous mutations was fully utilized, and the method of numerical expression of pathogenic synonymous mutations was studied at the DNA, RNA, and protein levels. In addition, the same data set and the same machine learning model are combined for testing, and then the feature selection method is used to remove irrelevant features from the extracted features to select a relatively good feature representation method. 3) A prediction method for pathogenic synonymous mutations with convolutional neural network was used as the basic model. The aim was to learn the representation method of pathogenic synonymous mutation data based on deep learning, especially the efficient implicit feature representation ability. The deep feature representation method of the biological characteristics of pathogenic synonymous mutations is also used to make up for the lack of feature representation ability of shallow learning. Then, deep network structure design and deep model training optimization strategies are studied to improve the robustness and generalization performance of the model. 4) The prediction method of pathogenic synonymous mutation based on ensemble learning is evaluated. To reduce the correlation of individual classifier results, a method suitable for training and learning is selected from the existing prediction methods to obtain the individual classifier. Then, randomization is introduced in the learning and training processes to obtain diverse individual classifiers. Finally, a learning algorithm for the integrated decision-making classifier is designed to construct a suitable secondary classifier to more effectively solve the problem of pathogenic synonymous mutation prediction. How to perform ensemble pruning after the generation of cascading ensemble classifiers should also be studied to further improve the predictive performance of the cascading ensemble learning classifier and obtain better pathogenic synonymous mutation predictions. 5) Result verification and algorithm software development are performed. The predictive analysis data obtained in 3 and 4 above are analyzed and the results with high reliability and potential clinical application value are selected to carry out molecular and biochemical experiments to determine the biological functions of synonymous mutations at the cellular level and to verify the accuracy of the prediction results.

Genome Sequencing

Methods to rapidly and comprehensively interpret the various new variations identified in genome sequencing are lacking. Therefore, it is not yet possible to associate variations with possible diseases or health issues, which greatly reduces the value of genome sequencing. A primary task of sequencing research is how to analyze the sequence data, especially the variation information that it contains. Prediction methods can effectively solve this problem (Castrense et al., 2019; Gang et al.,

2019; Liu et al., 2020a; Jin et al., 2021; Yin et al., 2021). The published methods include CADD (Kircher et al., 2014), MutationTaster2 (Cooper, 2014),Mut-Pred (Li et al., 2009),PolyPhen-2 (Adzhubei et al., 2010),PON-P2 (Niroula et al., 2015) and VEST (Carter et al., 2013). The aim of the project was to design a standard variation analysis framework for genomic variation prediction. Specifically, a variation tolerance prediction model is constructed based on the genome sequencing data and calculation method, and a relatively high-performance variation mechanism is also constructed based on the influence of protein solubility and metabolic stability. The following aspects are included. 1) A general prediction model of variation tolerance classification is constructed based on the existing variation data for all kinds of variations (e.g., replacement, insertion, and deletion) using only the sequence information as the basic and important step of the analysis services. A highly accurate predictive model ProtSol is constructed. New data will be collected and sorted out, the training data set will be integrated, and input features will be selected. Then, the classification algorithm will be optimized and a protein solubility impact prediction model with higher accuracy and greater generalization will be established for variation analysis. 3) A model ProtMS to predict the effect of variation on protein metabolism stability is constructed. Classification and regression models based on sequence information will be established to predict the impact of variations on metabolic stability. These models will serve as important parts of the mechanism analysis. In addition, browser/server architecture variant analysis software tools will be developed and released to provide online services for researchers and clinical medical staff.

Deep Mining of Structural Variation Information

Structural variations in the human genome can cause diseases (Feuk, et al., 2006). Structural variations include translocation, inversion, deletion, and duplication of genes, and accurate detection of genetic variations or genetic testing can contribute to the exploration and analysis of diseases and life processes. The obtained structural variant information can be applied, for example, to target drugs to tumors and to provide a reliable reference for clinical applications (Xu et al., 2018; Xue et al., 2018; Tang et al., 2019; Liu et al., 2020b; Zhang et al., 2020d; An and Yu, 2021; Liu et al., 2021b; Wang et al., 2021; Wu and Yu, 2021). Therefore, we need to mine structural variation information accurately. Published methods include DeepBind (Alipanahi et al., 2015), DeepVariant (Angermueller et al., 2017), and DeepCpG (Poplin et al., 2018). Deep mining of the human genomic structural variation information includes the following aspects. 1) Comprehensive characterization of features using three basic detection methods: sequence features related to structural variation from different aspects, definition of representation modes of different types of variation, and construction of classification and combined expression of feature descriptions. Then, the comprehensive characteristics are obtained by combining three types of detection methods (i.e., double-terminal mapping-based, split-comparison-based, and mapping depth-based methods) to form a complete comprehensive characterization of variation features. 2) Feature image generation whereby images are generated to discover comprehensive variation information from two main aspects, pixel composition and color expression. Pixel composition includes base stacking and fragment tiled, and color expression includes stacked base color expression and tiled fragment color expression. 3) Data amplification to ensure that the generated images can be used for deep learning training and recognition. The main purpose of data amplification is to avoid the problem of network overfitting or a decline in the recognition of minority samples. This step includes mainly studying the method of amplification of mutated images different from natural images; studying sufficiently clear and high-resolution amplified images, and studying the speed of image generation to ensure the network is better able to learn the mapping of complex functions. A schematic diagram of this method of deep mining of structural information is given in **Figure 2**.

LITERATURE CONTRIBUTION

Early studies of the human genome focused mainly on collecting data and understanding structural variation at the genome level. In this paper, the literature closely related to diseases and current medical fields was reviewed, including genome prediction methods and many aspects of genome variation. The genomic variants associated with the diseases are summarized in **Table 2**. Bioinformatics analysis studies of pathogenic synonymous mutations aim to integrate different data sources and numerical types and detect reliable characteristics for feature representation methods, and also to accurately characterize feature coding methods that are intrinsically linked to pathogenic synonymous mutations. For ensemble learning methods, it is necessary to propose a pruning mechanism for the primary classifier learning algorithm with adaptive learning ability. Genome sequencing studies aim to establish a set of standardized solutions and integrate existing databases and multiple prediction models. The goal is to solve the versatility and generalization of variation analysis models, to solve the data imbalance problem that is common in variation data, and to find sequence-based biological characteristics in different variation prediction models. The contribution of the deep mining of structural variation information is to develop a new way of generating variation feature images from sequence comparison text information, a method of data amplification for small samples of variation images to achieve a balance, and an accurate detection technology framework by digging deep into the genomic structural variation information. The proposed points can take genomic variation prediction research one step further, and provide medication recommendations for the treatment of specific diseases, thereby reducing the adverse effects on patients due to improper medication strategies. This is one of the reasons why genomic variation is an important area of study.

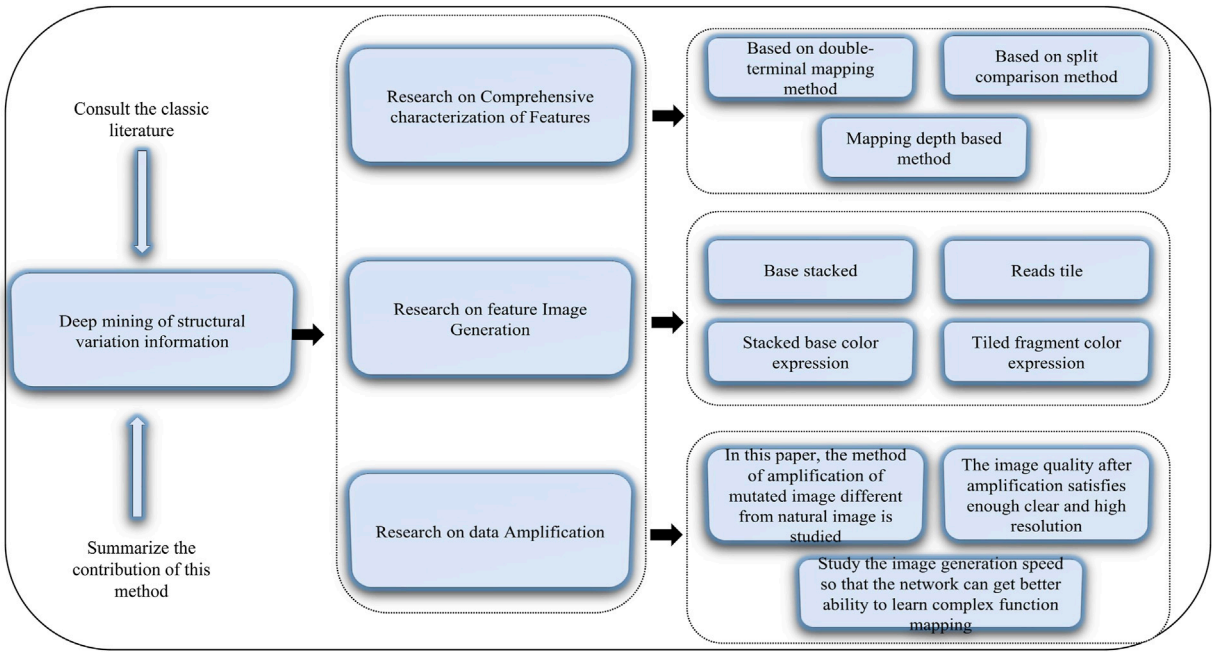


FIGURE 2 | Diagram of deep mining of structural variation information.

TABLE 2 | Summary of genomic variations associated with disease.

Disease	Causes	Result
Type 2 diabetes (Freemantle et al., 2005)	There were 139 common gene variants and 4 rare gene variants	Availability of Inhaled Insulin Promotes greater perceived acceptance of insulin therapy in Patients with type 2 diabetes
Neonatal epilepsy (Thuresson et al., 2016)	Whole gene repeats of SCN2A and SCN3A	Extra copy of SCN2A has an effect on epilepsy pathogenesis
Bladder cancer (Bonberg et al., 2013)	Copy number variation in GSTM1 gene	A loss of 9p21 was less predictive for detecting bladder cancer
Lung cancer (Yang et al., 2013)	Cnv-67048 variation on WWOX	be related with altered WWOX gene expression and exons absence in them
A wide variety of tumor (Abdel-Rahman et al., 2011)	BAP1 mutation	BAP1 is the candidate gene in only a small subset of hereditary UM, suggesting the contribution of other candidate genes.

CONCLUSION

The study of structural variations in genomes can promote research on genome evolution, significant biological phenotypic changes (Yang et al., 2020b; Yang et al., 2020c; Yin et al., 2020), the treatment of many diseases (Li et al., 2018; Yang et al., 2021b; Long et al., 2021), and recommendations for therapeutic drugs (Wei et al., 2014; Ding et al., 2020a; Ding et al., 2020b; Wang et al., 2020; Wei et al., 2020). The accurate prediction of genomic variation is of great importance to studies of many diseases, which indicates the significance of this literature review through which existing variation data were integrated and collected, and a tolerance classification model of various variations was constructed based on sequence information. Furthermore, all the literature is experimenting around key scientific issues. It is essential to accurately predict individual genomic variation events that are conducive to systematically inferring the process of variation formation, so that the results can be confidently used for the clinical application of precision medicine. Finally, accurate predictions also help in analyzing the functions of synonymous mutations and can guide relevant experiments. Therefore, genomic variation prediction is of great significance to drug design and precision medicine.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

We thank Margaret Biswas, PhD, from Liwen Bianji (Edanz) (www.liwenbianji.cn/) for editing the English text of a draft of this manuscript.

REFERENCES

Abdel-Rahman, M. H., Pilarski, R., Cebulla, C. M., Massengill, J. B., Christopher, B. N., Boru, G., et al. (2011). Germline BAP1 Mutation Predisposes to Uveal Melanoma, Lung Adenocarcinoma, Meningioma, and Other Cancers. *J. Med. Genet.* 48 (12), 856–859. doi:10.1136/jmedgenet-2011-100156

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nat Methods*. 7(4), 248–9. doi:10.1038/nmeth0410-248

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning. *Nat. Biotechnol.* doi:10.1038/nbt.3300

An, Q., and Yu, L. (2021). A Heterogeneous Network Embedding Framework for Predicting Similarity-Based Drug-Target Interactions. *Brief. Bioinformatics*. 22, bbab275. doi:10.1093/bib/bbab275

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). Erratum to: DeepCpG: Accurate Prediction of Single-Cell DNA Methylation States Using Deep Learning. *Genome Biol.* 18 (1), 90. doi:10.1186/s13059-017-1233-z

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry. *Nature* 456 (7218), 53–59. doi:10.1038/nature07517

Bonberg, N., Taeger, D., Gawrych, K., Johnen, G., Banek, S., Schwentner, C., et al. (2013). Chromosomal Instability and Bladder Cancer: the UroVysionTMtest in the UroScreen Study. *BJU Int.* 112 (4), E372–E382. doi:10.1111/j.1464-410x.2012.11666.x

Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N., and Brudno, M. (2013). Identification of Deleterious Synonymous Variants in Human Genomes. *Bioinformatics* 29, 1843–1850. doi:10.1093/bioinformatics/btt308

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., and Karchin, R. (2013). Identifying Mendelian Disease Genes with the Variant Effect Scoring Tool. *BMC Genomics* 14 (3), S3–S16. doi:10.1186/1471-2164-14-S3-S3

Castrense, S., Giulia, B., Samuele, B., Emidio, C., Pier, L. M., and Rita, C. (2019). Are Machine Learning Based Methods Suited to Address Complex Biological Problems? Lessons from CAGI-5 Challenges. *Hum. Mutat.* 40, 1455–1462. doi:10.1002/humu.23784

Cheng, L., Qi, C., Yang, H., Lu, M., Cai, Y., Fu, T., et al. (2021). gutMGene: a Comprehensive Database for Target Genes of Gut Microbes and Microbial Metabolites. *Nucleic Acids Res.* 9, gkab786. doi:10.1093/nar/gkab786

Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., and Zhang, X. (2021). Functional Alterations Caused by Mutations Reflect Evolutionary Trends of SARS-CoV-2. *Brief. Bioinformatics* 22 (2), 1442–1450. doi:10.1093/bib/bbab042

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics* 34 (11), 1953–1956. doi:10.1093/bioinformatics/bty002

Cheng, N., Li, M., Zhao, L., Zhang, B., Yang, Y., Zheng, C. H., et al. (2019). Comparison and Integration of Computational Methods for Deleterious Synonymous Mutation Prediction. *Brief. Bioinformatics*. 21, 970–981. doi:10.1093/bib/bbz047

Chu, Y., Wang, X., Dai, Q., Wang, Y., Wang, Q., Peng, S., et al. (2021). MDA-GCNFTG: Identifying miRNA-Disease Associations Based on Graph Convolutional Networks via Graph Sampling through the Feature and Topology Graph. *Brief Bioinform*.

Chu, Y., Kaushik, A. C., Wang, X., Wang, W., Zhang, Y., Shan, X., et al. (2021). DTI-CDF: a cascade Deep forest Model towards the Prediction of Drug-Target Interactions Based on Hybrid Features. *Brief Bioinform* 22 (1), 451–462. doi:10.1093/bib/bbz152

Cooper, N. D. (2014). *MutationTaster2: Mutation Prediction for the Deep-Sequencing Age [Letter]*.

Deng, L., Li, W., and Zhang, J. (2021). LDAH2V: Exploring Meta-Paths across Multiple Networks for lncRNA-Disease Association Prediction. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18 (4), 1572–1581. doi:10.1109/tcbb.2019.2946257

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. *Knowledge-Based Syst.* 204, 106254. doi:10.1016/j.knosys.2020.106254

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug-Target Interactions via Fuzzy Bipartite Local Model. *Neural Comput. Applic* 32, 10303–10319. doi:10.1007/s00521-019-04569-z

Fang, S., Pan, J., Zhou, C., Tian, H., He, J., Shen, W., et al. (2019). Circular RNAs Serve as Novel Biomarkers and Therapeutic Targets in Cancers. *Cgt* 19 (2), 125–133. doi:10.2174/1566523218666181109142756

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural Variation in the Human Genome. *Nat. Rev. Genet.* 7, 85–97. doi:10.1038/nrg1767

Freemantle, N., Blonde, L., Duhot, D., Hompesch, M., Eggertsen, R., Hobbs, F. D. R., et al. (2005). Availability of Inhaled Insulin Promotes Greater Perceived Acceptance of Insulin Therapy in Patients with Type 2 Diabetes. *Diabetes care* 28, 427–428. doi:10.2337/diacare.28.2.427

Gang, H., Liu, G., Zhang, M., Zhao, Y., Jiang, J., and Chen, S. (2019). Comprehensive Characterization of T-DNA Integration Induced Chromosomal Rearrangement in a Birch T-DNA Mutant. *BMC Genomics* 20 (1), 311. doi:10.1186/s12864-019-5636-y

Gelfman, S., Wang, Q., Mcsweeney, K. M., Ren, Z., La Carpia, F., Halvorsen, M., et al. (2017). Annotating Pathogenic Non-coding Variants in Genic Regions. *Nat. Commun.* 8 (1), 236. doi:10.1038/s41467-017-00141-2

Alkan, C., Coe, B., and Eichler, E. E. (2011). Genome Structural Variation Discovery and Genotyping. *Nat Rev Genet*. 12, 363–76. doi:10.1038/nrg2958

Guo, F., and Wang, L. (2011). "Computing the Protein Binding Sites,". *Bioinformatics Research and Applications*. Editors J. Chen and J. X. Wang (Changsha, China: Zelikovsky A), 6674, 25–36. doi:10.1007/978-3-642-21260-4_7

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020). TOOme: A Novel Computational Framework to Infer Cancer Tissue-Of-Origin by Integrating Both Gene Mutation and Expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

Hong, J., Luo, Y., Zhang, Y., Ying, J., Xue, W., Xie, T., et al. (2020). Protein Functional Annotation of Simultaneously Improved Stability, Accuracy and False Discovery Rate Achieved by a Sequence-Based Deep Learning. *Brief Bioinform* 21 (4), 1437–1447. doi:10.1093/bib/bbz081

Hu, Y., Qiu, S., and Cheng, L. (2021). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 7036592. doi:10.1155/2021/7036592

Hu, Y., Sun, J. Y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (1990). Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19 (1), 11.

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 Regulates TSPAN13 Expression and Contributes to Alzheimer's Disease with Sex Differences. *Brain* 143 (11), e95. doi:10.1093/brain/awaa302

Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E., and Kimchi-Sarfaty, C. (2014). Exposing Synonymous Mutations. *Trends Genet.* 30, 308-21. doi:10.1016/j.tig.2014.04.006

Jiang, Q., and Liu, G. (2016). Lack of Association between MC1R Variants and Parkinson's Disease in European Descent. *Ann. Neurol.* 79, 866–868. doi:10.1002/ana.24627

Jiang, Q., Jin, S., Jiang, Y., Liao, M., Feng, R., Zhang, L., et al. (2017). Alzheimer's Disease Variants with the Genome-wide Significance Are Significantly Enriched in Immune Pathways and Active in Immune Cells. *Mol. Neurobiol.* 54 (1), 594–600. doi:10.1007/s12035-015-9670-8

Jin, Q., Cui, H., Sun, C., Meng, Z., and Su, R. (2021). Free-form Tumor Synthesis in Computed Tomography Images via Richer Generative Adversarial Network. *Knowledge-Based Syst.* 218, 106753. doi:10.1016/j.knosys.2021.106753

Jin, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L., Su, R., et al. (2019). DUNet: A Deformable Network for Retinal Vessel Segmentation. *Knowledge-Based Syst.* 178, 149–162. doi:10.1016/j.knosys.2019.04.025

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nat. Genet.* 46 (3), 310–315. doi:10.1038/ng.2892

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., et al. (2009). Automated Inference of Molecular Mechanisms of Disease from Amino Acid Substitutions. *Bioinformatics* 25, 2744–2750. doi:10.1093/bioinformatics/btp528

Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., et al. (2017). NOREVA: Normalization and Evaluation of MS-based Metabolomics Data. *Nucleic Acids Res.* 45 (W1), W162–W170. doi:10.1093/nar/gkx449

Li, F., Zhou, Y., Zhang, X., Tang, J., Yang, Q., Zhang, Y., et al. (2020). SSizer: Determining the Sample Sufficiency for Comparative Biological Study. *J. Mol. Biol.* 432 (11), 3411–3421. doi:10.1016/j.jmb.2020.01.027

Li, Y. H., Li, X. X., Hong, J. J., Wang, Y. X., Fu, J. B., Yang, H., et al. (2020). Clinical Trials, Progression-Speed Differentiating Features and Swiftness Rule of the Innovative Targets of First-In-Class Drugs. *Brief. Bioinformatics* 21 (2), 649–662. doi:10.1093/bib/bby130

Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., et al. (2018). Therapeutic Target Database Update 2018: Enriched Resource for Facilitating Bench-To-Clinic Research of Targeted Therapeutics. *Nucleic Acids Res.* 46 (D1), D1121–D1127. doi:10.1093/nar/gkx1076

Liu, G., Hu, Y., Han, Z., Jin, S., and Jiang, Q. (2019). Genetic Variant Rs17185536 Regulates SIM1 Gene Expression in Human Brain Hypothalamus. *Proc. Natl. Acad. Sci. USA* 116 (9), 3347–3348. doi:10.1073/pnas.1821550116

Liu, G., Hu, Y., Jin, S., and Jiang, Q. (2017). Genetic Variant Rs763361 Regulates Multiple Sclerosis CD226 Gene Expression. *Proc. Natl. Acad. Sci. USA* 114 (6), E906–E907. doi:10.1073/pnas.1618520114

Liu, G., Jin, S., Hu, Y., and Jiang, Q. (2018). Disease Status Affects the Association between Rs4813620 and the Expression of Alzheimer's Disease Susceptibility geneTRIB3. *Proc. Natl. Acad. Sci. USA* 115 (45), E10519–E10520. doi:10.1073/pnas.1812975115

Liu, G., Wang, T., Tian, R., Hu, Y., Han, Z., Wang, P., et al. (2018). Alzheimer's Disease Risk Variant Rs2373115 Regulates GAB2 and NARS2 Expression in Human Brain Tissues. *J. Mol. Neurosci.* 66 (1), 37–43. doi:10.1007/s12031-018-1144-9

Liu, G., Zhang, Y., Wang, L., Xu, J., Chen, X., Bao, Y., et al. (2018). Alzheimer's Disease Rs11767557 Variant Regulates EPHA1 Gene Expression Specifically in Human Whole Blood. *Jad* 61 (3), 1077–1088. doi:10.3233/jad-170468

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., and Deng, L. (2020). DrugCombDB: a Comprehensive Database of Drug Combinations toward the Discovery of Combinatorial Therapy. *Nucleic Acids Res.* 48 (D1), D871–D881. doi:10.1093/nar/gkz1007

Liu, J., Liu, S., Liu, C., Zhang, Y., Pan, Y., Wang, Z., et al. (2021). Nabe: an Energetic Database of Amino Acid Mutations in Protein-Nucleic Acid Binding Interfaces. *Database (Oxford)* 2021, 2021. doi:10.1093/database/baab050

Liu, J., Su, R., Zhang, J., and Wei, L. (2021). Classification and Gene Selection of Triple-Negative Breast Cancer Subtype Embedding Gene Connectivity Matrix in Deep Neural Network. LID - Bbaa395 [pii] LID -. *Briefings in Bioinformatics*, 2021, 1477–4054. doi:10.1093/bib/bbaa395

Liu, Y., Huang, Y., Wang, G., and Wang, Y. (2020). A Deep Learning Approach for Filtering Structural Variants in Short Read Sequencing Data. *Brief Bioinform.*

Livingstone, M., Folkman, L., Yang, Y., Zhang, P., Mort, M., Cooper, D. N., et al. (2017). Investigating DNA-, RNA-, and Protein-Based Features as a Means to Discriminate Pathogenic Synonymous Variants. *Hum. Mutat.* 38 (10), 1336–1347. doi:10.1002/humu.23283

Long, J., Yang, H., Yang, Z., Jia, Q., Liu, L., Kong, L., et al. (2021). Integrated Biomarker Profiling of the Metabolome Associated with Impaired Fasting Glucose and Type 2 Diabetes Mellitus in Large-Scale Chinese Patients. *Clin. Transl Med.* 11 (6), e432. doi:10.1002/ctm2.432

Meyerson, M., Gabriel, S., Getz, G., Meyerson, M., Gabriel, S., and GAdvances, Getz. (2010). Advances in Understanding Cancer Genomes through Second-Generation Sequencing. *Nat. Rev. Genet.* 11 (10), 685–696. doi:10.1038/nrg2841

Niroula, A., Urolagin, S., and Vihinen, M. (2015). PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *Plos One* 10 (2), e0117380. doi:10.1371/journal.pone.0117380

Peng, L., Zhou, D., Liu, W., Zhou, L., Wang, L., Zhao, B., et al. (2020). Prioritizing Human Microbe-Disease Associations Utilizing a Node-Information-Based Link Propagation Method. *IEEE Access* 8, 31341–31349. doi:10.1109/access.2020.2972283

Poplin, R., Chang, P-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). Creating a Universal SNP and Small Indel Variant Caller with Deep Neural Networks. *bioRxiv*, 092890.

Shen, Y., Tang, J., and Guo, F. (2019). Identification of Protein Subcellular Localization via Integrating Evolutionary and Physicochemical Information into Chou's General PseAAC. *J. Theor. Biol.* 462, 230–239. doi:10.1016/j.jtbi.2018.11.012

Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2020). Empirical Comparison and Analysis of Web-Based Cell-Penetrating Peptide Prediction Tools. *Brief. Bioinformatics* 21 (2), 408–420. doi:10.1093/bib/bby124

Su, R., Liu, X., Jin, Q., Liu, X., and Wei, L. (2021). Identification of Glioblastoma Molecular Subtype and Prognosis Based on Deep MRI Features. *Knowledge-Based Syst.* 232, 107490. doi:10.1016/j.knosys.2021.107490

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A Deep forest Model to Predict Anti-cancer Drug Response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., et al. (2020). ANPELA: Analysis and Performance Assessment of the Label-free Quantification Workflow for Metaproteomic Studies. *Brief. Bioinformatics* 21 (2), 621–636. doi:10.1093/bib/bby127

Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., et al. (2019). Simultaneous Improvement in the Precision, Accuracy, and Robustness of Label-free Proteome Quantification by Optimizing Data Manipulation Chains*. *Mol. Cell Proteomics* 18 (8), 1683–1699. doi:10.1074/mcp.ra118.001169

Teng, H., Wei, W., Li, Q., Xue, M., Shi, X., Li, X., et al. (2020). Prevalence and Architecture of Posttranscriptionally Impaired Synonymous Mutations in 8,320 Genomes across 22 Cancer Types. *Nucleic Acids Res.* 48 (3), 1192–1205. doi:10.1093/nar/gkaa019

Thuresson, A. C., Van Buggenhout, G., Sheth, F., Kamate, M., Andrieux, J., Clayton Smith, J., et al. (2016). Whole Gene Duplication of SCN2A and SCN3A Is Associated with Neonatal Seizures and a normal Intellectual Development. *Clin. Genet.* 91 (1), 106–110. doi:10.1111/cge.12797

Wang, J., Liu, X., Shen, S., Deng, L., and Liu, H. (2021). DeepDDS: Deep Graph Neural Network with Attention Mechanism to Predict Synergistic Drug Combinations. *Brief. Bioinformatics*. doi:10.1093/bib/bbab390

Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., et al. (2020). Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics. *Nucleic Acids Res.* 48 (D1), D1031–D1041. doi:10.1093/nar/gkz981

Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2020). Comparative Analysis and Prediction of Quorum-sensing Peptides Using Feature Representation Learning and Machine Learning Algorithms. *Brief. Bioinformatics* 21 (1), 106–119.

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a Sequence-Based Predictor Using Effective Feature Representation to Improve the Prediction of Anti-cancer Peptides. *Bioinformatics* 34 (23), 4007–4016. doi:10.1093/bioinformatics/bty451

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. *Ieee/acm Trans. Comput. Biol. Bioinf.* 11 (1), 192–201. doi:10.1109/tcbb.2013.146

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast Prediction of Protein Methylation Sites Using a Sequence-Based Feature Selection Technique. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (4), 1264–1273. doi:10.1109/tcbb.2017.2670558

Wu, X., and Yu, L. (2021). *EPSOL: Sequence-Based Protein Solubility Prediction Using Multidimensional Embedding*. Oxford, England: Bioinformatics.

Wu, Y., Lu, X., Shen, B., and Zeng, Y. (2019). The Therapeutic Potential and Role of miRNA, lncRNA, and circRNA in Osteoarthritis. *Cgt* 19 (4), 255–263. doi:10.2174/1566523219666190716092203

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D.-Q. (2018). PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors from Protein Sequences Using a Stacked Ensemble Method. *Front. Microbiol.* 9, 2571. doi:10.3389/fmicb.2018.02571

Xu, L., Liang, G., Liao, C., Chen, G. D., Chang, C. C., and k-Skip-n-Gram-Rf (2019). K-Skip-N-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification. *Front. Genet.* 10 (33), 33. doi:10.3389/fgene.2019.00033

Xu, L., Liang, G., Wang, L., and Liao, C. (2018). A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides. *Genes* 9 (3), 158. doi:10.3390/genes9030158

Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018). What Contributes to Serotonin-Norepinephrine Reuptake Inhibitors' Dual-Targeting Mechanism? the Key Role of Transmembrane Domain 6 in Human Serotonin and Norepinephrine Transporters Revealed by Molecular Dynamics Simulation. *ACS Chem. Neurosci.* 9 (5), 1128–1140. doi:10.1021/acschemneuro.7b00490

Yang, H., Ding, Y., Tang, J., and Guo, F. (2021). Drug-disease Associations Prediction via Multiple Kernel-Based Dual Graph Regularized Least Squares. *Appl. Soft Comput.* 112, 107811. doi:10.1016/j.asoc.2021.107811

Yang, H., Ding, Y., Tang, J., and Guo, F. (2021). Identifying Potential Association on Gene-Disease Network via Dual Hypergraph Regularized Least Squares. *BMC Genomics* 22 (1), 605. doi:10.1186/s12864-021-07864-z

Yang, L., LiuLiu, B., Huang, B., Deng, J., Li, H., Yu, B., et al. (2013). A Functional Copy Number Variation in the WWOX Gene Is Associated with Lung Cancer Risk in Chinese. *Hum. Mol. Genet.* 22 (9), 1886–1894. doi:10.1093/hmg/ddt019

Yang, Q., Hong, J., Li, Y., Xue, W., Li, S., Yang, H., et al. (2020). A Novel Bioinformatics Approach to Identify the Consistently Well-Performing Normalization Strategy for Current Metabolomic Studies. *Brief. Bioinformatics* 21 (6), 2142–2152. doi:10.1093/bib/bbz137

Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., Li, X., et al. (2020). Consistent Gene Signature of Schizophrenia Identified by a Novel Feature Selection Strategy from Comprehensive Sets of Transcriptomic Data. *Brief. Bioinformatics* 21 (3), 1058–1068. doi:10.1093/bib/bbz049

Yang, Q., Wang, Y., Zhang, Y., Li, F., Xia, W., Zhou, Y., et al. (2020). NOREVA: Enhanced Normalization and Evaluation of Time-Course and Multi-Class Metabolomic Data. *Nucleic Acids Res.* 48 (W1), W436–W448. doi:10.1093/nar/gkaa258

Yin, J., Li, F., Zhou, Y., Mou, M., Lu, Y., Chen, K., et al. (2021). INTEDE: Interactome of Drug-Metabolizing Enzymes. *Nucleic Acids Res.* 49 (D1), D1233–D1243. doi:10.1093/nar/gkaa755

Yin, J., Sun, W., Li, F., Hong, J., Li, X., Zhou, Y., et al. (2020). VARIDT 1.0: Variability of Drug Transporter Database. *Nucleic Acids Res.* 48 (D1), D1042–D1050. doi:10.1093/nar/gkz779

Yu, L., Xia, M., and An, Q. (2021). A Network Embedding Framework Based on Integrating Multiplex Network for Drug Combination Prediction. *Brief. Bioinformatics*. doi:10.1093/bib/bbab364

Yu, L., Wang, M., Yang, Y., Xu, F., Zhang, X., Xie, F., et al. (2021). Predicting Therapeutic Drugs for Hepatocellular Carcinoma Based on Tissue-specific Pathways. *Plos Comput. Biol.* 17 (2), e1008696. doi:10.1371/journal.pcbi.1008696

Zhang, J., Zhang, Z., Pu, L., Tang, J., and Guo, F. (2020). AIEpred: an Ensemble Predictive Model of Classifier Chain to Identify Anti-inflammatory Peptides. *Ieee/acm Trans. Comput. Biol. Bioinform*, 1. doi:10.1109/TCBB.2020.2968419

Zhang, S., Su, M., Sun, Z., Lu, H., and Zhang, Y. (2020). The Signature of Pharmaceutical Sensitivity Based on ctDNA Mutation in Eleven Cancers. *Exp. Biol. Med. (Maywood)* 245 (8), 720–732. doi:10.1177/1535370220906518

Zhang, T., Hu, Y., Wu, X., Ma, R., Jiang, Q., and Wang, Y. (2016). Identifying Liver Cancer-Related Enhancer SNPs by Integrating GWAS and Histone Modification ChIP-Seq Data. *Biomed. Res. Int.* 2016, 2395341. doi:10.1155/2016/2395341

Zhang, T., Wu, Y., Lan, Z., Shi, Q., Yang, Y., Guo, J., et al. (2017). Syntool: A Novel Region-Based Intolerance Score to Single Nucleotide Substitution for Synonymous Mutations Predictions Based on 123,136 Individuals. *Biomed. Res. Int.* 2017, 5096208. doi:10.1155/2017/5096208

Zhang, X., Li, M., Lin, H., Rao, X., Feng, W., Yang, Y., et al. (2017). regSNPs-Splicing: a Tool for Prioritizing Synonymous Single-Nucleotide Substitution. *Hum. Genet.* 136 (Suppl. 9), 1279–1289. doi:10.1007/s00439-017-1783-x

Zhang, Z.-M., Tan, J.-X., Wang, F., Dao, F.-Y., Zhang, Z.-Y., and Lin, H. (2020). Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method. *Front. Bioeng. Biotechnol.* 8, 254. doi:10.3389/fbioe.2020.00254

Zhang, Z.-M., Wang, J.-S., Zulfiqar, H., Lv, H., Dao, F.-Y., and Lin, H. (2020). Early Diagnosis of Pancreatic Ductal Adenocarcinoma by Combining Relative Expression Orderings with Machine-Learning Method. *Front. Cel Dev. Biol.* 8, 582864. doi:10.3389/fcell.2020.582864

Zhou, L.-Y., Qin, Z., Zhu, Y.-H., He, Z.-Y., and Xu, T. (2019). Current RNA-Based Therapeutics in Clinical Trials. *Cgt* 19 (3), 172–196. doi:10.2174/1566523219666190719100526

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Lin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 25 November 2021
doi: 10.3389/fcell.2021.789587

Colon Cancer-Related Genes Identification and Function Study Based on Single-Cell Multi-Omics Integration

Xuepu Sun[1†], Yu Guo[2†], Yu Zhang[3†], Peng Zhao[4], Zhaoqing Wang[4], Zheng Wei[1] and Haiquan Qiao[1*]*

[1]The First Affiliated Hospital of Harbin Medical University, Harbin, China, [2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [3]Department of Neurosurgery, General Hospital of Heilongjiang Province Land Reclamation Bureau, Harbin, China, [4]School of Life Science and Technology, Harbin Institute of Technology, Harbin, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

***Correspondence:**
Zheng Wei
zhengwei202110@163.com
Haiquan Qiao
haiquanqiao668@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *05 October 2021*
Accepted: *01 November 2021*
Published: *25 November 2021*

Citation:
Sun X, Guo Y, Zhang Y, Zhao P, Wang Z, Wei Z and Qiao H (2021) Colon Cancer-Related Genes Identification and Function Study Based on Single-Cell Multi-Omics Integration. Front. Cell Dev. Biol. 9:789587. doi: 10.3389/fcell.2021.789587

Transcriptomes and DNA methylation of colon cancer at the single-cell level are used to identify marker genes and improve diagnoses and therapies. Seven colon cancer subtypes are recognized based on the single-cell RNA sequence, and the differentially expressed genes regulated by dysregulated methylation are identified as marker genes for different types of colon cancer. Compared with normal colon cells, marker genes of different types show very obvious specificity, especially upregulated genes in tumors. Functional enrichment analysis for marker genes indicates a possible relation between colon cancer and nervous system disease, moreover, the weak immune system is verified in colon cancer. The heightened expression of markers and the reduction of methylation in colon cancer promote tumor development in an extensive mechanism so that there is no biological process that can be enriched in different types.

Keywords: colon cancer, single cell, transcriptome, methylation, comprehensive analysis, marker genes

INTRODUCTION

Colorectal cancer is the most common gastrointestinal tumor with high appearance, recurrence, and death. Up to 70% occurs in the colon, and according to the statistics of 36 cancers in 185 countries in 2018, colorectal cancer resulted in 881,000 deaths, of which over 550,000 cases were due to colon cancer. In addition, more than 1,000,000 new colon cancer patients were confirmed during that year, justifying why the screening and treatment of colon cancer are hot issues worldwide (Bailey et al., 2015; Chen et al., 2016; Bray et al., 2018; Fang et al., 2019; Liang et al., 2019; Liu et al., 2020; Hu et al., 2021a). With the development of different kinds of treatment technologies, the 5-year survival rate of early-stage colon cancer patients has increased up to 90%, however, many urgent problems remain to be solved: for example, the median survival time of colon cancer patients with metastasis is less than 2 years, and 5-year overall survival rate is only 50–65%. Early screening and diagnosis for colon cancer mainly depend on electronic colorectal mirror surgery and pathology, which is expensive and has poor prognosis (Kiriyama et al., 2012; Chiba et al., 2014; Missiaglia et al., 2014; Fakih, 2015; Miller et al., 2016; Sewda et al., 2016; Marmol et al., 2017; White et al., 2017; Zhao et al., 2017; Hu et al., 2020; Mo et al., 2020; Hu et al., 2021a; Hu et al., 2021b). As a result, molecular biomarkers are vital for verifying the mechanism that promotes colon cancer development and improving early screening, diagnosis, and treatment for colon cancer.

More and more research has revealed that abnormal epigenetic modifications contribute to the occurrence and development of colon cancer by changing the expression of genes associated with tumors. This has meant that the stable and critical epigenetic modification style, methylation, has become most viable candidate for colon cancer biomarker research (Cheng et al., 2018; Hong, 2018; Tang et al., 2018; Wang et al., 2019). The abnormal DNA methylation pattern associated with colon cancer has been verified (Simmer et al., 2012; Ao et al., 2019; Zuo et al., 2020). High and low methylated CpG sites have been identified using the data of Illumina Human Methylation 450K (Naumov et al., 2013). The Wnt signaling pathway (Novellasdemunt et al., 2015; Sheng et al., 2021), Rap1 signaling (Zhang et al., 2017), and many other signaling pathways have been confirmed to have an effect on colon cancer, at the same time, *GDNF* (Zhong et al., 2018), *BMP3* (Park et al., 2016), and a few other genes have been regarded as the signs of malignant colon cancer, while *NEU1* (Uemura et al., 2009) as well as others suppress metastasis.

A few genes have been recognized as the marker genes for colon cancer, however, the number of marker genes is not enough for choosing efficient target genes, and tumor heterogeneity has had an influence on the result. Considering the reasons mentioned above, in this study, we attempt to identify marker genes for colon cancer by the comprehensive analysis of transcriptomes and methylation at the single-cell level, and conduct functional enrichment analysis for the makers.

METHODS

Dataset

The data used in this research are from a previous study (Bian et al., 2018) and are the **Supplementary Material** of the accession number GSE97693 in the Gene Expression Omnibus (GEO) database (Barrett et al., 2005) (https://www.ncbi.nlm.nih.gov/geo/).

Data Preprocessing

Colon cancer cells which possess transcriptomes with methylomics and normal cells are selected for further analysis. The genes expressed in all cells are reserved in the analysis for RNA, in contrast, because of the poor quality of methylation sequencing, the genes methylated in over 50% of normal cells and cancer cells are reserved for further methylation analysis. The downloaded transcriptomes are normalized to FPKM or TPM, so the data are renormalized, and the expression level of each gene is represented uniformly by the expression level of each gene taking up the percentage of the average expression level of all genes in the cell. The DNA methylation level is represented by the average methylation level of CpG loci in the promoter region.

Differentially Expressed Genes

Differentially expressed genes are defined by p-value <0.01 [Wilcox test (Hiller, 1988)] and $|\log2(FC)|>2$ in this study. This experiment includes two parts: comparison between normal cells and tumor cells, and comparison between normal cells and different types of colon cancer cells. According to RNA, colon cancer cells are clustered into subtypes by Seurat (Butler et al., 2018) (version 3.0.1), which is a package in R specially designed for single-cell transcriptomic data across different conditions, technologies, and species.

Differentially Methylated Genes

In this study, the differentially methylated genes are defined by the p-value threshold of the Wilcox test (Hiller, 1988) which is 0.01 and the absolute value of difference threshold which is 0.2. This module includes two comparisons: normal cells versus colon cancer cells and normal colon cells versus different types of cancer cells. The subtypes are consistent with the result of transcriptomic data clusters by default.

Differentially Expressed and Methylated Genes

According to previous studies, DNA methylation is an important factor of changing the expression of genes. Gene expression is usually negatively regulated by the DNA methylation level. So the methylation and transcriptome genes with negative correlations are recognized as the marker genes for colon cancer as well as subtypes.

Functional Enrichment Analysis

Functional enrichment analysis is performed for the marker genes to study their mechanism of affecting colon cancer development via the package in R called ClusterProfiler (version 3.12.0) (Yu et al., 2012). The R package ClusterProfiler is designed for comparing biological themes among genes clusters, such as those in Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and DO.

RESULTS

Subtypes of Colon Cancer Cells

Based on the Seurat R package, principle component analysis (PCA) of colon cancer cell transcriptomes is performed and the top five principle components are picked for later t-distributed stochastic neighbor embedding (t-SNE) (**Figures 1A,B**). A total of 607 colon cancer cells are clustered into 7 subtypes (**Figure 1C**). Contrasting with all other subtypes, differentially expressed genes in each subtype are identified and the top 10 cluster the colon cancer cells very well (**Figure 1D**; **Table 1**).

Study for Differentially Expressed Genes

The Wilcox test and fold change test are carried out between the 607 colon cancer cells and 20 normal cells with transcriptomes. In total, 331 differentially expressed genes are identified, 146 of which are upregulated in colon cancer (**Figures 2A,B**). To remove the influence of tumor heterogeneity, comparisons between normal cells and each subtype are performed. The result shows there are more differently expressed genes in each subtype in general, especially the genes of upregulated expression in tumors (**Figures 2A,B**), furthermore, variation in

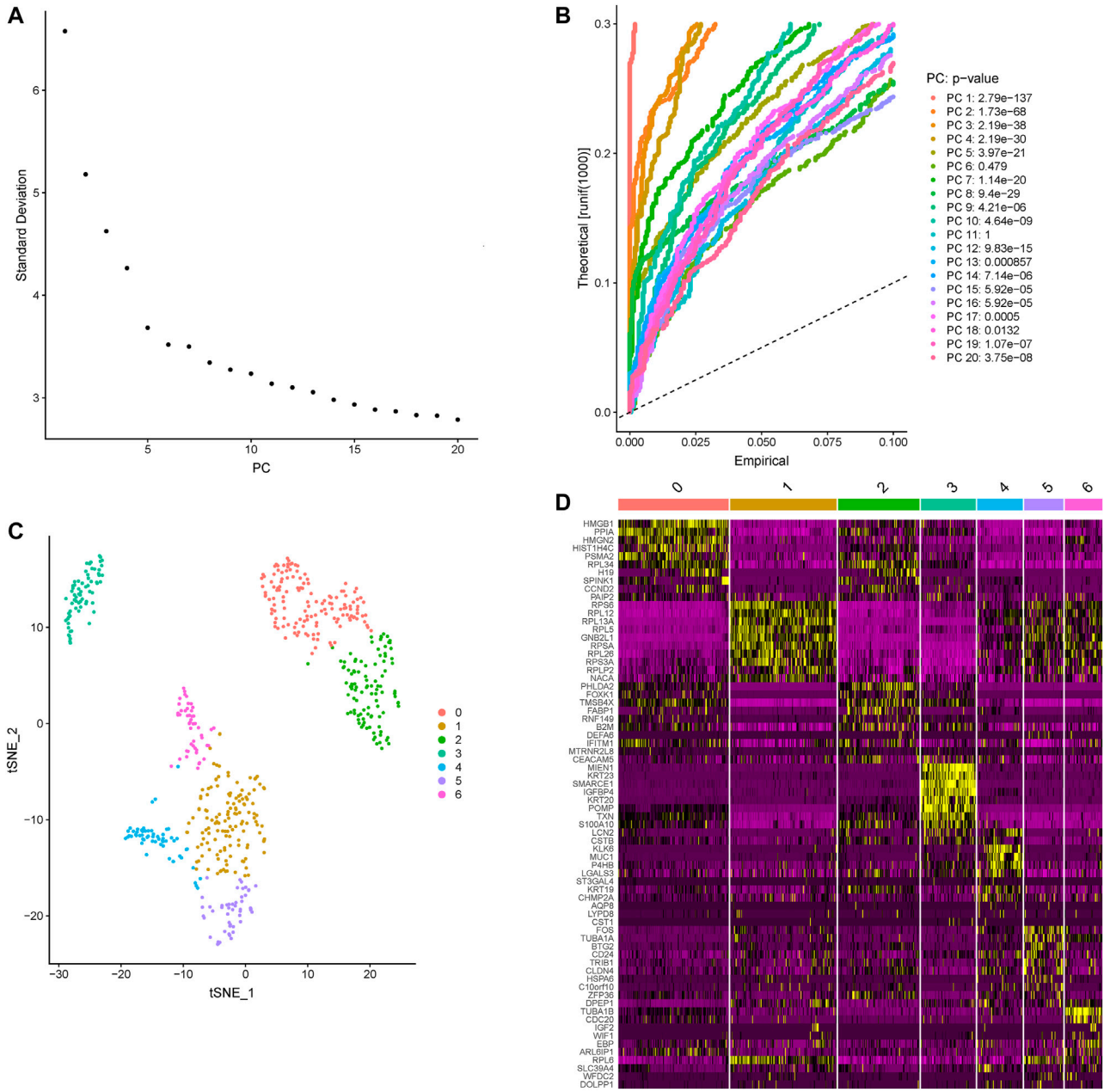


FIGURE 1 | The subtypes clustered across all colon cancer cells. **(A)** The standard deviations of the top 20 principle components are shown, and the elbow corresponds well with the significant dims. **(B)** Statistical significance of the PCA scores of the top 20 PCs, the *p*-values are shown on the right. **(C)** The clusters of colon cancer are marked by color. **(D)** The heatmap of the top 10 differentially expressed genes in each subtype, yellow and purple represent high and low expression separately, the subtypes are marked by number.

the amount of upregulated genes in subtypes is significant, this suggests that the genes whose expression is upregulated in tumors are significantly affected by subtypes.

At the same time, functional enrichment analysis indicates the genes of upregulated expression in colon cancer cells are enriched in a variety of biological processes, besides this, these processes are very different between subtypes. The findings above verify the diversities between subtypes and the different mechanisms of colon cancer development in subtypes (**Supplementary Figure S1**; **Table 2**). However, the genes of downregulated expression in colon cancer are enriched in fewer processes and for some subtypes, the enriched biological processes are extremely similar (**Supplementary Figure S2**; **Table 2**). The notable result in the KEGG function analysis shows that a few downregulated expression genes in some subtypes are enriched in Alzheimer disease,

TABLE 1 | The top 10 differentially expressed genes in each subtype.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
HMGB1	RPS6	PHLDA2	MIEN1	KLK6	FOS	TUBA1B
PPIA	RPL12	FOXK1	KRT23	MUC1	TUBA1A	CDC20
HMGN2	RPL13A	TMSB4X	SMARCE1	P4HB	BTG2	IGF2
HIST1H4C	RPL5	FABP1	IGFBP4	LGALS3	CD24	WIF1
PSMA2	GNB2L1	RNF149	KRT20	ST3GAL4	TRIB1	EBP
RPL34	RPSA	B2M	POMP	KRT19	CLDN4	ARL6IP1
H19	RPL26	DEFA6	TXN	CHMP2A	HSPA6	RPL6
SPINK1	RPS3A	IFITM1	S100A10	AQP8	C10orf10	SLC39A4
CCND2	RPLP2	MTRNR2L8	LCN2	LYPD8	ZFP36	WFDC2
PAIP2	NACA	CEACAM5	CSTB	CST1	DPEP1	DOLPP1

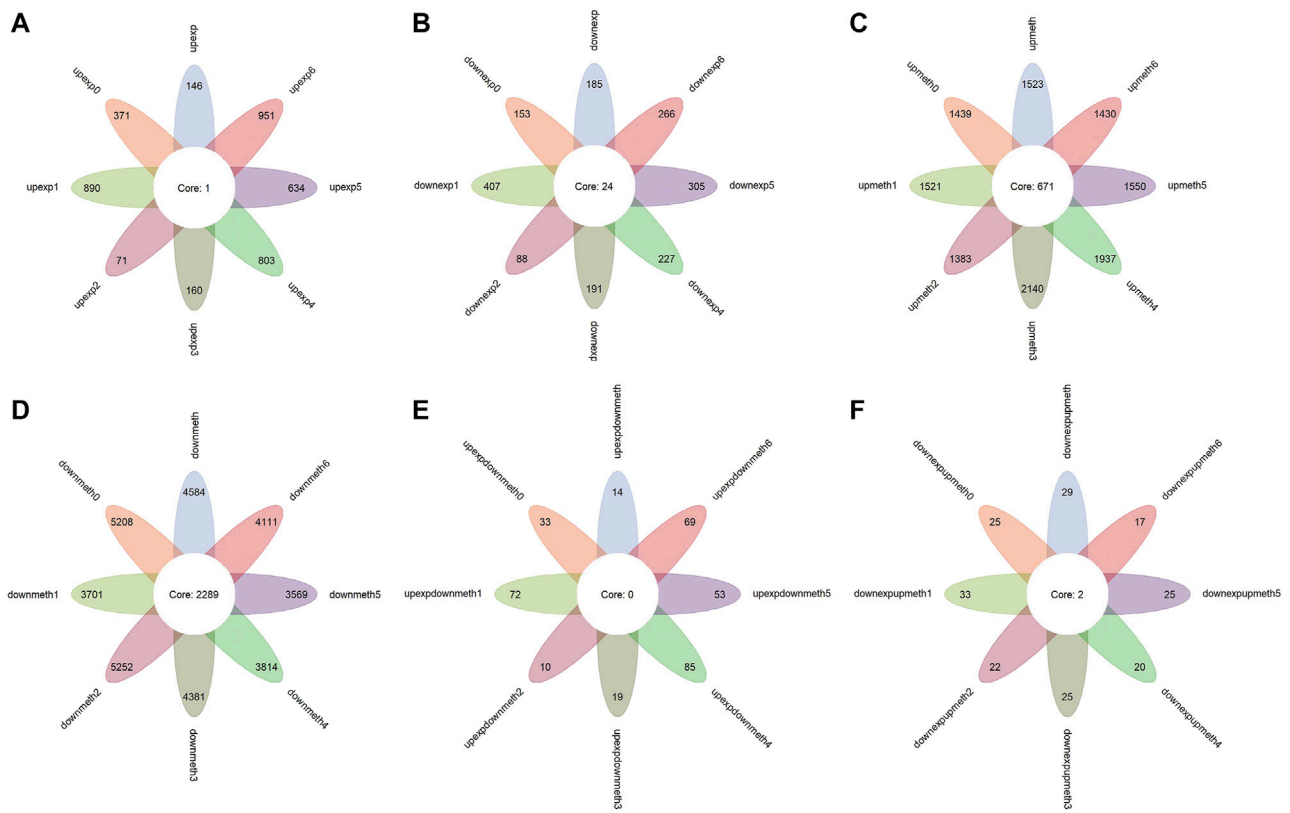


FIGURE 2 | Functional enrichment analysis for dysregulated genes. The color of the point means the significance (*p*-value). The size of the point means the gene number. **(A)** The biological process enrichment of upregulated expressed genes in colon cancer and subtypes. **(B)** The biological process enrichment of downregulated expressed genes in colon cancer and subtypes. **(C)** The KEGG pathway enrichment of downregulated expressed genes in colon cancer and subtypes. **(D)** The biological process enrichment of upregulated methylated genes in colon cancer and subtypes. **(E)** The biological process enrichment of downregulated methylated genes in colon cancer and subtypes. **(F)** The KEGG pathway of downregulated methylated genes in colon cancer and subtypes. The biological process enrichment of upregulated expressed and downregulated methylated genes in colon cancer and subtypes. The biological process enrichment of downregulated expressed and upregulated methylated genes in colon cancer and subtypes.

Huntington disease, and Parkinson disease, which reminds us of the possible relationship between the colon and nervous system disease (**Supplementary Figure S3**; **Table 2**).

Study of Differentially Methylated Genes

The *p*-value of the Wilcox test and the absolute value of difference between normal cells and colon cancer cells are verified. Overall, 1,523 genes of upregulated methylation and 4,584 genes of downregulated methylation are identified (**Figures 2C,D**). Comparing normal colon cells and colon cancer cells of different types, the result shows that a similar number of differentially methylated genes are recognized in subtypes (**Figures 2C,D**), which suggests the pattern of dysregulated methylation in colon cancer is similar.

TABLE 2 | The enriched term number of function analysis for dysregulated genes.

		Colon cancer	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Biological process	Upexp	3	209	86	42	36	156	0	321
	Downexp	13	13	151	24	11	110	161	143
	upmeth	27	32	1	1	157	139	0	8
	downmeth	17	27	11	32	26	26	10	14
	upexpdownmeth	1	8	0	0	7	0	1	1
	downexpupmeth	25	9	40	39	3	0	73	34
KEGG	downexp	4	5	34	2	5	21	20	15
	downmeth	4	9	1	9	9	6	1	3

In addition, the biological process functional enrichment shows that the genes of upregulated methylation in colon cancer mainly act on two fields: first, influencing different signaling transmission using different means, such as the neuroactive ligand–receptor interaction, calcium signaling pathway, pathways verified to be related to colon cancer, for example, the cAMP signaling pathway (Leiphrakpam et al., 2018), and cell adhesion molecules (CAMs) (Hara et al., 2019); second, cell–cell adhesion *via* plasma–membrane adhesion molecules, which is enriched in almost all subtypes. Another important finding is that some marker genes have an effect on cognition, learning or memory, and regulation of the neurological system process, which also suggests that colon cancer may be related to nervous system disease, such as Alzheimer disease and Parkinson disease (**Supplementary Figure S4**). In both biological process and KEGG pathways, the functions of downregulated methylation genes in tumors are highly consistent, including the perception of smell, cytokine–cytokine receptor interaction, and a few others (**Supplementary Figures S5, 6**).

Study of Marker Genes

In light of existing studies, abnormal methylation change leading to dysregulated expression of genes is a vital early event in the development of certain kinds of cancers, and, in general, RNA and methylation have a negative relationship. Therefore, 14 genes with upregulated expression together with downregulated methylation and 29 genes with downregulated expression as well as upregulated methylation are identified as marker genes for colon cancer (**Figures 2E,F**). The number of upregulated expression and downregulated methylation genes in subtypes has a large fluctuation, but the number of downregulated expression and upregulated methylation genes in subtypes is almost same (**Figures 2E,F**), which is consistent with the result of differentially expressed genes. In addition, this indicates that heterogeneity in colon cancer is more reflected in genes whose expression is upregulated.

In function analysis, upregulated expression and downregulated methylation genes promote colon cancer in extensive mechanisms so that there is no term enriched in some subtypes, moreover, the mechanisms are different in subtypes (**Supplementary Figure S7**). The functions of downregulated expression and upregulated methylation genes are mainly associated with immunity, such as innate immune response–activating signal transduction, activation of innate immune response, neutrophil activation involved in immune response, and negative regulation of interleukin−1 secretion (**Supplementary Figure S8**), which is in line with our study.

DISCUSSION

According to existing studies, some genes and pathways have been confirmed to be related to colon cancer. However, those studies are mostly dependent on the data of bulk sample sequencing. Therefore, in this study, the single-cell transcriptomes and methylation of colon cancer are used to remove the influence of tumor heterogeneity. Differentially expressed and methylated genes are identified as marker genes for subtypes. Biological process analysis identifies several processes, which suggest the relationship between colon cancer and diseases of the nervous system (Alzheimer disease, Parkinson disease). In addition, the intra-tumor heterogeneity of colon cancer is more reflected in upregulated expressed genes. Clinically, this finding provides possible targeted therapeutic approaches for colon cancer patients. However, the small number of cells and the high intra-tumor heterogeneity mean that the findings need to be verified in larger cohorts.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/geo/, GSE97693.

AUTHOR CONTRIBUTIONS

XS wrote the paper and did the experiments. YG and YZ provided ideas of this work. ZW, PZ and ZW revised this manuscript and guided how to do experiments. HQ supervised this work. All authors contributed to the article and approved the submitted version.

FUNDING

This project is supported by the Youth program of the National Natural Science Foundation of China (no. 61801147).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.789587/full#supplementary-material

Supplementary Figure S1 | The results of biological process enrichment analysis for upregulated expressed genes in colon cancer and subtypes.

Supplementary Figure S2 | The results of biological process enrichment analysis for downregulated expressed genes in colon cancer and subtypes.

Supplementary Figure S3 | The results of KEGG pathway enrichment analysis for downregulated expressed genes in colon cancer and subtypes.

Supplementary Figure S4 | The results of biological process enrichment analysis for upregulated methylated genes in colon cancer and subtypes.

Supplementary Figure S5 | The results of biological process enrichment analysis for downregulated methylated genes in colon cancer and subtypes.

Supplementary Figure S6 | The results of KEGG pathway enrichment analysis for upregulated methylated genes in colon cancer and subtypes.

Supplementary Figure S7 | The results of biological process enrichment analysis for upregulated expressed and downregulated methylate genes in colon cancer and subtypes.

Supplementary Figure S8 | The results of biological process enrichment analysis for downregulated expressed and upregulated methylated genes in colon cancer and subtypes.

REFERENCES

Ao, C., Jin, S., Lin, Y., and Zou, Q. (2019). Review of Progress in Predicting Protein Methylation Sites. *Coc* 23, 1663–1670. doi:10.2174/1385272823666190723141347

Bailey, C. E., Hu, C.-Y., You, Y. N., Bednarski, B. K., Rodriguez-Bigas, M. A., Skibber, J. M., et al. (2015). Increasing Disparities in the Age-Related Incidences of colon and Rectal Cancers in the United States, 1975-2010. *JAMA Surg.* 150, 17–22. doi:10.1001/jamasurg.2014.1756

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., et al. (2005). NCBI GEO: Mining Millions of Expression Profiles-Ddatabase and Tools. *Nucleic Acids Res.* 33, D562–D566. doi:10.1093/nar/gki022

Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., et al. (2018). Single-cell Multiomics Sequencing and Analyses of Human Colorectal Cancer. *Science* 362, 1060–1063. doi:10.1126/science.aao3791

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A. Cancer J. Clinic.* 68, 394–424. doi:10.3322/caac.21492

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species. *Nat. Biotechnol.* 36, 411–420. doi:10.1038/nbt.4096

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer Statistics in China, 2015. *CA: A Cancer J. Clinic.* 66, 115–132. doi:10.3322/caac.21338

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics* 34, 1953–1956. doi:10.1093/bioinformatics/bty002

Chiba, H., Takahashi, A., Inamori, M., Goto, T., Ohata, K., Matsuhashi, N., et al. (2014). Early colon Cancer Presenting as Intussusception and Successfully Treated Using Endoscopic Submucosal Dissection. *Endoscopy* 46, E326–E327. doi:10.1055/s-0034-1377224

Fakih, M. G. (2015). Metastatic Colorectal Cancer: Current State and Future Directions. *J. Clin. Oncol.* 33, 1809–1824. doi:10.1200/jco.2014.59.7633

Fang, S., Pan, J., Zhou, C., Tian, H., He, J., Shen, W., et al. (2019). Circular RNAs Serve as Novel Biomarkers and Therapeutic Targets in Cancers. *Curr. Gene Ther.* 19, 125–133. doi:10.2174/1566523218666181109142756

Hara, M., Takeba, Y., Iiri, T., Ohta, Y., Ootaki, M., Watanabe, M., et al. (2019). Vasoactive Intestinal Peptide Increases Apoptosis of Hepatocellular Carcinoma by Inhibiting the cAMP/Bcl-xL Pathway. *Cancer Sci.* 110, 235–244. doi:10.1111/cas.13861

Hiller, E. (1988). Statistical Comparison of Multiple Dependent Samples-Tthe Friedman Test and the Wilcoxon and Wilcox Dij Test. *Z. Arztl Fortbild (Jena)* 82, 885–886.

Hong, S. N. (2018). Genetic and Epigenetic Alterations of Colorectal Cancer. *Intest Res.* 16, 327–337. doi:10.5217/ir.2018.16.3.327

Hu, Y., Qiu, S., and Cheng, L. (2021a). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 7036592. doi:10.1155/2021/7036592

Hu, Y., Sun, J.-y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (2021b). rs1990622 Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19, 11. doi:10.1186/s12916-020-01883-5

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 Regulates TSPAN13 Expression and Contributes to Alzheimer's Disease with Sex Differences. *Brain* 143, e95. doi:10.1093/brain/awaa302

Kiriyama, S., Saito, Y., Yamamoto, S., Soetikno, R., Matsuda, T., Nakajima, T., et al. (2012). Comparison of Endoscopic Submucosal Dissection with Laparoscopic-Assisted Colorectal Surgery for Early-Stage Colorectal Cancer: a Retrospective Analysis. *Endoscopy* 44, 1024–1030. doi:10.1055/s-0032-1310259

Leiphrakpam, P. D., Brattain, M. G., Black, J. D., and Wang, J. (2018). TGFβ and IGF1R Signaling Activates Protein Kinase A through Differential Regulation of Ezrin Phosphorylation in colon Cancer Cells. *J. Biol. Chem.* 293, 8242–8254. doi:10.1074/jbc.ra117.001299

Liang, Y., Zhang, C., and Dai, D.-Q. (2019). Identification of Differentially Expressed Genes Regulated by Methylation in colon Cancer Based on Bioinformatics Analysis. *World J. Gastroenterol.* 25, 3392–3407. doi:10.3748/wjg.v25.i26.3392

Liu, W., Haran, J. P., Ash, A. S., Allison, J. J., Ye, S., Tjia, J., et al. (2020). High-dimensional Causal Mediation Analysis with a Large Number of Mediators Clumping at Zero to Assess the Contribution of the Microbiome to the Risk of Bacterial Pathogen Colonization in Older Adults. *Curr. Bioinform* 15, 671–696. doi:10.2174/1574893614666191115123219

Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., and Rodriguez Yoldi, M. J. (2017). Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *Int. J. Mol. Sci.* 18, 197. doi:10.3390/ijms18010197

Miller, K. D., Siegel, R. L., Lin, C. C., Mariotto, A. B., Kramer, J. L., Rowland, J. H., et al. (20162016). Cancer Treatment and Survivorship Statistics, 2016. *CA: A Cancer J. Clinic.* 66, 271–289. doi:10.3322/caac.21349

Missiaglia, E., Jacobs, B., D'ario, G., Di Narzo, A. F., Soneson, C., Budinska, E., et al. (2014). Distal and Proximal colon Cancers Differ in Terms of Molecular, Pathological, and Clinical Features. *Ann. Oncol.* 25, 1995–2001. doi:10.1093/annonc/mdu275

Mo, F., Luo, Y., Fan, D., Zeng, H., Zhao, Y., Luo, M., et al. (2020). Integrated Analysis of mRNA-Seq and miRNA-Seq to Identify C-MYC, YAP1 and miR-3960 as Major Players in the Anticancer Effects of Caffeic Acid Phenethyl Ester in Human Small Cell Lung Cancer Cell Line. *Curr. Gene Ther.* 20, 15–24. doi:10.2174/1566523220666200523165159

Naumov, V. A., Generozov, E. V., Zaharjevskaya, N. B., Matushkina, D. S., Larin, A. K., Chernyshov, S. V., et al. (2013). Genome-scale Analysis of DNA Methylation in Colorectal Cancer Using Infinium HumanMethylation450 BeadChips. *Epigenetics* 8, 921–934. doi:10.4161/epi.25577

Novellasdemunt, L., Antas, P., and Li, V. S. W. (2015). Targeting Wnt Signaling in Colorectal Cancer. A Review in the Theme: Cell Signaling: Proteins, Pathways and Mechanisms. *Am. J. Physiol. Cel Physiol.* 309, C511–C521. doi:10.1152/ajpcell.00117.2015

Park, S.-K., Song, C. S., Yang, H.-J., Jung, Y. S., Choi, K. Y., Koo, D. H., et al. (2016). Field Cancerization in Sporadic colon Cancer. *Gut and Liver* 10, 773–780. doi:10.5009/gnl15334

Sewda, K., Coppola, D., Enkemann, S., Yue, B., Kim, J., Lopez, A. S., et al. (2016). Cell-surface Markers for colon Adenoma and Adenocarcinoma. *Oncotarget* 7, 17773–17789. doi:10.18632/oncotarget.7402

Sheng, Y., Jiang, Y., Yang, Y., Li, X., Qiu, J., Wu, J., et al. (2021). CNA2Subpathway: Identification of Dysregulated Subpathway Driven by Copy Number Alterations in Cancer. *Brief Bioinform.* 22, bbaa413. doi:10.1093/bib/bbaa413

Simmer, F., Brinkman, A. B., Assenov, Y., Matarese, F., Kaan, A., Sabatino, L., et al. (2012). Comparative Genome-wide DNA Methylation Analysis of Colorectal Tumor and Matched normal Tissues. *Epigenetics* 7, 1355–1367. doi:10.4161/epi.22562

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor Origin Detection with Tissue-specific miRNA and DNA Methylation Markers. *Bioinformatics* 34, 398–406. doi:10.1093/bioinformatics/btx622

Uemura, T., Shiozaki, K., Yamaguchi, K., Miyazaki, S., Satomi, S., Kato, K., et al. (2009). Contribution of Sialidase NEU1 to Suppression of Metastasis of Human colon Cancer Cells through Desialylation of Integrin β4. *Oncogene* 28, 1218–1229. doi:10.1038/onc.2008.471

Wang, Q., Yang, M., Zhang, Y., Zhong, L., and Zheng, X. (2019). Novel Combination Oncolytic Adenoviral Gene Therapy Armed with Dm-dNK and CD40L for Breast Cancer. *Curr. Gene Ther.* 19, 54–65. doi:10.2174/1566523219666190307094713

White, A., Joseph, D., Rim, S. H., Johnson, C. J., Coleman, M. P., and Allemani, C. (2017). Colon Cancer Survival in the United States by Race and Stage (2001-2009): Findings from the CONCORD-2 Study. *Cancer* 123 (Suppl. 24), 5014–5036. doi:10.1002/cncr.31076

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zhang, Y. L., Wang, R. C., Cheng, K., Ring, B. Z., and Su, L. (2017). Roles of Rap1 Signaling in Tumor Cell Migration and Invasion. *Cancer Biol. Med.* 14, 90–99. doi:10.20892/j.issn.2095-3941.2016.0086

Zhao, T., Zhang, N., Zhang, Y., Ren, J., Xu, P., Liu, Z., et al. (2017). A Novel Method to Identify Pre-microRNA in Various Species Knowledge Base on Various Species. *J. Biomed. Semant* 8, 30. doi:10.1186/s13326-017-0143-z

Zhong, F., Zhang, W., Cao, Y., Wen, Q., Cao, Y., Lou, B., et al. (2018). LncRNA NEAT1 Promotes Colorectal Cancer Cell Proliferation and Migration via Regulating Glial Cell-Derived Neurotrophic Factor by Sponging miR-196a-5p. *Acta Biochim. Biophys. Sin.* 50, 1190–1199. doi:10.1093/abbs/gmy130

Zuo, Y., Song, M., Li, H., Chen, X., Cao, P., Zheng, L., et al. (2020). Analysis of the Epigenetic Signature of Cell Reprogramming by Computational DNA Methylation Profiles. *Curr. Bioinform.* 15, 589–599. doi:10.2174/1574893614666190919103752

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a shared affiliation with the authors XS, ZW, and HQ at the time of the review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Sun, Guo, Zhang, Zhao, Wang, Wei and Qiao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 29 November 2021
doi: 10.3389/fcell.2021.758776

Comprehensive Analysis of Subtype-Specific Molecular Characteristics of Colon Cancer: Specific Genes, Driver Genes, Signaling Pathways, and Immunotherapy Responses

Fangjie Hu[1†], Jianyi Wang[2†], Minghui Zhang[3†], Shuoshuo Wang[2†], Lingyu Zhao[2†], Hao Yang[4], Jinrong Wu[5*] and Binbin Cui[1*]*

[1] Department of Colorectal Surgery, Harbin Medical University Cancer Hospital, Harbin Medical University, Harbin, China, [2] Department of Pathology, Harbin Medical University, Harbin, China, [3] Department of Oncology, Chifeng City Hospital, Chifeng, China, [4] Department of Radiation Oncology, Inner Mongolia Cancer Hospital & Affiliated People's Hospital of Inner Mongolia Medical University, Hohhot, China, [5] Department of Anaesthesiology, The First Affiliated Hospital of Harbin Medical University, Harbin, China

Colon cancer is a complex, heterogeneous disease. The Colorectal Cancer Subtyping Consortium reported a novel classification system for colon cancer in 2015 to better understand its heterogeneity. This molecular classification system divided colon cancer into four distinct consensus molecular subtypes (CMS 1, 2, 3, and 4). However, the characteristics of different colon cancer molecular subtypes have not been fully elucidated. This study comprehensively analyzed the molecular characteristics of varying colon cancer subtypes using multiple databases and algorithms, including The Cancer Genome Atlas (TCGA) database, DriverDBv3 database, CIBERSORT, and MCP-counter algorithms. We analyzed the alterations in the subtype-specific genes of different colon cancer subtypes, such as the RNA levels and DNA alterations, and showed that specific subtype-specific genes significantly affected prognosis. We also explored the changes in colon cancer driver genes and representative genes of 10 signaling pathways in different subtypes. We identified genes that were altered in specific subtypes. We further detected the infiltration of 22 immune cell types in four colon cancer subtypes and the infiltration level of primary immune cells among these subtypes. Additionally, we explored changes in immune checkpoint genes (ICGs) and immunotherapy responses among different colon cancer subtypes. This study may provide clues for the molecular mechanism of tumorigenesis and progression in colon cancer. It also offers potential biomarkers and targets for the clinical diagnosis and treatment of different colon cancer subtypes.

Keywords: colon cancer, molecular subtypes, driver gene, immune infiltration, immune checkpoint

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Zhiguang Li,
National Institutes of Health (NIH), United States
Guomin Shen,
Henan University of Science and Technology, China

****Correspondence:***
Binbin Cui
13351112888@163.com
Jinrong Wu
wujinrong@hrbmu.edu.cn
Hao Yang
haoyang050201@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *15 August 2021*
Accepted: *12 October 2021*
Published: *29 November 2021*

Citation:
Hu F, Wang J, Zhang M, Wang S, Zhao L, Yang H, Wu J and Cui B (2021) Comprehensive Analysis of Subtype-Specific Molecular Characteristics of Colon Cancer: Specific Genes, Driver Genes, Signaling Pathways, and Immunotherapy Responses. Front. Cell Dev. Biol. 9:758776. doi: 10.3389/fcell.2021.758776

INTRODUCTION

Colon cancer is the third most common cancer and second leading cause of cancer-related death globally (Sung et al., 2021). Clinically, colon cancer is primarily classified according to its histopathological features, including tumor size, grade, and disease stage. However, this classification method ignores the heterogeneity of colon cancer. The molecular characteristics of colon cancer tissues of the same pathological type may show significant differences, preventing traditional pathological classifications from accurately distinguishing the biological aspects of colon cancer. From 2012 to 2014, six teams reported representative molecular subtype systems (Schlicker et al., 2012; Budinska et al., 2013; De Sousa et al., 2013; Marisa et al., 2013; Sadanandam et al., 2013). However, because of differences in patient cohorts, sequencing platforms, bioinformatics analysis methods, and data analysis in different studies, each subtype system has different interpretations of the molecular classification of colon cancer. In 2015, based on six subtyping systems, the Colorectal Cancer Subtyping Consortium identified four consensus molecular subtypes (CMS1, CMS2, CMS3, and CMS4), providing the most robust classification system for colon cancer to date (Guinney et al., 2015). The implementation of molecular classification in the clinical decision-making of colon cancer is crucial to solve various clinical problems in colon cancer progression. Presently, different treatment strategies can be formulated according to different molecular classifications in the clinic, including applying microsatellite instability (MSI) and BRAF mutations to predict clinical treatment and prognosis (Van Cutsem et al., 2016; Benson et al., 2018, 2020).

The present study explored the molecular characteristics of four colon cancer subtypes, such as the RNA level, DNA level, and immune infiltration level. We also explored the differences in driver genes and representative genes of 10 signaling pathways in the different colon cancer subtypes and assessed their influence on patient prognosis. Additionally, we evaluated the immune therapy response of four subtypes, including the immune checkpoint gene expression level and immunophenoscore. We aimed to explore the potential differences between different subtypes and identify potential biological therapeutic targets to provide new ideas for the clinical treatment of colon cancer.

MATERIALS AND METHODS

Data Source

We downloaded colon cancer data (41 normal samples and 471 colon cancer samples) from the UCSC Xena database[1] across the following four genomic platforms: RNA expression, gene mutations, copy number variation, and single nucleotide polymorphisms (SNPs).

Classifier Analysis of Colon Cancer

The CMSclassifier developed by Bionetworks was applied to classify the four molecular subtypes of colon cancer samples—CMS1, CMS2, CMS3, and CMS4 (Guinney et al., 2015). The DESeq R package was used to analyze the differential expression of mRNAs, lncRNAs, and miRNAs between different colon cancer subtypes, and a p-value < 0.01 was considered statistically significant (Anders and Huber, 2010). Differentially expressed RNAs were identified in cancer samples and normal samples. Next, each subtype-specific RNA was identified by calculating the differential expression of each gene between a certain subtype and other subtypes and considering the intersection of the differentially expressed genes between every two groups as the gene specifically expressed by the subtype. For example, the analysis of the differences was performed between CMS1 and CMS2, between CMS1 and CMS3, and between CMS1 and CMS4. Next, the intersection of genes with significant differences was finally defined as specific RNA of CMS1. Regarding the cutoff parameters of the differentially expressed genes, a p-value < 0.01 and $\log_2$FC > 1.5 or $\log_2$FC < -1.5 were defined as significantly different genes. The screening methods for specifically expressed lncRNAs and miRNAs were consistent with those described above. Hierarchical clustering analysis was conducted using the R package pheatmap. Each cluster was calculated using the average expression level of subtype-specific RNAs in different subtypes. Metascape[2] was used to perform gene ontology (GO) analysis.

Prognostic Analysis of the Four Colon Cancer Subtypes

Kaplan-Meier survival curves were used to explore the potential link between subtype-specific genes and OS. The median was used as the cutoff for high or low expression of subtype-specific RNAs.

Analysis of Driver Genes in Each Subtype

Fifteen recognized driver gene prediction algorithms were used to identify the driver genes using the DriverDBv3 database[3] (Liu et al., 2020). The driver gene represents a gene detected by more than five algorithms. The expression levels of the driver genes were subjected to variance analysis, and we analyzed the alterations in driver genes in different subtypes.

DNA Alterations Among the Four Subtypes

We analyzed the DNA alterations (copy number variation, gene mutations, and SNPs) of different genes in four colon cancer subtypes, such as subtype-specific RNA, driver genes, and representative genes from 10 oncogenic pathways.

Tumor Mutation Burden of Different Colon Cancer Subtypes

The tumor mutation burden (TMB) is the total number of mutations per million bases in tumor tissue, such as somatic gene coding errors, base substitutions, insertions, and deletions (Lawlor et al., 2021). The TMB data of colon cancer were downloaded from the TCGA database. The colon cancer samples of each subtype were divided into low- and high-TMB groups

[1]https://xenabrowser.net/datapages/

[2]http://metascape.org/gp/index.html

[3]http://ngs.ym.edu.tw/driverdb

based on the median values. We compared the difference in survival between the low- and high-TMB groups of each subtype using Kaplan–Meier analysis, and the *p*-value was calculated using the log-rank test.

Representative Gene Analysis in Classic Pan-Cancer Pathways

The classic pan-cancer atlas was downloaded from PathwayMapper[4] and included 10 pathways: (1) cell cycle, (2) Hippo, (3) Myc, (4) Notch, (5) NRF2, (6) PI3K, (7) RTK/RAS, (8) TGF-β, (9) TP53, and (10) Wnt signaling (Bahceci et al., 2017). All mutation data of colon cancer samples were obtained from the cBioPortal database[5]. We evaluated the gene mutations and copy number variations (CNVs) of all essential representative genes in the 10 pathways. The sum of gene mutation and copy number variation was identified as the altered frequency of representative genes in the pathway.

Analysis of Immune Cell and Fibroblast Infiltration Levels

CIBERSORT and MCP-counter algorithms were used to predict the levels of infiltrating immune cells and fibroblasts in colon cancer samples (Newman et al., 2015; Becht et al., 2016). Pearson correlation analysis was used to assess the correlation between immune cells and fibroblasts.

Immune Therapy Score of Different Colon Cancer Subtypes

We used the immunophenoscore (IPS) from The Cancer Immunome Atlas to assess the immunotherapy response among the four subtypes of colon cancer (Charoentong et al., 2017). The immunophenoscore score primarily comprises four parts: MHC molecules (MHCs), effector cells (ECs), immune checkpoints (CPs), and immunosuppressive cells (SCs). Based on the expression of representative genes or gene sets of immune imaging, the IPS was calculated on a scale of 0–10. A higher IPS indicates more immunogenic tumors and a well-predicted response to anti-PD-1/PD-L1 therapy. The R code used is available at GitHub.[6]

Immune Checkpoint Gene Analysis in Colon Cancer

Differential expression analysis of the immune checkpoint genes (ICGs) in each subtype of colon cancer was performed using the DESeq R package. A p-value < 0.01 was considered statistically significant. The samples of each subtype were divided into low and high expression groups according to the median values. Kaplan–Meier analysis was used to compare the difference in survival between the low and high expression groups of each subtype, and the *p*-value was calculated using the log-rank test.

[4]http://www.pathwaymapper.org/

[5]https://www.cbioportal.org/

[6]https://github.com/icbi-lab/Immunophenogram

RESULTS

Identification of Molecular Subtype-Specific RNAs in Colon Cancer

First, we predicted the molecular subtypes of 471 colon cancer samples using the CMSclassifier method. The number of samples of each colon cancer subtype was as follows: CMS1: 70; CMS2: 134; CMS3: 109; CMS4: 158 (**Figure 1A**). We separately counted the number of RNAs specifically expressed in different subtypes (**Figure 1B** and **Supplementary Table 1**). Next, we explored the changes in the expression of specific mRNAs among different subtypes by performing hierarchical clustering analysis. Specific mRNAs identified in each subtype were divided into clusters with different biological characteristics (**Figure 1C**). The right panel shows the average value of each cluster, showing a significant difference among the four subtypes. We also analyzed the expression levels of lncRNAs and miRNAs in different subtypes (**Supplementary Figures 1**, **2**). Based on these results, the expression level of RNAs specific to different subtypes was significantly different in a single subtype compared with other subtypes.

Gene Mutations Among the Colon Cancer Subtypes

To further explore the specific changes in different molecular subtypes in colon cancer, we evaluated the gene mutation status among the four subtypes of colon cancer, such as DNA mutations, CNVs, and SNPs. Missense mutations accounted for the highest proportion of all mutation types (**Figure 2A**), with the CMS2 subtype demonstrating the highest probability of missense mutations (85.95%) among the four subtypes. In the CMS1 subtype, the frameshift deletion mutation rate was the highest among the four types, at 13.98%, while the nonsense mutation rate was the lowest at 4.79%. **Figure 2B** shows the CNVs among the four different types. The CNV level of the four subtypes showed the lowest amplification in the CMS1 subtype and the highest deletions in the CMS4 subtype. Next, we compared the changes in SNPs among the different subtypes. The C > T transversion ratio among the four subtypes was the highest, exceeding 50%, reaching 57.63% in the CMS1 type (**Figure 2C**). The C > A and T > C conversion ratios were also higher in the four subtypes. Among them, the C > A conversion ratio was higher in the CMS3 type, and T > C was higher in CMS1. The CMS4 subtype showed more T > G transversions than the other subtypes. The proportion of CMS2 in the T > A and C > G subtypes was higher than that in the other subtypes. These results suggest specific differences in the gene mutation levels among the different subtypes. Next, we performed DNA mutation analysis on the subtype-specific genes obtained in Result 1. Nine genes were identified with a higher proportion of amplification and deletion—MYH7B, R3HDML, F7, SLED1, MSH4, IGF2, IGF2-AS, XKR4, and miR-6848 (**Figure 2D**). Further analysis of the correlation between the copy number variation of these genes and their RNA expression levels showed that most genes are related to their copy number variation. Among the four subtypes,

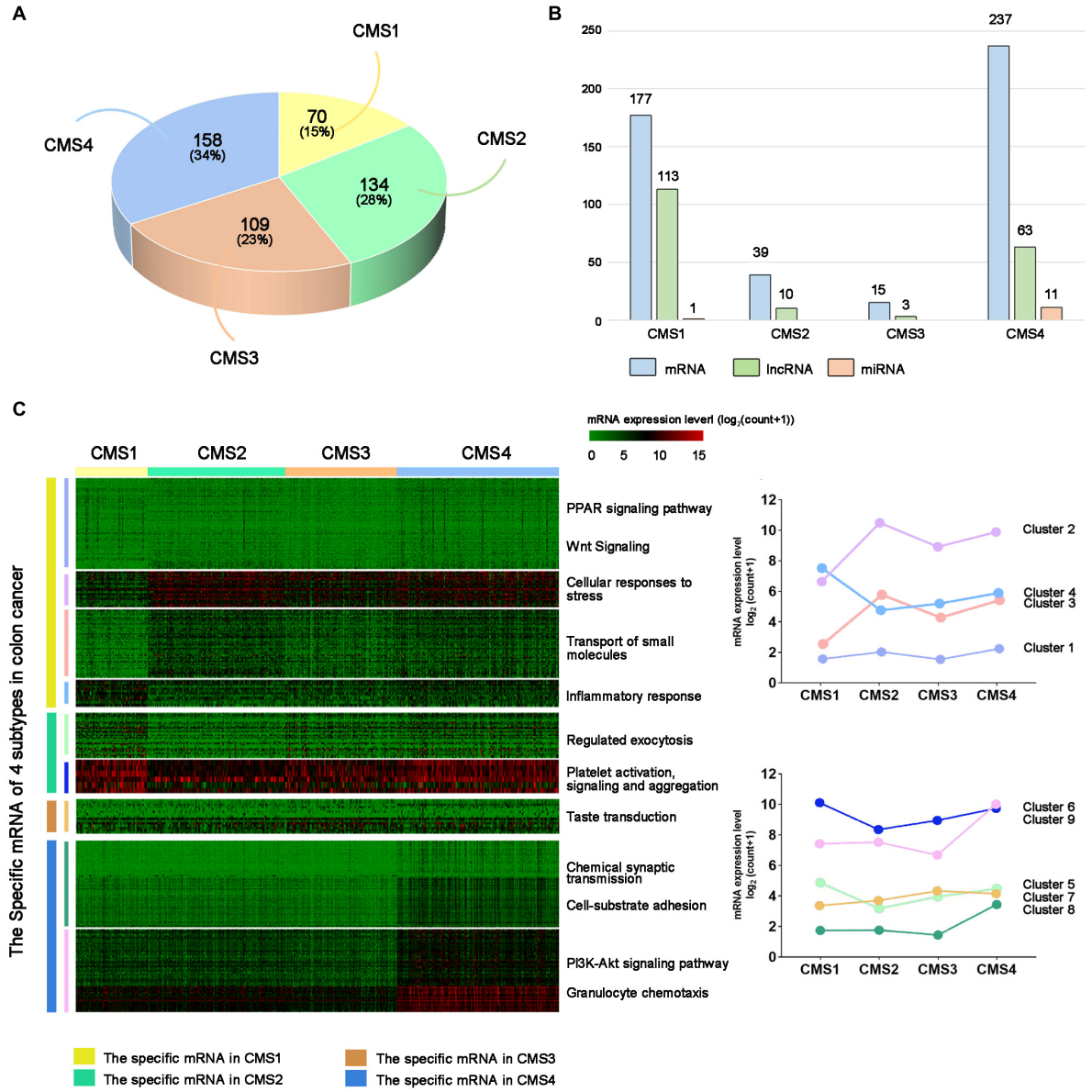


FIGURE 1 | Subtype-specific RNA analysis of colon cancer. **(A)** Statistical analysis of colon cancer samples from each subtype. **(B)** Number of subtype-specific RNAs identified in each subtype. **(C)** The left panel shows hierarchical clustering analysis and GO analysis of subtype-specific RNAs in each subtype. Red indicates a high expression level, and green indicates a low expression level. The right panel shows the average values of subtype-specific RNAs in different clusters of the four subtypes.

the expression level of R3HDML was significantly positively correlated with its DNA amplification (**Figures 2D,E**).

Subtype-Specific Genes Predicts the Prognosis of Patients With Colon Cancer

The above results suggest that the expression levels and mutations of specific RNAs in different subtypes show significant differences. Therefore, we speculate that changes in these genes may predict the prognosis of patients with different subtypes of colon cancer. Using the Kaplan–Meier method, we compared the overall survival of patients between the subtype-specific RNA high expression and low expression groups. The expression levels of many subtype-specific RNAs predicted the prognosis of patients with a single subtype. For example, in CMS1, the survival period of patients with high levels of CLDN8 was significantly shorter than that of patients with low levels of CLDN8. In CMS4, patients had lower levels of AP003548.1, and their survival time was longer (**Figure 3A** and **Supplementary Figure 3**; $p < 0.05$).

Additionally, we divided the samples of different types into a gene alteration group and a no-alteration group according

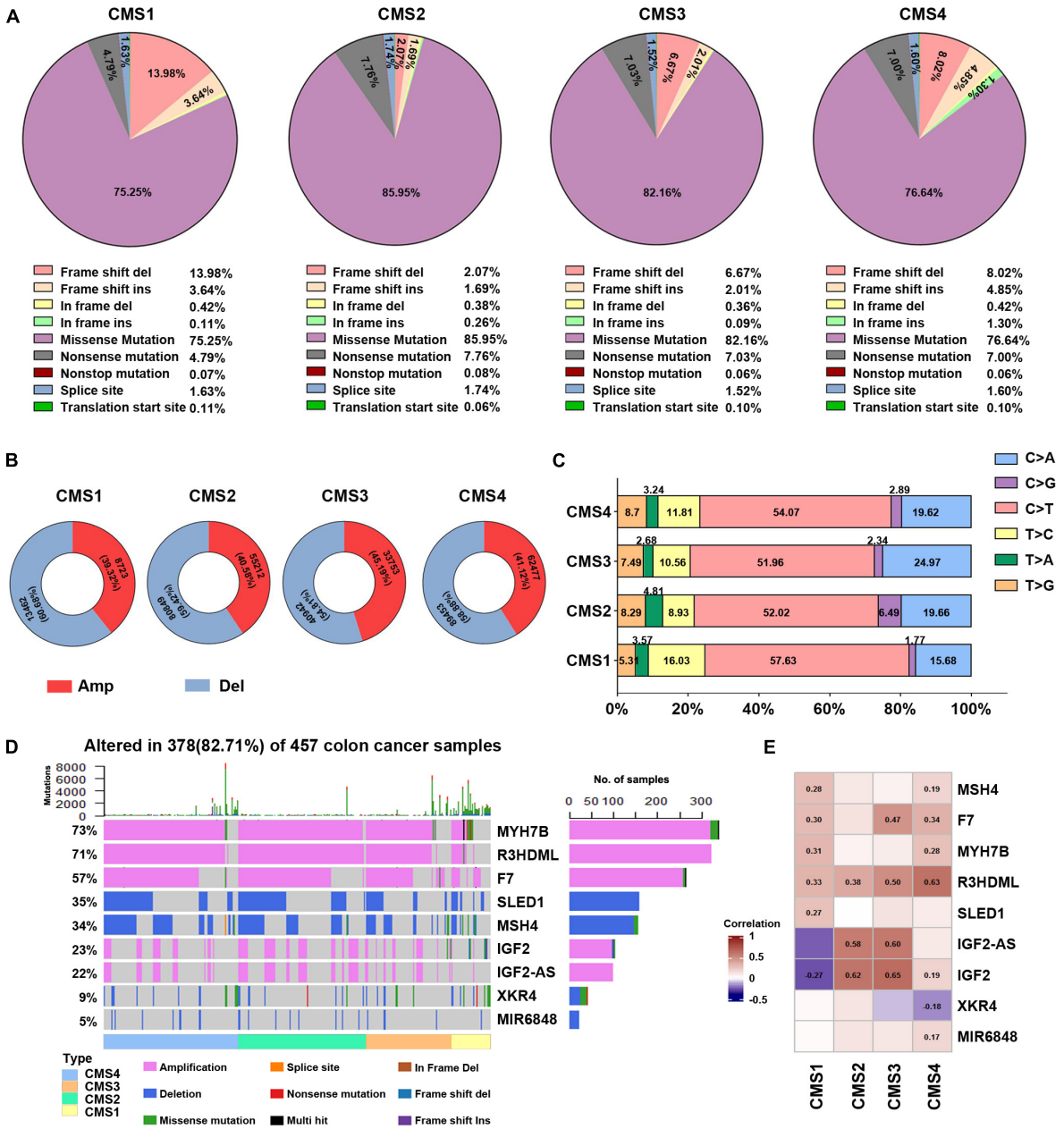


FIGURE 2 | Analysis of DNA alterations in the four colon cancer subtypes. **(A)** Different classified categories of DNA mutations in each colon cancer subtype. **(B)** Number of copy number variations (CNVs) in each subtype. Red represents amplification, and blue indicates deletion. **(C)** Change in the SNV transversion ratios in the different subtypes. **(D)** Waterfall chart of the mutations in different subtype-specific genes. **(E)** Correlation analysis of subtype-specific RNAs and their modifications in the different subtypes. The numbers in the figure are correlation coefficients, where a negative value represents a negative correlation and a positive value represents a positive correlation.

to whether a mutation was present in the subtype-specific gene. Furthermore, we analyzed whether subtype-specific genetic changes affect the prognosis of patients. In the CMS1 sample, the survival rate of the abnormally changed FSTL4 and LY6G6E groups was significantly shorter than that of the entire group. By contrast, the unusually altered FSTL4 group in the CMS3 sample had longer survival rates than the no-alteration group (**Figure 3B**). These results indicate that the expression levels and mutations of these subtype-specific genes can be used to predict the prognosis of patients with different types.

Identification of Driver Genes in the Colon Cancer Subtypes

Tumor driver genes play critical roles in the occurrence and development of tumors. To further investigate the molecular

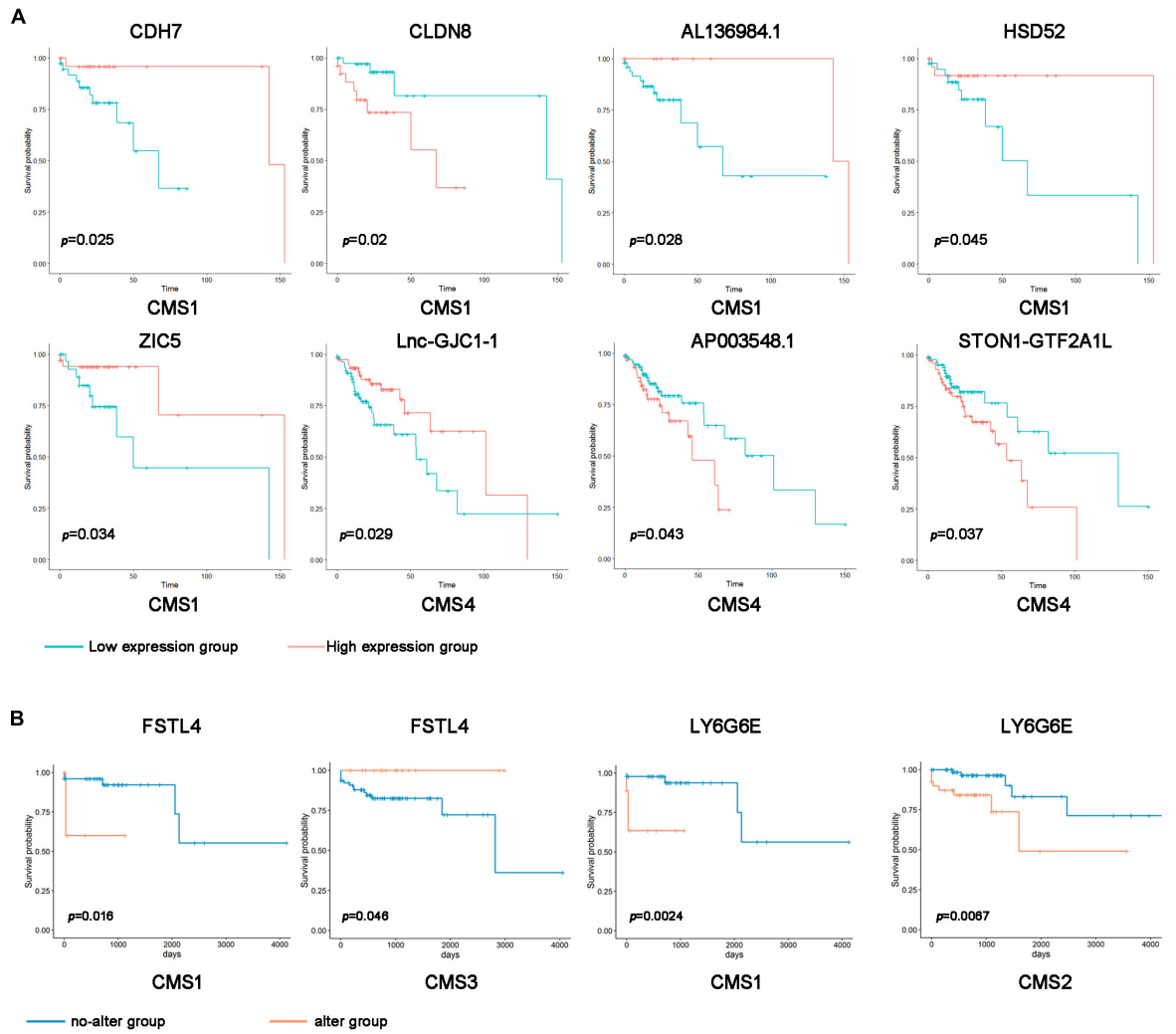


FIGURE 3 | Correlation of subtype-specific RNA expression with overall survival in colon cancer patients. **(A)** Kaplan–Meier survival curves were generated for subtype-specific RNAs by comparing groups with high (red line) and low (blue line) gene expression. $p < 0.05$ according to the log-rank test. **(B)** Kaplan–Meier survival curve of the subtype-specific genes with gene mutations. Orange indicates the gene alteration group, and blue indicates the no-alteration group.

features of different subtypes, we predicted 30 driver genes in colon cancer samples using the DriverDBv3 online database to explore the molecular characteristics of different subtypes (**Supplementary Table 2**). Differential analysis of the expression levels of driver genes in different subtypes revealed that the expression levels of almost all driver genes were significantly higher in CMS4 samples than in other subtype samples (**Figure 4A** and **Supplementary Figure 4A**). The expression of most driver genes was lower in CMS3. AXIN2, NEB, NRAS, RP1, RNF43, and other driver genes were significantly lower in CMS1 than in the other subtypes (**Figure 4A** and **Supplementary Figure 4A**). These results indicate that different driver genes may ultimately play different roles in the different subtypes. Regarding the different subtypes, driver genes may become decisive biomarkers and potential therapeutic targets.

Next, we further analyzed the mutations of driver genes in the different subtypes. Most colon cancer samples showed abnormal changes in the APC, SMAD4, ARID1A, TP53, FBXW7, KRAS, RNF43, NRAS, ATM, and PIK3CA genes (**Figure 4B**). The BRAF gene is mainly expressed as missense mutations in CMS1 subtypes, and its mutation rate is significantly higher than that of the other subtypes. Except for CMS1, the mutation frequency of APC in the three other subtypes was more than 50%, and the mutation frequency of APC was the highest in CMS2. Missense mutations of KRAS and PIK3CA are more widespread in the CMS3 subtype, and the TP53 mutation is the most common in the CMS4 subtype. CMS4 subtype patients also had a higher proportion of SMAD4, ARID1A, and NRAS deletions. Next, we analyzed the correlation between the expression level of a driver gene and its mutation. The expression levels of many driver genes were positively correlated with their gene deletions, such as TGIF1, SMAD4, and ARID1A (**Figure 4C** and **Supplementary Figures 4A,B**). Additionally, to further study the potential cancer drivers in colon cancer progression, we evaluated 10 classical

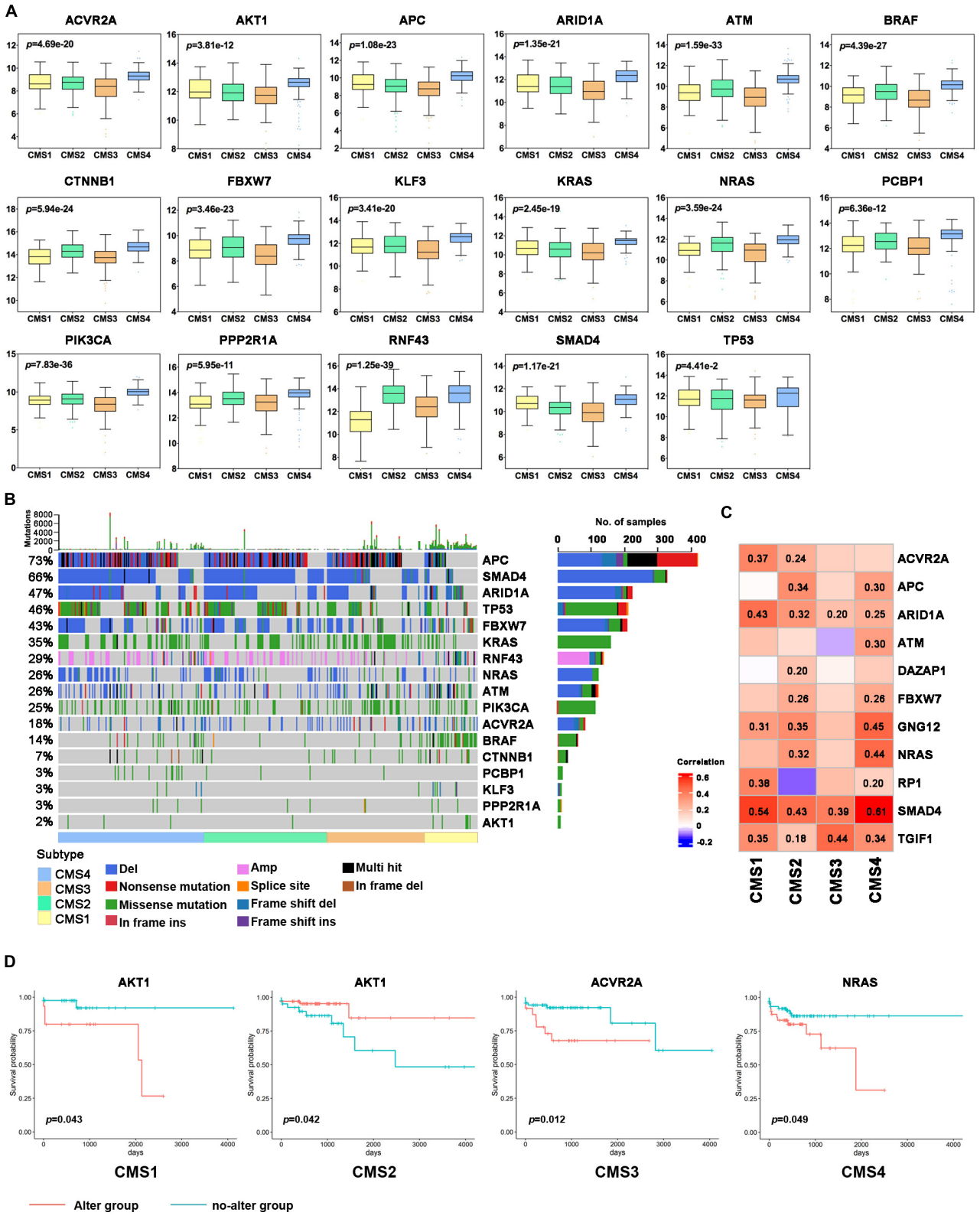


FIGURE 4 | Analysis of driver genes in the different colon cancer subtypes. **(A)** Expression levels of the driver genes in each subtype. **(B)** Waterfall chart of the frequency of mutations in the driver genes in each subtype. **(C)** Correlation analysis of subtype-specific RNAs and their modifications in the different subtypes. **(D)** Kaplan–Meier survival curve of the driver genes with gene mutations. Red indicates the gene alteration group, and blue indicates the no-alteration group.

signaling pathways with frequently altered genes and essential cancer relative genes explored from these pathways derived from TCGA data. It was showed that critical oncogenes and tumor suppressor genes had higher mutation rates, such as KRAS, BRAF, APC, AMER1, TCF7L2, PIK3CA, FBXW7, DCHS2, FAT1, TP53, SMAD4, and ATM (**Figure 5**).

To further understand the significance of driving gene changes on the prognosis of different subtypes, we evaluated the alter of driver genes on the prognosis of patients with colon cancer. It was shown that the patients with CMS1 presenting alter of AKT1 had a shorter OS, and AKT1 no-alter of CMS2 had a prolonged OS (**Figure 4D**; $p < 0.05$). The survival rate of the abnormally changed KLF3 group in the CMS1 subtype was significantly shorter than that of the no-alteration group. The survival rate of the abnormally altered PCBP1 group in the CMS3 subtype was markedly shorter than that of the no-alteration group (**Supplementary Figure 4C**; $p < 0.05$).

Differences in Immune Cell Infiltration Among the Different Subtypes

Immune cells are considered to be an important component of tumor microenvironment (TME), which plays an important role in tumor genesis, outcome, and treatment (especially immunotherapy) (Zhang et al., 2021). In addition to the analysis of subtype-specific genes alteration, an in-depth study of the clinical significance of immune cell infiltration levels in different subtypes is helpful to determine a clinical diagnosis and patient treatment. Therefore, we investigated the different invasion levels of 22 immune cell types in the 4 colon cancer subtypes. The infiltration level of these immune cell types was significantly different in the four subtypes of colon cancer (**Figure 6A**). CMS1 is an MSI immune subtype with the characteristics of immune infiltration and activation. The infiltration levels of CD8 T cells, T follicular helper cells, M1 macrophages, neutrophils, and NK activated cells in the CMS1 subtype were significantly higher than those in the other three subtypes. In the CMS2 subtype, the levels of infiltrating CD4 memory-activated cells and NK resting cells were higher than those in the other three subtypes. The CD4 memory resting cells were the highest in the CMS3 subtypes. The levels of infiltrating M0 macrophages and M2 macrophages were more elevated in the CMS4 subtype than in the other subtypes. Next, we examined the difference in the levels of infiltrating fibroblasts in the four subtypes, among which the levels of infiltrating fibroblasts in CMS4 were higher than those in the other subtypes. Thus, we further assessed the correlation between 22 immune cell types and fibroblasts in the different subtypes. Many T cells were negatively correlated with fibroblasts in the different subtypes, such as T follicular helper cells, which were negatively correlated with fibroblasts in CMS1 (**Figure 6B**). CD4 activated and resting T cells showed a negative correlation with fibroblasts in CMS4. Naïve CD4 cells were also negatively correlated with fibroblasts in CMS2 and CMS3.

Furthermore, we compared the TMB levels among the four subtypes and analyzed the correlation between the TMB and immune cell infiltration level. The level of TMB in the CMS1 subtype was significantly higher than that in the other subtypes (**Figure 6A**). Correlation analysis showed a significant correlation between the TMB and infiltration levels of specific immune cell types in the different subtypes. For example, the infiltration level of B memory cells was significantly positively correlated with the TMB in CMS1; T follicular helper cells and mast activated cells showed a positive correlation with the TMB in CM2. In CMS3, M1, and M2 macrophages showed a positive correlation with the TMB (**Supplementary Figure 5**).

Because the level of immune cell infiltration in different subtypes is significantly different, we speculate that its infiltration may be related to the prognosis of patients with different subtypes. Therefore, according to the level of immune cell infiltration, samples of various types were divided into high infiltration level and low infiltration level groups. Kaplan–Meier analysis was used to compare the prognostic differences between the two groups (**Figure 7A**). A low degree of resting NK cell and Treg cell infiltration indicated a better prognosis for patients with the CMS4 subtype. The group with high M0 macrophage infiltration and the CMS1 subtype showed a more extended survival period; high M1 macrophage infiltration in the CMS3 subtype group correlated with a better prognosis. We also analyzed whether the TMB level in the different subtypes affects the prognosis of patients. Only patients with CMS4 significantly differed in OS between the high TMB and low TMB groups (**Figure 7B**.).

Immunophenogram Analysis Predicts the Response of the Different Molecular Subtypes to Immunotherapy

Immune checkpoint genes play crucial roles in evading self-reactivity and represent a new target to develop cancer treatments (Schoenfeld and Hellmann, 2020; Mauri et al., 2021). First, based on the published literature, we screened 28 ICGs and evaluated the differences in their expression levels among the different subtypes (**Figure 8A**). We performed hierarchical cluster analysis on the expression levels of these ICGs and evaluated their overall expression levels in the 4 subtypes based on the average expression of 28 genes. The expression levels of most genes in CMS1 and CMS4 were higher. The first cluster of genes showed a higher expression level in CMS4, and the fourth cluster of genes showed a higher expression level in CMS1 (**Figure 8A**). PDL1, LAG3, KLRC1, CD94, CD244, and IDO1 expression was higher in the CMS1 subtypes (**Supplementary Figure 6**; $p < 0.0001$). PD1, PDL2, CD2, CD40, CD47, CD80, CD86, CD96, and CD200R1 were higher in the CMS4 subtype than in the other subtypes. The increased expression of ICGs in CMS1 and CMS4 suggests that these two types may respond better to immunotherapy. To further verify this hypothesis, we divided different types of samples into high-expression and low-expression groups according to the expression levels of ICGs and compared the OS between the two groups of samples. The patients with higher expression levels of immune checkpoints had a longer OS. For example, patients with higher levels of VTCN1 expression had a better OS rate in CMS1 (**Figure 8B**).

Next, we used the immune score to predict the response of the different subtype samples to PD1/PDL1 treatment. In the

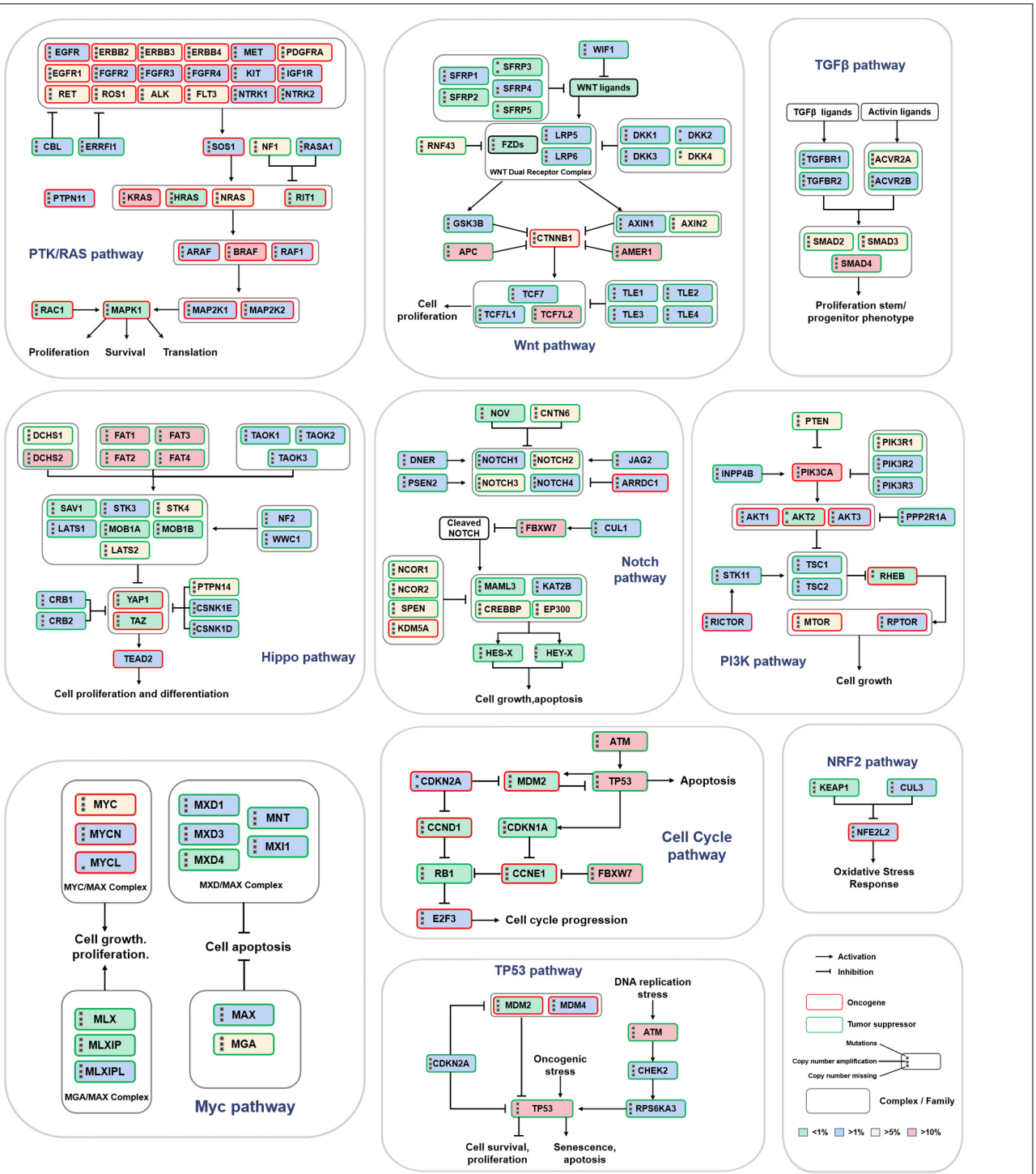


FIGURE 5 | Mutation frequency of key genes in 10 classic signaling pathways. The three black dots on the left side of the gene name represent gene mutations, copy number amplifications and deletions. The red border represents oncogenes, and the green border indicates tumor suppressor genes. The intensity of the color represents the mutation frequency.

PD1-positive group, the IPS of CMS1 was higher than that of the other subtypes (**Figure 8C**). However, the IPS of CMS4 was lower in these groups. The above results indicate that CMS1 patients may have a higher positive response to immunotherapy, such as anti-PD-1/PD-L1 therapy or anti-PD-1/PD-L1 combined with anti-CTLA4 treatment.

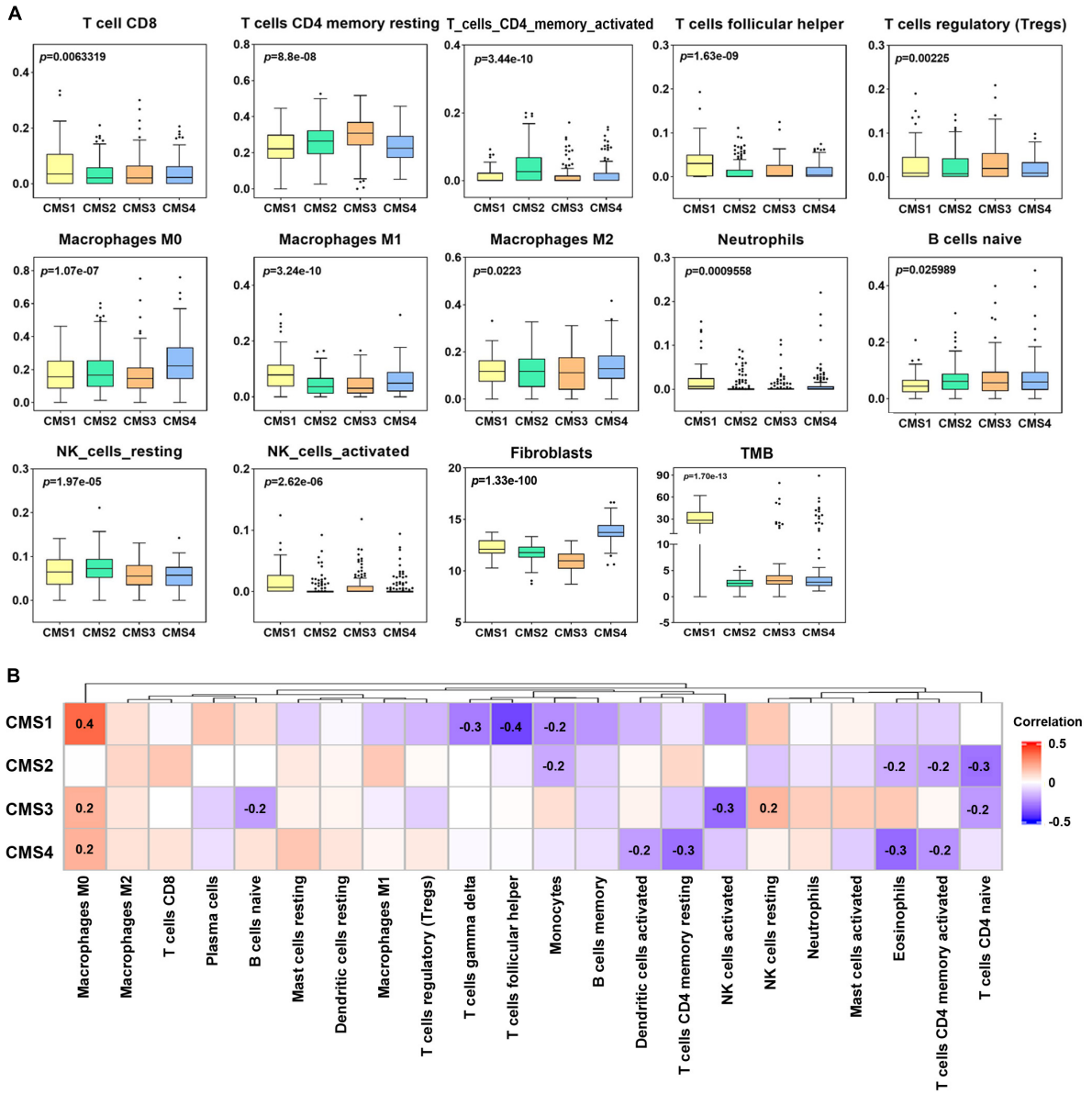


FIGURE 6 | Levels of immune cell infiltration in different colon cancer subtypes. **(A)** Differential analysis of the immune cell infiltration level, fibroblasts and tumor mutation burden (TMB) in different colon cancer subtypes ($p < 0.05$). **(B)** Correlation analysis between fibroblasts and the levels of twenty-two infiltrating immune cell types in the different colon cancer subtypes. Red indicates a positive correlation, and blue indicates a negative correlation. The numerical value represents the degree of correlation ($p < 0.05$).

DISCUSSION

The high heterogeneity of colon cancer remains a challenge to treat patients with different types of colon cancer. This study further interpreted the specific molecular characteristics of different molecular subtypes by analyzing the RNA levels, DNA alterations, and immune cell infiltration levels in the four molecular subtypes of colon cancer. First, by examining the RNA levels in the different molecular types, RNAs with specific expression levels in the different subtypes were screened, including mRNAs, lncRNAs, and miRNAs. These subtype-specific RNAs may be biomarkers for patients with different colon cancer subtypes' clinical diagnosis, treatment, and prognosis. For example, compared with the other subtypes, TDGF1 expression was lower in CMS1. Low TDGF1 expression results in a worse prognosis of patients with the CMS1 subtype (**Supplementary Figure 3**). Similarly, Sun et al. (2020) reported that TDGF1, an immune-related gene (IRG), is a low-risk prognosis-related gene in colon cancer. Besides, we found that patients with high expression of CLDN8 have a worse prognosis in the CMS1 subtype. Claudin 8 (CLDN8) is a complete membrane protein that constitutes the tight junction

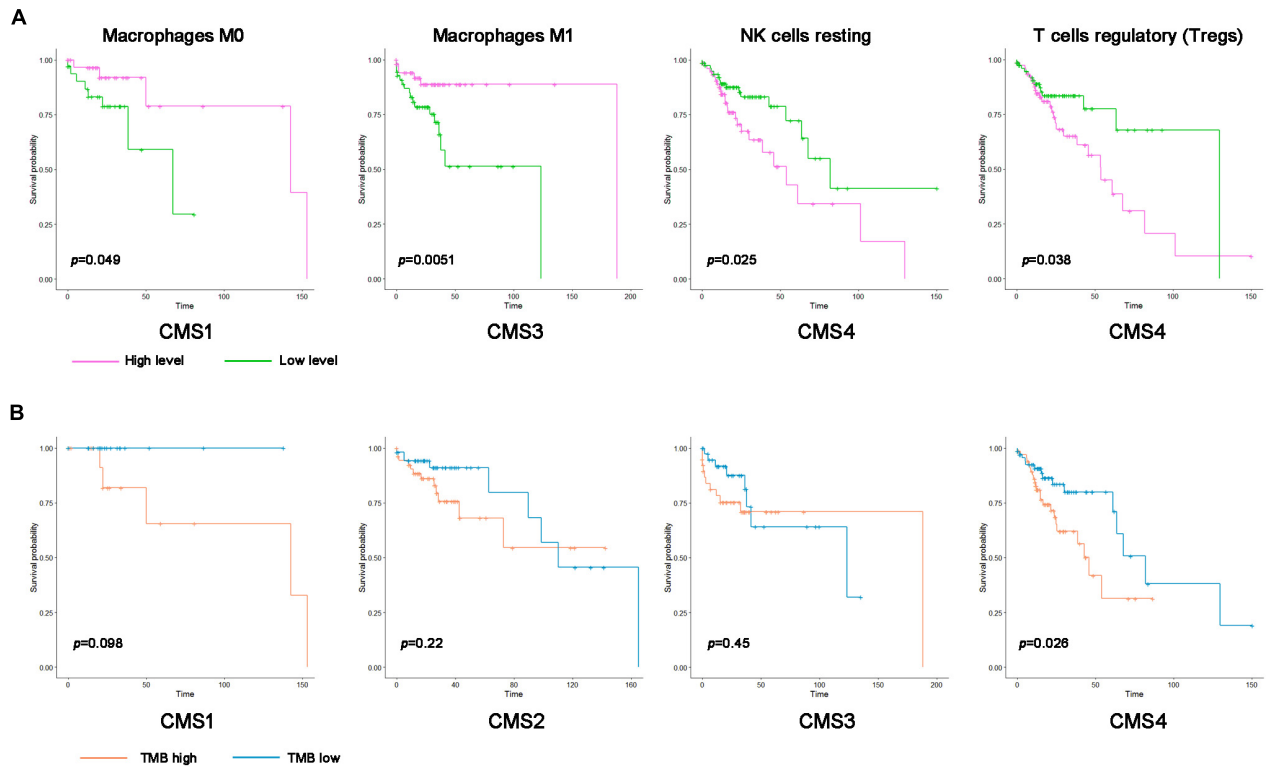


FIGURE 7 | Kaplan–Meier survival curve of immune cells and the tumor mutation burden (TMB) in the different colon cancer subtypes. **(A)** Kaplan–Meier survival curve of the immune cell infiltration level in the different colon cancer subtypes ($p < 0.05$). Pink represents a high infiltration level, and green represents a low infiltration level. **(B)** Kaplan–Meier survival curve of the TMB in the different colon cancer subtypes. Blue represents the high TMB group, and orange represents the low TMB group.

of the cell membrane. Cheng et al. (2019) have shown that CLDN8 promotes the proliferation, migration, and invasion of CRC cancer cells by activating MAPK/ERK signaling. Moreover, Gröschl et al. (2013) found that DUSP4 expression was lower in MSS colon cancer samples than MSI-H colon cancer samples by microarray analysis. This result is consistent with our research, which showed low expression of DUSP4 in CMS2 (**Supplementary Figure 8**). Intelectin-1 (ITLN1) is a novel adipokine that is used as a marker of metabolic disorders. Many studies reported that it might be an effect as an inhibitor in multiple cancers, including gastric cancer, ovarian cancer, and colon cancer. Katsuya reported that colon cancer patients with low ITLN1 expression had a higher grade of metastasis (M grade) than colon cancer patients with retained ITLN1 expression. And patients with retained ITLN1 expression had a better prognosis (Katsuya et al., 2020). This study showed that ITLN1 was significantly higher in CMS3 (a metabolic subtype) than other subtypes (**Supplementary Figure 8**). Furthermore, an increased expression of ITLN1 showed a better prognosis in colon cancer patients (**Supplementary Figure 8**). Next, we further analyzed the mutations of subtype-specific RNA, and the expression level of many subtype-specific RNAs was significantly correlated with DNA mutations. For example, the expression level of the specific gene MSH4 in CMS1 is positively correlated with DNA amplification. The expression levels of the IGF2 and IGF2-AS genes in the CMS2 subtype are significantly positively correlated with DNA amplification. These results also indicate that changes in specific genes in different subtypes may help explain the molecular characteristics of different subtypes.

In addition to the impact of subtype-specific genes on the different subtypes, driver gene changes also play essential roles. Many studies have shown that genetic alterations in driver genes contribute to colon cancer development and malignant progression (Nakayama and Oshima, 2019). In 2012, a comprehensive analysis of colon cancer using TCGA found that 16% of colon cancer cases contained hypermutations in microsatellite regions and were classified as having an MSI phenotype (Cancer Genome Atlas Network, 2012a). The other 84% of colon cancer samples were identified as microsatellite stable (MSS) and showed more mutations (Cancer Genome Atlas Network, 2012a). The most frequently mutated genes in colon cancer are APC, KRAS, SMAD4, and TP53 (Huang et al., 2018; Yaeger et al., 2018). Mutations in these genes occur in approximately 10–80% of colon cancers and play crucial roles in CRC metastasis (Cancer Genome Atlas Network, 2012a; Giannakis et al., 2016). Additionally, many other driver genes are frequently mutated in colon cancer, such as ARID1A, SOX9, FAM123B, BCL9L, RBM10, CTCF, and KLF5. In the present study, APC (73%), SMAD4 (66%), ARID1A (47%), and TP53 (46%) were mutated in many colon cancer samples. CMS2, as

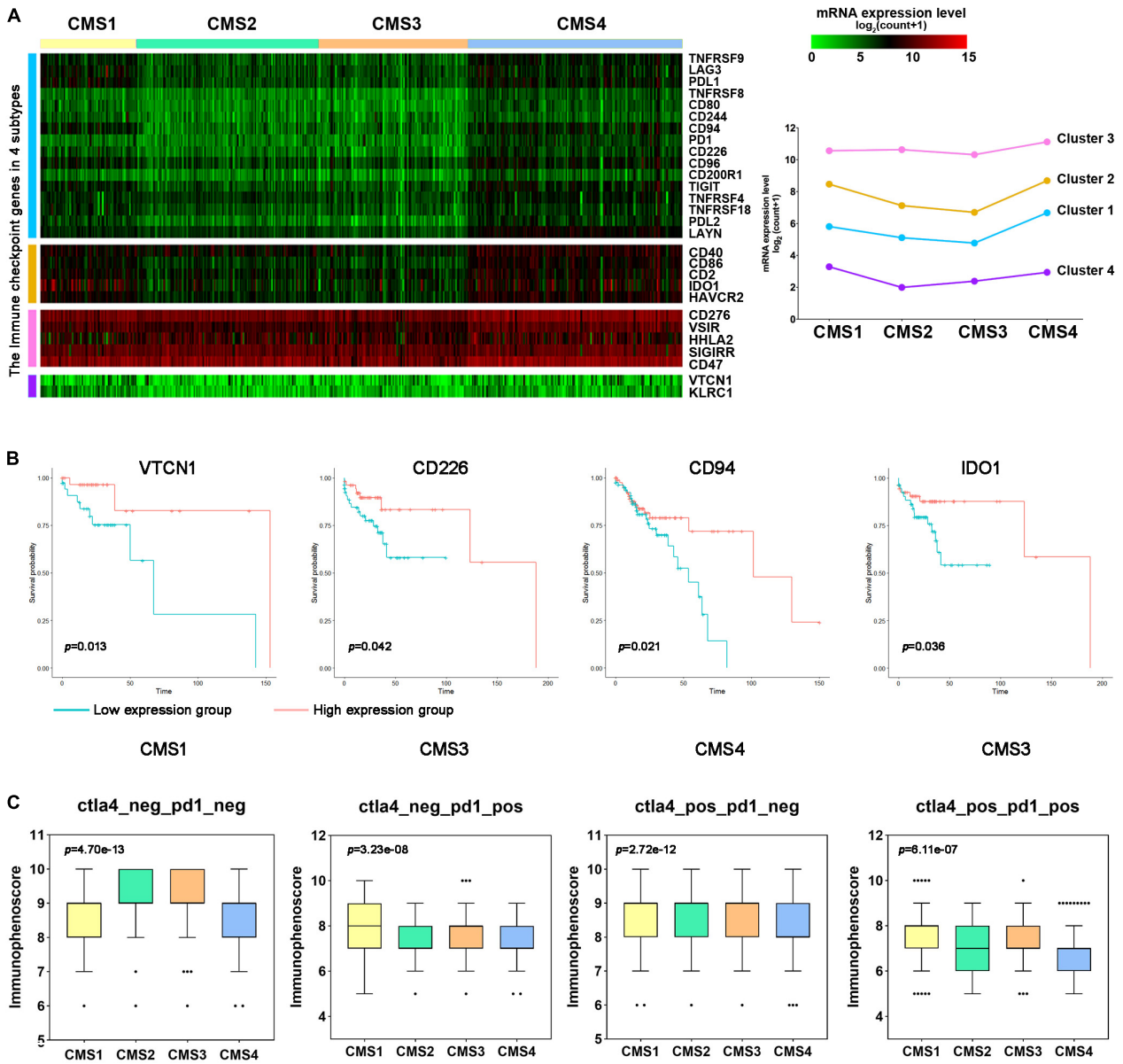


FIGURE 8 | Immune checkpoint gene expression and immune response prediction in each subtype of colon cancer. **(A)** Hierarchical clustering analysis of the mRNA expression levels for immune inhibitory checkpoint-related genes. **(B)** Kaplan–Meier survival curves of immune checkpoint genes (ICGs) in patients with different colon cancer subtypes. Red represents the high expression group, and blue represents the low expression group. **(C)** Comparison of the IPS between patients with each subtype of colon cancer in the CTLA4-negative/positive or PD-1-negative/positive groups. CTLA4 positivity or PD1 positivity represents anti-CTLA4 or anti-PD-1/PD-L1 therapy, respectively.

the most common chromosomal unstable subtype, has significant somatic CNVs. In addition to the higher mutation levels of APC and TP53, our analysis showed that the copy number deletion of TGIF1 in CMS2 was significantly higher than that in the other subtypes. Abnormally high expression of TGIF1 is found in colorectal cancer tissues and promotes the progression of CRC (Shah et al., 2019; Li et al., 2020). And our research showed that ARID1A deletion is more common in CMS4 than in other subtypes. ARID1A is essential for maintaining intestinal stem cells and intestinal homeostasis in mice, and its deletion can induce the occurrence of CRC in mice (Mathur et al., 2017; Hiramatsu et al., 2019). Erfani also proved that ARID1A downregulation might promote CRC metastasis and epithelial cell movement by decreasing EMT-related protein E-cadherin (Erfani et al., 2021). In addition to influencing colon cancer progression, abnormally altered driver genes may help explain the differences among the four subtypes in response to clinical treatment. The mutation frequencies of TP53 were 56, 32, and 59% in the CMS2, CMS3, and CMS4 subtypes, respectively. Based on this result, TP53 mutation indicates

a robust carcinogenic effect on most colon cancers (Cancer Genome Atlas Network, 2012b). Therefore, further research into driver gene mutations in colon cancer may have good therapeutic value for targeted therapy.

The proportion of MSI in the CMS1 subtype is significantly higher than that of other subtypes. 15–20% of CMS1 tumors represent MSI characteristics. MSI tumor's core area and surrounding area are rich in tumor-infiltrating cytotoxic T lymphocytes (CTL) compared with MSS tumor (De Smedt et al., 2015). In this study, numerous immune cells in CMS1, including CD8 + T cells, were significantly higher than other subtypes, consistent with previous research. Several studies have confirmed that immune cell infiltration is especially related to the better prognosis of MSI tumors (Deschoolmeester et al., 2011). Furthermore, the study indicated the presence of CD8 + TILs on the periphery of the tumor was significantly associated with a better prognosis (Park et al., 2016). Therefore, TIL is a critical prognostic component of CMS1. Besides, A high level of M1 macrophage infiltration in patients with the CMS3 subtype indicated a longer survival time. High Treg and NK cell infiltration in the CMS4 subtype indicated a worse prognosis for patients with colon cancer. CAFs were significantly different among the four subtypes of colon cancer and correlated with immune cell infiltration. In the CMS1 immune subtype, CAFs were negatively correlated with T cell infiltration. We assessed the correlation between immune cells and the TMB. M1 and M2 macrophages showed a positive correlation with the TMB in the CMS3 subtype (**Supplementary Figure 5**). These results indicated that immune cells were correlated with the survival of patients with different subtypes of colon cancer.

Since ctLA-4 inhibitors were approved for melanoma in 2011, PD1/PD-L1 inhibitors have now been used to treat more than 20 different cancers, including colon cancer. In the present study, we analyzed 29 ICGs and found that most were expressed at higher levels in CMS4 than in the other subtypes, such as PD1, PD-L1, PD-L2, CD2, and CD40. Most ICGs showed low mutation levels in colon cancer cases, except CD2, VTCN1, and LAYN (**Supplementary Figure 7**). This result suggests the stability of immune therapy targeting ICGs. The CMS1 subtype responds well to anti-PD-1/anti-CTLA4 combination treatment but not to anti-CTLA4 monotherapy. However, the other subtypes did not respond to either anti-CTLA4 or anti-PD-1 therapy. These features suggest that the CMS1 subtype is susceptible to immunotherapy. Significant immune responses were observed in patients with different colon cancer subtypes, a finding that may be related to the complex cellular components in tumor tissues. In the future, the effect of tumor immunotherapy may be improved by changing the ratio of immune cells. The interplay mechanisms of tumor cells and immune cells in different subtypes are essential to predict the therapeutic response and prognosis. They may also contribute to developing an individualized treatment plan for patients with different subtypes of colon cancer.

In summary, we investigated the specific molecular characteristics of the four subtypes of colon cancer from multiple perspectives. The findings may provide a theoretical basis to identify patients more likely to benefit from immunotherapy and offer potential biomarkers for future immunotherapy.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

JYW and FH designed and supervised this study. FH drafted the manuscript. JYW and SW illustrated the figures for the manuscript and analyzed the data. LZ and MZ searched the database. BC, HY, and JRW reviewed and edited the manuscript. All authors approved the final manuscript.

FUNDING

This study was supported by the Beijing Medical Award Foundation (YXJL-2019-0072-0023 to BC), Nn10 Program of Harbin Medical University Cancer Hospital (Nn102017-02 to BC), the Natural Science Foundation of Inner Mongolia (No. 2020MS08084 to MZ), and the Inner Mongolia Science and Technology Plan Project (Nos. 201802133 and 2020GG0297 to MZ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.758776/full#supplementary-material

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Bahceci, I., Dogrusoz, U., La, K. C., Babur, O., Gao, J., and Schultz, N. (2017). PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data. *Bioinformatics* 33, 2238–2240. doi: 10.1093/bioinformatics/btx149

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* 17:218. doi: 10.1186/s13059-016-1070-5

Benson, A. B., Venook, A. P., Al-Hawary, M. M., Arain, M. A., Chen, Y. J., Ciombor, K. K., et al. (2020). NCCN guidelines insights: rectal cancer, Version 6.2020. *J. Natl. Compr. Cancer Netw.* 18, 806–815. doi: 10.6004/jnccn.2020.0032

Benson, A. B., Venook, A. P., Al-Hawary, M. M., Cederquist, L., Chen, Y. J., Ciombor, K. K., et al. (2018). NCCN guidelines insights: colon cancer, version 2.2018. *J. Natl. Compr. Cancer Netw.* 16, 359–369. doi: 10.6004/jnccn.2018.0021

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K. O., et al. (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* 231, 63–76. doi: 10.1002/path.4212

Cancer Genome Atlas Network (2012a). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337. doi: 10.1038/nature11252

Cancer Genome Atlas Network (2012b). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* 18, 248–262. doi: 10.1016/j.celrep.2016.12.019

Cheng, B., Rong, A., Zhou, Q., and Li, W. (2019). CLDN8 promotes colorectal cancer cell proliferation, migration, and invasion by activating MAPK/ERK signaling. *Cancer Manag. Res.* 11, 3741–3751. doi: 10.2147/CMAR.S189558

De Smedt, L., Lemahieu, J., Palmans, S., Govaere, O., Tousseyn, T., Van Cutsem, E., et al. (2015). Microsatellite instable vs stable colon carcinomas: analysis of tumour heterogeneity, inflammation and angiogenesis. *Br. J. Cancer* 113, 500–509. doi: 10.1038/bjc.2015.213

De Sousa, E. M. F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L. P., et al. (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.* 19, 614–618. doi: 10.1038/nm.3174

Deschoolmeester, V., Baay, M., Lardon, F., Pauwels, P., and Peeters, M. (2011). Immune cells in colorectal cancer: prognostic relevance and role of MSI. *Cancer Microenviron.* 4, 377–392. doi: 10.1007/s12307-011-0068-5

Erfani, M., Zamani, M., Hosseini, S. Y., Mostafavi-Pour, Z., Shafiee, S. M., Saeidnia, M., et al. (2021). ARID1A regulates E-cadherin expression in colorectal cancer cells: a promising candidate therapeutic target. *Mol. Biol. Rep.* 48, 6749–6756. doi: 10.1007/s11033-021-06671-9

Giannakis, M., Mu, X. J., Shukla, S. A., Qian, Z. R., Cohen, O., Nishihara, R., et al. (2016). Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep.* 15, 857–865. doi: 10.1016/j.celrep.2016.03.075

Gröschl, B., Bettstetter, M., Giedl, C., Woenckhaus, M., Edmonston, T., Hofstadter, F., et al. (2013). Expression of the MAP kinase phosphatase DUSP4 is associated with microsatellite instability in colorectal cancer (CRC) and causes increased cell proliferation. *Int. J. Cancer* 132, 1537–1546. doi: 10.1002/ijc.27834

Guinney, J., Dienstmann, R., Wang, X., de Reynies, A., Schlicker, A., Soneson, C., et al. (2015). The consensus molecular subtypes of colorectal cancer. *Nat. Med.* 21, 1350–1356. doi: 10.1038/nm.3967

Hiramatsu, Y., Fukuda, A., Ogawa, S., Goto, N., Ikuta, K., Tsuda, M., et al. (2019). Arid1a is essential for intestinal stem cells through Sox9 regulation. *Proc. Natl. Acad. Sci. U.S.A.* 116, 1704–1713. doi: 10.1073/pnas.1804858116

Huang, D., Sun, W., Zhou, Y., Li, P., Chen, F., Chen, H., et al. (2018). Mutations of key driver genes in colorectal cancer progression and metastasis. *Cancer Metastasis Rev.* 37, 173–187. doi: 10.1007/s10555-017-9726-5

Katsuya, N., Sentani, K., Sekino, Y., Yamamoto, Y., Kobayashi, G., Babasaki, T., et al. (2020). Clinicopathological significance of intelectin-1 in colorectal cancer: intelectin-1 participates in tumor suppression and favorable progress. *Pathol. Int.* 70, 943–952. doi: 10.1111/pin.13027

Lawlor, R. T., Mattiolo, P., Mafficini, A., Hong, S. M., Piredda, M. L., Taormina, S. V., et al. (2021). Tumor mutational burden as a potential biomarker for immunotherapy in pancreatic cancer: systematic review and still-open questions. *Cancers* 13:3119. doi: 10.3390/cancers13133119

Li, J., Hu, M., Liu, N., Li, H., Yu, Z., Yan, Q., et al. (2020). HDAC3 deteriorates colorectal cancer progression via microRNA-296-3p/TGIF1/TGFbeta axis. *J. Exp. Clin. Cancer Res.* 39:248. doi: 10.1186/s13046-020-01720-w

Liu, S. H., Shen, P. C., Chen, C. Y., Hsu, A. N., Cho, Y. C., Lai, Y. L., et al. (2020). DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res.* 48, D863–D870. doi: 10.1093/nar/gkz964

Marisa, L., de Reynies, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., et al. (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.* 10:e1001453. doi: 10.1371/journal.pmed.1001453

Mathur, R., Alver, B. H., San Roman, A. K., Wilson, B. G., Wang, X., Agoston, A. T., et al. (2017). ARID1A loss impairs enhancer-mediated gene regulation and drives colon cancer in mice. *Nat. Genet.* 49, 296–302. doi: 10.1038/ng.3744

Mauri, G., Bonazzina, E., Amatu, A., Tosi, F., Bencardino, K., Gori, V., et al. (2021). The evolutionary landscape of treatment for BRAF(V600E) mutant metastatic colorectal cancer. *Cancers* 13:137. doi: 10.3390/cancers13010137

Nakayama, M., and Oshima, M. (2019). Mutant p53 in colon cancer. *J. Mol. Cell Biol.* 11, 267–276. doi: 10.1093/jmcb/mjy075

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Park, J. H., Powell, A. G., Roxburgh, C. S., Horgan, P. G., McMillan, D. C., and Edwards, J. (2016). Mismatch repair status in patients with primary operable colorectal cancer: associations with the local and systemic tumour environment. *Br. J. Cancer* 114, 562–570. doi: 10.1038/bjc.2016.17

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* 19, 619–625. doi: 10.1038/nm.3175

Schlicker, A., Beran, G., Chresta, C. M., McWalter, G., Pritchard, A., Weston, S., et al. (2012). Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med. Genomics* 5:66. doi: 10.1186/1755-8794-5-66

Schoenfeld, A. J., and Hellmann, M. D. (2020). Acquired resistance to immune checkpoint inhibitors. *Cancer Cell* 37, 443–455. doi: 10.1016/j.ccell.2020.03.017

Shah, A., Melhuish, T. A., Fox, T. E., Frierson, H. F. Jr., and Wotton, D. (2019). TGIF transcription factors repress acetyl CoA metabolic gene expression and promote intestinal tumor growth. *Genes Dev.* 33, 388–402. doi: 10.1101/gad.320127.118

Sun, Y. L., Zhang, Y., Guo, Y. C., Yang, Z. H., and Xu, Y. C. (2020). A prognostic model based on the immune-related genes in colon adenocarcinoma. *Int. J. Med. Sci.* 17, 1879–1896. doi: 10.7150/ijms.45813

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 Countries. *CA Cancer J. Clin.* 71, 209–249. doi: 10.3322/caac.21660

Van Cutsem, E., Cervantes, A., Adam, R., Sobrero, A., Van Krieken, J. H., Aderka, D., et al. (2016). ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. *Ann. Oncol.* 27, 1386–1422. doi: 10.1093/annonc/mdw235

Yaeger, R., Chatila, W. K., Lipsyc, M. D., Hechtman, J. F., Cercek, A., Sanchez-Vega, F., et al. (2018). Clinical sequencing defines the genomic landscape of metastatic colorectal cancer. *Cancer Cell* 33, 125.e3–136.e3. doi: 10.1016/j.ccell.2017.12.004

Zhang, Z., Bao, S., Yan, C., Hou, P., Zhou, M., and Sun, J. (2021). Computational principles and practice for decoding immune contexture in the tumor microenvironment. *Brief. Bioinform.* 22:bbaa075. doi: 10.1093/bib/bbaa075

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Hu, Wang, Zhang, Wang, Zhao, Yang, Wu and Cui. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

REVIEW
published: 06 December 2021
doi: 10.3389/fcell.2021.803608

Research on the Computational Prediction of Essential Genes

Yuxin Guo[1,2,3,4], *Ying Ju*[5], *Dong Chen*[6]* *and Lihong Wang*[7]*

[1]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [2]Key Laboratory of Computational Science and Application of Hainan Province, Haikou, China, [3]Key Laboratory of Data Science and Intelligence Education, Hainan Normal University, Ministry of Education, Haikou, China, [4]School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [5]School of Informatics, Xiamen University, Xiamen, China, [6]College of Electrical and Information Engineering, Quzhou University, Quzhou, China, [7]Beidahuang Industry Group General Hospital, Harbin, China

Genes, the nucleotide sequences that encode a polypeptide chain or functional RNA, are the basic genetic unit controlling biological traits. They are the guarantee of the basic structures and functions in organisms, and they store information related to biological factors and processes such as blood type, gestation, growth, and apoptosis. The environment and genetics jointly affect important physiological processes such as reproduction, cell division, and protein synthesis. Genes are related to a wide range of phenomena including growth, decline, illness, aging, and death. During the evolution of organisms, there is a class of genes that exist in a conserved form in multiple species. These genes are often located on the dominant strand of DNA and tend to have higher expression levels. The protein encoded by it usually either performs very important functions or is responsible for maintaining and repairing these essential functions. Such genes are called persistent genes. Among them, the irreplaceable part of the body's life activities is the essential gene. For example, when starch is the only source of energy, the genes related to starch digestion are essential genes. Without them, the organism will die because it cannot obtain enough energy to maintain basic functions. The function of the proteins encoded by these genes is thought to be fundamental to life. Nowadays, DNA can be extracted from blood, saliva, or tissue cells for genetic testing, and detailed genetic information can be obtained using the most advanced scientific instruments and technologies. The information gained from genetic testing is useful to assess the potential risks of disease, and to help determine the prognosis and development of diseases. Such information is also useful for developing personalized medication and providing targeted health guidance to improve the quality of life. Therefore, it is of great theoretical and practical significance to identify important and essential genes. In this paper, the research status of essential genes and the essential genome database of bacteria are reviewed, the computational prediction method of essential genes based on communication coding theory is expounded, and the significance and practical application value of essential genes are discussed.

Keywords: communication coding, database, essential genes, machine learning, minimum feature group

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Leyi Wei,
Shandong University, China
Wen Zhang,
Huazhong Agricultural University, China

***Correspondence:**
Dong Chen
peakgrin@outlook.com
Lihong Wang
2975241625@qq.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *28 October 2021*
Accepted: *22 November 2021*
Published: *06 December 2021*

Citation:
Guo Y, Ju Y, Chen D and Wang L (2021) Research on the Computational Prediction of Essential Genes. Front. Cell Dev. Biol. 9:803608. doi: 10.3389/fcell.2021.803608

INTRODUCTION

With the smooth progress of the Human Genome Project, more and more new genes have been discovered, and the study of gene function has become a major issue in the field of life sciences. Genes support the basic structure and performance of life. It stores all the information about the race, blood type, gestation, growth, apoptosis and other processes of life. The environment and genetics jointly determine important physiological processes such as reproduction, cell division, and protein synthesis. Genes are involved in diverse phenomena such as birth, growth, decline, illness, aging, and death (Yang et al., 2015; Hu et al., 2020; Cheng et al., 2021a). Except for some viruses whose genes are composed of ribonucleic acid (RNA) (Sun et al., 2013; Riaz and Li, 2019), in most organisms, genes are composed of deoxyribonucleic acid (DNA) arranged linearly on chromosomes. The existence of a simple life form requires at least 265 to 350 genes, whose functions are affected by internal factors and external environmental factors.

The impact of human genome research on medicine has gradually entered the clinic and changed the traditional medical model. Human genomic research has had a major impact on medicine. Genomic medicine will gradually be applied in clinical practice, and will change the traditional medical model (Liu et al., 2016a; Zhou et al., 2019; Zou, 2019; Chen et al., 2020a; Liu et al., 2020a; Li et al., 2021; Qi et al., 2021). Many believe that it will trigger the next medical revolution (Zeng et al., 2022). Genomic medicine aims to prevent diseases and apply specific gene therapies to address existing genetic defects. In-depth clinical genomic research heralds the arrival of a new era of medicine with a focus on drug optimization, and the prediction and prevention of diseases based on the genome of each patient. Genomic research has identified preferred targets for the development of new drugs (Grazziotin et al., 2015; Yan et al., 2019; Yu et al., 2020a; Zeng et al., 2020a; Liu et al., 2020b; Zhang et al., 2020; Zhuang et al., 2020; Yan et al., 2021a; Yan et al., 2021b; Deng et al., 2021; Shang et al., 2021).

So far, there are dozens of experimental methods to identify essential genes. The methods to realize essential gene identification are mainly divided into two kinds, one is to determine by means of wet experiment; The second is to use computational biology to predict the necessary genes identified by experiments. Wet experiments include: In 1995, Itaya used induced mutations to detect essential genes of *Bacillus subtilis*, Venter used global transposon mutations to identify essential genes of *Mycoplasma genitalium*, targeted gene knockout, transposon mutations, genetic imprinting, and scattered Shotgun method, RNA interference and CRISPR technology (Liu et al., 2016b; Fang et al., 2019a; Ru et al., 2019; Chen et al., 2021). And some researchers have made progress in different fields through the study of essential genes using different wet experiments. Such as: Uddin et al. conducted a comparative genomics analysis and docking studies on *Acinetobacter baumannii*, and predicted and analyzed its non-host essential genes to screen for new drug candidates (Cheng et al., 2018; Uddin et al., 2019; Wang et al., 2020). Wang et al. used the CRISPR (clustered regularly interspaced short palindromic repeats) system to construct a genome-wide single-guide RNA library and screened for genes essential for the survival and proliferation of human cancer cell lines, thus providing a cancer cell identification strategy (Lander et al., 2015). In the field of energy and chemical industries, Voshol et al. created a transposons library of *Synechococcus elongatus* PCC 7942, and then screened it to identify new target genes. The aim was to improve the production capacity of fatty acids and hydrocarbons, and the gene encoding the GTP-binding protein Era was identified as an essential gene in this process (Voshol et al., 2015; Hu et al., 2021a). Another research group suppressed genes in the chloroplasts of *Chlamydomonas* to study its essential signaling pathways, regulatory circuits, and gene functions (Rochaix and Ramundo, 2015; Fang et al., 2019b; Cheng et al., 2020; Wang et al., 2021a). Such analyses can shed light on the growth and photosynthetic processes of photosynthetic organisms, so that strategies can be developed to enhance the photosynthetic ability of engineered bacteria. Therefore, studies on essential genes have theoretical significance and also have practical application value. Therefore, it is of great theoretical significance and practical value to study essential genes.

The disadvantages of wet experiments are that they are expensive, time-consuming, inconsistent in their accuracy, and they can give different experimental results. In addition, wet experimental methods are not applicable to some bacteria. At the same time, as research progresses, it is necessary to obtain essential genes on a genome-wide scale in order to obtain as complete a set of data as possible. This presents a serious challenge to the determination of essential genes by wet assay (Yang et al., 2014; Tavasolian et al., 2020). Therefore, some scientists began to establish predictive models with higher accuracy based on the essential gene information and the biological characteristics of essential genes, combined with computer science and various mathematical algorithms, so as to realize the rapid identification of essential genes. Therefore, greatly reducing unnecessary time and capital consumption.

ESSENTIAL GENE DATABASE

Bacteria are abundant both in terms of the number of species and the size of populations. Only a small proportion of bacteria have been fully sequenced. Currently, the following databases provide essential genetic data:

1) NCBI: Holds genomic data for more than 35,000 bacteria. Essential/non-essential genes have been fully determined for only a few bacteria.
2) The Database of Essential Genes (DEG): This database has been developed and maintained by Tianjin University. The latest version (DEG 15.2) contains essential genetic information for 41 kinds of bacteria (48 groups) (Luo et al., 2014), 26 essential gene datasets for eukaryotes and one essential gene dataset for archaea.
3) CEG (Cluster of Essential Genes): Holds information about essential homologous gene clusters for 29 kinds of bacteria (Ye et al., 2013).

4) OGEE (Online Gene Essentiality Database): Holds information about essential genes for 39 kinds of bacteria (Chen et al., 2011).
5) PEC database: Holds information for essential genes in *Escherichia coli*, including relevant structural information and gene function information (Rivas et al., 2015).

METHODS FOR IDENTIFYING ESSENTIAL GENES

The study of essential genes can also obtain potential drug targets, which can be used to develop antibacterial drugs to resist the invasion of pathogenic microorganisms. Therefore, it is of great practical value to study the theoretical and computational prediction methods of essential genes. Theoretical prediction is a common method of comparative genomics, which is to understand the characteristics of genes to be tested by comparing with the structure of known genes. The disadvantage of this method is that it can not accurately determine all the necessary genes of a certain microorganism, and it is more difficult to analyze eukaryotes. Computational prediction is a potential computational prediction, which uses a variety of biological characteristics of the research object, combined with statistical and classification prediction algorithms, to predict essential genes. With the rapid development of technology, omics data such as protein structure and interaction are widely used in essential gene prediction analysis. Therefore, computational biology methods can be used to provide suitable candidate target genes. The main steps are as follows: First, algorithms and tools are developed to predict essential genes of pathogenic bacteria through computational modeling. This step yields a set of candidate essential genes. Second, an essential gene prediction model is constructed. The model includes information on conserved regions of genes, and this allows for the identification of essential genes showing homology among multiple species. Next, the sequences of the obtained essential genes are compared with the genome sequences of humans or other mammals, and those that are too similar are filtered out. The ones that remain are thus identified as candidate targets for the development of new therapeutic drugs. These genes can be targeted to design and screen effective therapeutic drugs. Compared with traditional drug design strategies, the computational model is more directed, so it can shorten the development time for new drugs (Yu et al., 2020b; Zeng et al., 2020b; Huo et al., 2020; Li et al., 2020; Cheng et al., 2021b; Wang et al., 2021b; Dong et al., 2021). For example, Paul et al. used a biological metabolic network of leishmaniasis and deletion mutations designed using bioinformatics methods to identify a collection of essential proteins for this disease (Ao et al., 2021; Hu et al., 2021b), and this method was more than five times more efficient than the method of randomly selecting potential drug targets (Stanly Paul et al., 2014; Chiu et al., 2020; Wang et al., 2020). As full genome sequences are obtained for more organisms, and with the continuing development of functional genomics research, more attention is being paid to the relationships among genes (Wang et al., 2019), proteins, and phenotypes (Kitano, 2002). Proteins are indispensable components of cellular structures, and participate in a wide range of processes that affect growth, physiology, and development (Eisenberg et al., 2000). In addition, proteins function in particular subcellular compartments (Huh et al., 2003).

Gene Computation and Prediction Methods Based on Communication Coding Theory

In 1995, Itaya studied the minimum number of chromosomal loci required for the survival of *B. subtilis* using an induced mutation method. Mutations of only six out of 79 randomly selected chromosomal loci prevented the formation of bacterial colonies. Thus, it was inferred that genes at these six loci were essential for the survival of *B. subtilis* (Itaya, 1995). Later, other researchers determined the essential genes of *Mycoplasma* genitalium and explored the composition of the minimum gene set with experimental methods (HutchisonPeterson et al., 1999). Ongoing research on the minimum gene set has continuously discovered new essential genes, thus opening the door to essential gene characterization and further in-depth research.

Hwang et al. studied the topological properties of essential and non-essential genes in protein interaction networks with *Saccharomyces cerevisiae* and *E. coli* as the research objects. Then, they predicted essential genes using machine learning methods and on the basis of protein interaction networks and sequence information (Hwang et al., 2009; Song et al., 2021). Deng et al. used machine learning methods, combined with three types of 8 characteristic parameters such as subcellular localization, phylogenetic information and expression levels, and co-expression networks to predict the essential genes of 4 species including *E. coli*. However, the effectiveness of this method relies on similar feature distributions in the research objects (Deng et al., 2011). Yang et al. studied human essential genes by constructing a classifier based on 28 protein interaction network topological features and 22 biological features (Yang et al., 2014). Yu et al. identified six fractal features of DNA and protein sequences of DEG bacteria, and classified sequences using naïve Bayes and random forest methods. Arun et al. studied the relationships among gene conservation, repetition, constitutive expression, and gene essentiality in *E. coli* (Arun et al., 2016). Although those studies made considerable progress, there were still some problems: few species were studied; the universality of prediction models and analysis results and prediction accuracy needed to be improved; and characteristic parameters that effectively describe essential genes needed to be identified.

Considering that information transmission and coding are similar between biological systems and modern communication systems, communication coding theory can be a useful tool for essential gene analysis and prediction of gene functions. The definition of coding problem in DNA computation was first proposed by Garzon and Deaton et al. in literature network, and then the complete definition of coding problem in DNA

computation was given in 2004 after refining and summarizing (Garzon and Deaton, 2004).

The steps involved in essential gene prediction based on the theory of communication coding are as follows: First, a genetic sequence analysis model is constructed based on communication coding theory. The model includes sequence analysis based on simple coding models (such as block codes, convolutional codes) and cascaded and mixed coding models. In error correction coding, convolutional codes (based on channel coding methods but with better performance), have been proven to be as effective as block codes in theory and in practice. The next step is to extract the characteristic parameters that describe essential genes. The combination of the theoretical model of communication coding for analyzing genetic sequences and genetic database information can reveal correlations between information units in DNA sequences at different scales in the coding sense, and extract the characteristic parameters of essential genes. Then, based on the features of the coding meaning and the omics data, the necessary gene calculation and prediction are carried out and combined with the determined gene data in the database, the performance optimization parameters of the established model are evaluated. Finally, the noise or redundancy is removed from the acquired features, and the key features are screened out, which are called "essential features," and the feature set formed by these "essential features" is called "minimum feature group." The "minimum feature group" should be constructed within the range of existing features. This will simplify and optimize the model and the analytical process, and improve its analytical efficiency and universality while ensuring prediction accuracy.

Analysis and Design of Gene Computational Prediction Models Based on Communication Coding Theory

According to the research of gene prediction method based on communication coding theory in 3.1 summary, we use the following research steps to establish a prediction model:

1. Check published studies to understand the current research situation.
2. Collect and preprocess datasets from open databases such as DEG, NCBI, KEGG, and Uniprot, and clean and sort the collected data.
3. Model research and feature extraction. This step includes encoding - based sequence analysis model and feature extraction. 1) Sequence analysis model based on coding; If the genetic sequence is analyzed as a sequence with certain information and coding characteristics, the biological sequence (DNA, RNA, or amino acid sequence) needs to be expressed in a form that is suitable for analysis, regardless of which analytical model is used (Dao et al., 2021). In this context, researchers have proposed digital mapping methods, graphic expression methods, and geometric expression methods. Taking a DNA sequence as an example, the digital mapping method expresses the four bases A, G, C, and T as different values, in a form that is convenient for computer processing, including integer expression and plural expression (Rosen, 2006; Liu et al., 2021a). The graphic and geometric expression methods express each base as a vector in space, and then connect the corresponding vectors to map a curve in the space. These methods allow for visualization of the sequence structure, and the local and overall characteristics of the sequence. Visualization of sequences is also useful for sequence comparisons and sequence similarity analyses. Once the biological sequence has been expressed and visualized, models based on convolutional code and block code models can be designed, adjusted, and modified according to analytical results and biological significance. Models can be analyzed using cascaded codes and mixed codes on the basis of simple model analysis. 2) Extract features of essential genes: According to different observation scales, the basic information units are set as bases, codons, genes, or the whole genome. Then, the association between essential genes and information units is investigated. This results in the extraction of characteristic parameters that reflect the essentiality of genes in combination with the parameters of the coding model (code length, generator matrix, constraint length, and code distance).
4. Computationally predict essential genes: After completing the tasks in Step 3, essential genes are predicted using machine learning algorithms. The best model is selected through performance evaluation. Redundant features are removed and essential features are screened to identify the "minimum feature group."
5. Conduct performance testing and feedback optimization: Test the analytical performance of the prediction model based on cross validation and other methods (Jiang et al., 2013; Xu et al., 2021), using the essential/non-essential genes listed in DEG and other databases as training samples and verification samples.
6. Develop software to implement analysis and prediction algorithms and visualize the results. Package the final model into an executable program module with a functional interactive webpage. Establish an online service platform to provide online data analysis services.

The process of the research method is illustrated as a roadmap in **Figure 1**.

KEY PROTEIN IDENTIFICATION METHOD

A gene, the basic unit of heredity, is the DNA fragment required to produce a functional RNA or a polypeptide chain that forms into a functional protein. Therefore, genes and proteins are inextricably linked. For example, to explore the characteristics of essential genes, it is necessary to integrate and summarize the genomic metabolome, proteome and other omics data, and extract their combined data features, including metabolic pathways, evolutionary conservation, protein domains, and protein interactions (Wang et al., 2018; Zhang et al., 2018; Chen et al., 2019; Chen et al., 2020b; Yu et al., 2020c). Thus,

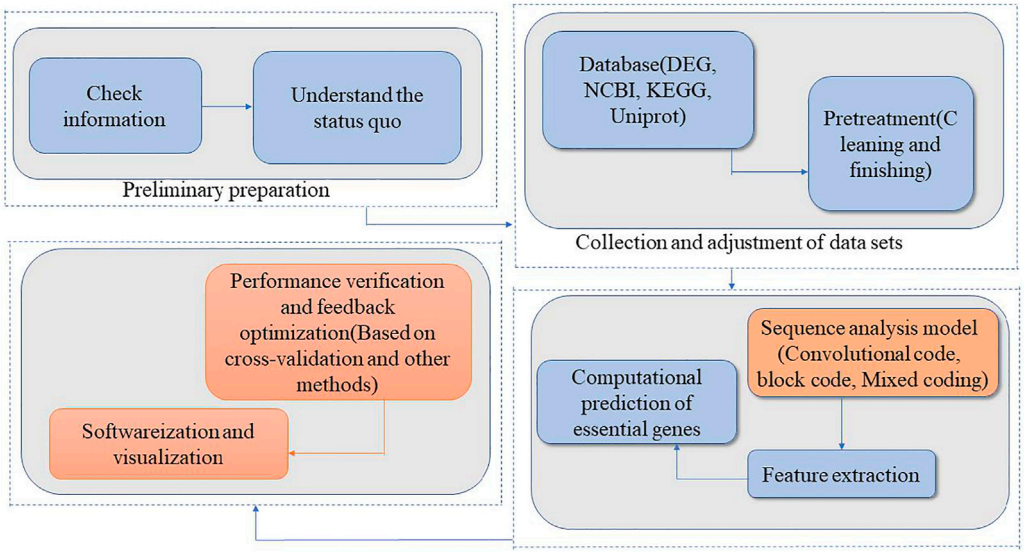


FIGURE 1 | Technical roadmap of gene computational prediction model based on communication coding theory.

it is important to identify proteins and collect all relevant information about their sequence, structure, localization, and function.

Protein is the material basis of life, is an organic macromolecule, is the basic organic matter that constitutes cells, and is the main undertaker of life activities (Wei et al., 2014; Wei et al., 2017; Guo et al., 2020; Tao et al., 2020; Wei et al., 2020; Guo et al., 2021). The human body contains many types of proteins with different properties and functions. All proteins have something in common: they are composed of 20 amino acids in different proportions, and they are constantly metabolized and renewed in the body. Many methods for identifying protein complexes based on protein interaction networks have been proposed. Some proteins are essential for the survival and reproduction of organisms, and for essential biological processes. Without them, organisms will have serious defects in their growth and development, or may even die. Thus, these are known as essential proteins. At the same time, research show many proteins perform specific biological functions only after they participate in the formation of protein complexes and interact with other proteins in the complex, suggesting that protein interactions are related to protein complexes. Therefore, this part takes the key proteins as the research object and uses the methods based on the subcellular protein interaction network and subcellular importance to identify the key proteins (Cheng et al., 2021a; Zulfiqar et al., 2021).

The centrality algorithm, a method used to identify key proteins in a protein-protein interaction network, is often used to quantify the contribution of a specific node in the graph and its impact on the network and it's based on the law of central-lethality (Liu et al., 2020c; Zhai et al., 2020). Nodes with higher centrality values are usually regarded as key nodes in the network. Commonly used graph-based centrality algorithms include degree centrality, betweenness centrality, tight centrality, subgraph centrality, node cluster centrality, and local average connection centrality. However, this method cannot fully assess the criticality of proteins because it does not take temporal and spatial characteristics of protein interactions into account. Therefore, the subcellular location information can be used to construct a subcellular protein interaction network, and a key protein identification method based on the subcellular protein interaction network (LSED) is proposed. LSED is first based on a given global protein interaction network and protein The subcellular location information of each subcellular compartment is constructed to construct a subcellular compartment protein interaction network (PSLIN). Then, its credibility is calculated based on the size of the protein interaction network in each subcellular interval. Next, the centrality scores of the proteins in the subnets of the protein interaction network for each subcellular interval are calculated with a centrality method. Finally, the Localization-specific Centrality Score is calculated for each protein based on its centrality score in the protein interaction network for different subcellular intervals and the credibility of the network. By comparing the prediction accuracy of LSED method on PSLIN with that of centrality method on global protein interaction network, a multi-species average accuracy index (AKAcc) can be proposed, which can more comprehensively evaluate the prediction accuracy of various methods on multiple species. The key protein recognition methods mentioned in this paper are shown in **Table 1**.

Because different subcellular intervals differ in importance, protein interactions in different subcellular intervals also differ in their importance. Therefore, the importance of protein interactions can be estimated based on the importance of subcellular intervals. A Centrality-based Independent Cascade (CIC) based on the importance of subcellular intervals has been proposed to detect essential proteins. For two interacting proteins

TABLE 1 | Key protein identification methods.

Methods	Principle
Degree centrality	Based on protein interactions, a reliable PPI network can be constructed by combining other protein biological information. Subsequently, key protein identification is carried out through the centrality method related to network topology
Betweenness centrality	
Tight centrality	
Subgraph centrality	
Node cluster centrality	
Local average connection centrality	
PSLIN	Use subcellular location information to construct a subcellular protein interaction network
LSED	Key protein identification method based on protein interaction network in subcellular compartment
CIC	Centrality method based on the importance of subcellular intervals

TABLE 2 | Identification methods of protein complexes.

Methods	Specific type
Graph clustering method	RNSC algorithm based on graph partition
	Density-based local search algorithm MCODE
Hierarchical clustering method based on similarity or distance	Condensation algorithm
	Division algorithm

u and v, the importance of the interaction (u,v) is defined as the maximum importance of the subcellular interval where the two proteins cooccur. A weighted protein interaction network is constructed based on the importance value of the interaction, and the CIC centrality method is used to estimate the criticality of proteins in this network. Thus, the CIC centrality score of a protein depends on the importance of its interactions in different subcellular intervals. Finally, the predictive performance of the CIC method is compared with those of other centrality algorithms. In addition, researchers conducted experiments on the protein interaction network of yeast, human, mouse and fruit fly to compare the prediction performance of CIC method with other centrality methods, including topology-based centrality methods and methods integrating other biological knowledge.

There are close relationships between protein complexes and essential proteins. Studies have shown that the criticality of a protein is not only determined by a single protein node, but often by the function of protein complexes. According to experimental data, essential proteins tend to aggregate in large amounts in certain complexes. Machine learning methods can be used to identify such protein complexes. At present, there is no strict mathematical expression or unified definition for the subgraph model corresponding to protein complexes and functional modules in interaction networks. However, some graph-clustering methods are effective for identifying protein complexes and functional modules, such as the graph-clustering algorithm RNSC, and density-based algorithms such as MCODE. Some studies have found that most proteins in the same protein complex have similar or identical functions, so some hierarchical clustering methods based on similarity or distance are also used to identify protein complexes. Hierarchical clustering algorithm can be divided into two categories, condensation algorithm and splitting algorithm. Among them, the G-N algorithm, a classic splitting algorithm, removes the edge with the highest betweenness and recalculates the betweenness of all edges. When at least two of the generated subgraphs meet the definition of the module after removing the edges, a corresponding phylogenetic tree can be drawn. And the HC-PIN method uses weighted edge clustering coefficient to do fast agglomerative clustering. The clustering idea is to merge the two clusters if there is an edge with the highest edge clustering coefficient between two clusters. At the same time, **Table 2** shows various methods for identifying protein complexes.

EXISTING PROBLEMS AND RESEARCH PROSPECTS

In this paper, we have discussed some of the methods used to analyze and process biological sequence information, and introduced the use of coding theory in communication engineering to calculate and predict essential genes. However, some problems are yet to be solved. First, limited data are available for the preliminary analysis of essential genes. The DEG database contains information for a limited number of species, and this makes it difficult to test and validate analytical models. In the identification and function prediction of proteins, it is seldom possible to determine protein function based on a single experiment, because protein function is also affected by environmental factors. In addition, due to various biological, budgetary, or ethical factors, experiments cannot be performed on some organisms. Experiments conducted *in vitro* may not truly reflect the activity of proteins in the body. Moreover, the datasets in biological databases have problems such as noise, deviation, and missing data (Su et al., 2019; Liu et al., 2020d; Liu et al., 2021b; Su et al., 2021).

There are many commonalities between the transmission of genetic information and communication information. Therefore, methods based on communication coding theory can be applied to identify essential genes in bacteria and other organisms. This

provides a new perspective and approach for analyzing and predicting essential genes. The information obtained in these analyses has applications in disease prevention, diagnosis, and treatment.

AUTHOR CONTRIBUTIONS

Conceptualization, DC and LW; data collection or analysis, YG and YJ; validation, YG and YJ; writing—original draft preparation, YG and YJ; writing—review and editing, YG. All authors have read and agreed to the published version of the manuscript.

FUNDING

The work was supported by the National Natural Science Foundation of China (62072385), and the Special Science Foundation of Quzhou (2021D004).

ACKNOWLEDGMENTS

Thanks to the guidance of my tutor and the joint efforts of other authors, the success of this article is the result of everyone's joint efforts.

REFERENCES

Ao, C., Yu, L., and Zou, Q. (2021). Prediction of Bio-Sequence Modifications and the Associations with Diseases. *Brief. Funct. genomics* 20 (1), 1–18. doi:10.1093/bfgp/elaa023

Arun, P., Miryala, S. K., Chattopadhyay, S., Thiyyagura, K., Bawa, P., Bhattacharjee, M., et al. (2016). Identification and Functional Analysis of Essential, Conserved, Housekeeping and Duplicated Genes. *FEBS Lett.* 590, 1428–1437. doi:10.1002/1873-3468.12192

Chen, C., Zhang, Q., Ma, Q., and Yu, B. (2019). LightGBM-PPI: Predicting Protein-Protein Interactions through LightGBM with Multi-Information Fusion. *Chemometrics Intell. Lab. Syst.* 191, 54–64. doi:10.1016/j.chemolab.2019.06.003

Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P. J., Ma, Q., et al. (2020). Improving Protein-Protein Interactions Prediction Accuracy Using XGBoost Feature Selection and Stacked Ensemble Classifier. *Comput. Biol. Med.* 123, 103899. doi:10.1016/j.compbiomed.2020.103899

Chen, S., Liu, Z., Li, M., Huang, Y., Wang, M., Zeng, W., et al. (2020). Potential Prognostic Predictors and Molecular Targets for Skin Melanoma Screened by Weighted Gene Co-expression Network Analysis. *Curr. Gene Ther.* 20 (1), 5–14. doi:10.2174/1566523220666200516170832

Chen, W.-H., Minguez, P., Lercher, M. J., and Bork, P. (2011). OGEE: an Online Gene Essentiality Database. *Nucleic Acids Res.* 40 (D1), D901–D906. doi:10.1093/nar/gkr986

Chen, Y., Ma, T., Yang, X., Wang, J., Song, B., and Zeng, X. (2021). MUFFIN: Multi-Scale Feature Fusion for Drug–Drug Interaction Prediction. *Bioinformatics* 37, 2651–2658. doi:10.1093/bioinformatics/btab169

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics* 34 (11), 1953–1956. doi:10.1093/bioinformatics/bty002

Cheng, L., Qi, C., Yang, H., Lu, M., Cai, Y., Fu, T., et al. (2021). gutMGene: a Comprehensive Database for Target Genes of Gut Microbes and Microbial Metabolites. *Nucleic Acids Res.*, gkab786. doi:10.1093/nar/gkab786

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a Comprehensive Database for Dysbiosis of the Gut Microbiota in Disorders and Interventions. *Nucleic Acids Res.* 48 (D1), D554–D560. doi:10.1093/nar/gkz843

Cheng, Y., Gong, Y., Liu, Y., Song, B., and Zou, Q. (2021). Molecular Design in Drug Discovery: a Comprehensive Review of Deep Generative Models. *Brief. Bioinform.* 22, bbab344. doi:10.1093/bib/bbab344

Chiu, T. P., Xin, B., Markarian, N., Wang, Y., and Rohs, R. (2020). TFBSshape: an Expanded Motif Database for DNA Shape Features of Transcription Factor Binding Sites. *Nucleic Acids Res.* 48 (D1), D246–D255. doi:10.1093/nar/gkz970

Dao, F. Y., Lv, H., Zhang, D., Zhang, Z. M., Liu, L., and Lin, H. (2021). DeepYY1: a Deep Learning Approach to Identify YY1-Mediated Chromatin Loops. *Brief Bioinform* 22 (4). bbaa356, doi:10.1093/bib/bbaa356

Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., et al. (2011). Investigating the Predictability of Essential Genes across Distantly Related Organisms Using an Integrative Approach. *Nucleic Acids Res.* 39 (3), 795–807. doi:10.1093/nar/gkq784

Deng, L., Li, W., and Zhang, J. (2021). LDAH2V: Exploring Meta-Paths across Multiple Networks for lncRNA-Disease Association Prediction. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18 (4), 1572–1581. doi:10.1109/tcbb.2019.2946257

Dong, J., Zhao, M., Liu, Y., Su, Y., and Zeng, X. (2021). Deep Learning in Retrosynthesis Planning: Datasets, Models and Tools. *Brief. Bioinform.*, bbab391. doi:10.1093/bib/bbab391

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000). Protein Function in the post-genomic Era. *Nature* 405 (6788), 823–826. doi:10.1038/35015694

Fang, M., Lei, X., and Guo, L. (2019). A Survey on Computational Methods for Essential Proteins and Genes Prediction. *Curr. Bioinformatics* 14 (3), 211–225. doi:10.2174/1574893613666181112150422

Fang, S., Pan, J., Zhou, C., Tian, H., He, J., Shen, W., et al. (2019). Circular RNAs Serve as Novel Biomarkers and Therapeutic Targets in Cancers. *Curr. Gene Ther.* 19 (2), 125–133. doi:10.2174/1566523218666181109142756

Garzon, M. H., and Deaton, R. J. (2004). Codeword Design and Information Encoding in DNA Ensembles. *Nat. Comput.* 3 (3), 253–292. doi:10.1023/b:naco.0000036818.27537.c9

Grazziotin, A. L., Vidal, N. M., and Venancio, T. M. (2015). Uncovering Major Genomic Features of Essential Genes in Bacteria and a Methanogenic Archaea. *Febs J.* 282, 3395–3411. doi:10.1111/febs.13350

Guo, Y., Yan, K., Lv, H., and Liu, B. (2021). PreTP-EL: Prediction of Therapeutic Peptides Based on Ensemble Learning. *Brief Bioinform* 22, bbab358. doi:10.1093/bib/bbab358

Guo, Z., Wang, P., Liu, Z., and Zhao, Y. (2020). Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction. *Front. Bioeng. Biotechnol.* 8, 584807. doi:10.3389/fbioe.2020.584807

Hu, Y., Qiu, S., and Cheng, L. (2021). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 7036592. doi:10.1155/2021/7036592

Hu, Y., Sun, J. Y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (2021). rs1990622 Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19 (1), 11. doi:10.1186/s12916-020-01883-5

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 Regulates TSPAN13 Expression and Contributes to Alzheimer's Disease with Sex Differences. *Brain.* 143 (11), e95. doi:10.1093/brain/awaa302

Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., et al. (2003). Global Analysis of Protein Localization in Budding Yeast. *Nature* 425, 686–691. doi:10.1038/nature02026

Huo, Y., Xin, L., Kang, C., Wang, M., Ma, Q., and Yu, B. (2020). SGL-SVM: A Novel Method for Tumor Classification via Support Vector Machine with Sparse Group Lasso. *J. Theor. Biol.* 486, 110098. doi:10.1016/j.jtbi.2019.110098

HutchisonIII, Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., et al. (1999). Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science* 286 (5447), 2165–2169. doi:10.1126/science.286.5447.2165

Hwang, Y.-C., Lin, C.-C., Chang, J.-Y., Mori, H., Juan, H.-F., and Huang, H.-C. (2009). Predicting Essential Genes Based on Network and Sequence Analysis. *Mol. Biosyst.* 5 (12), 1672–1678. doi:10.1039/b900611g

Itaya, M. (1995). An Estimation of Minimal Genome Size Required for Life. *Febs Lett.* 362 (3), 257–260. doi:10.1016/0014-5793(95)00233-y

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting Human microRNA-Disease Associations Based on Support Vector Machine. *Ijdmb* 8 (3), 282–293. doi:10.1504/ijdmb.2013.056078

Kitano, H. (2002). Computational Systems Biology. *Nature* 420 (6912), 206–210. doi:10.1038/nature01254

Lander, E. S., Wang, T., Hughes, N. W., and Jenny, W. (2015). Identification and Characterization of Essential Genes in the Human Genome. *Science* 350, 1096–1101. doi:10.1126/science.aac7041

Li, F., Luo, M., Zhou, W., Li, J., Jin, X., Xu, Z., et al. (2020). Single Cell RNA and Immune Repertoire Profiling of COVID-19 Patients Reveal Novel Neutralizing Antibody. *Protein Cell* 12, 751–755. doi:10.1007/s13238-020-00807-6

Li, H.-L., Pang, Y.-H., and Liu, B. (2021). BioSeq-BLM: a Platform for Analyzing DNA, RNA and Protein Sequences Based on Biological Language Models. *Nucleic Acids Res.*, gkab829. doi:10.1093/nar/gkab829

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol. Ther. - Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., and Deng, L. (2020). DrugCombDB: a Comprehensive Database of Drug Combinations toward the Discovery of Combinatorial Therapy. *Nucleic Acids Res.* 48 (D1), D871–D881. doi:10.1093/nar/gkz1007

Liu, J., Liu, S., Liu, C., Zhang, Y., Pan, Y., Wang, Z., et al. (2021). Nabe: an Energetic Database of Amino Acid Mutations in Protein–Nucleic Acid Binding Interfaces. *Database* 2021, baab050. doi:10.1093/database/baab050

Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., Zou, Q., et al. (2020). Computational Methods for Identifying the Critical Nodes in Biological Networks. *Brief. Bioinform.* 21 (2), 486–497. doi:10.1093/bib/bbz011

Liu, X., Yang, J., Zhang, Y., Fang, Y., Wang, F., Wang, J., et al. (2016). A Systematic Study on Drug-Response Associated Genes Using Baseline Gene Expressions of the Cancer Cell Line Encyclopedia. *Sci. Rep.* 6, 22811. doi:10.1038/srep22811

Liu, Y., Chen, D., Su, R., Chen, W., and Wei, L. (2020). iRNA5hmC: The First Predictor to Identify RNA 5-Hydroxymethylcytosine Modifications Using Machine Learning. *Front. Bioeng. Biotechnol.* 8, 227. doi:10.3389/fbioe.2020.00227

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2016). Inferring microRNA-Disease Associations by Random Walk on a Heterogeneous Network with Multiple Data Sources. *Ieee/acm Trans. Comput. Biol. Bioinform* 14 (4), 905–915. doi:10.1109/TCBB.2016.2550432

Liu, Y., Zhang, X., Zou, Q., and Zeng, X. (2021). Minirmd: Accurate and Fast Duplicate Removal Tool for Short Reads via Multiple Minimizers. *Bioinformatics* 37 (11), 1604–1606. doi:10.1093/bioinformatics/btaa915

Luo, H., Lin, Y., Gao, F., Zhang, C.-T., and Zhang, R. (2014). DEG 10, an Update of the Database of Essential Genes that Includes Both Protein-Coding Genes and Noncoding Genomic Elements: Table 1. *Nucl. Acids Res.* 42, D574–D580. doi:10.1093/nar/gkt1131

Qi, C., Wang, C., Zhao, L., Zhu, Z., Wang, P., Zhang, S., et al. (2021). SCovid: Single-Cell Atlases for Exposing Molecular Characteristics of COVID-19 across 10 Human Tissues. *Nucleic Acids Res.*, gkab881. doi:10.1093/nar/gkab881

Riaz, F., and Li, D. (2019). Non-coding RNA Associated Competitive Endogenous RNA Regulatory Network: Novel Therapeutic Approach in Liver Fibrosis. *Curr. Gene Ther.* 19 (5), 305–317. doi:10.2174/1566523219666191107113046

Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., et al. (2015). Effect of Predicted Protein-Truncating Genetic Variants on the Human Transcriptome. *Science* 348 (6235), 666–669. doi:10.1126/science.1261877

Rochaix, J.-D., and Ramundo, S. (2015). Conditional Repression of Essential Chloroplast Genes: Evidence for New Plastid Signaling Pathways. *Biochim. Biophys. Acta (Bba) - Bioenerg.* 1847 (9), 986–992. doi:10.1016/j.bbabio.2014.11.011

Rosen, G. (2006). Examining Coding Structure and Redundancy in DNA. *IEEE Eng. Med. Biol. Mag.* 25 (1), 62–68. doi:10.1109/memb.2006.1578665

Ru, X., Cao, P., Li, L., and Zou, Q. (2019). Selecting Essential MicroRNAs Using a Novel Voting Method. *Mol. Ther. - Nucleic Acids* 18, 16–23. doi:10.1016/j.omtn.2019.07.019

Shang, Y., Gao, L., Zou, Q., and Yu, L. (2021). Prediction of Drug-Target Interactions Based on Multi-Layer Network Representation Learning. *Neurocomputing* 434, 80–89. doi:10.1016/j.neucom.2020.12.068

Song, B., Li, F., Liu, Y., and Zeng, X. (2021). Deep Learning Methods for Biomedical Named Entity Recognition: a Survey and Qualitative Comparison. *Brief. Bioinform.* 22, bbab282. doi:10.1093/bib/bbab282

Stanly Paul, M. L., Kaur, A., Geete, A., and Elizabeth Sobhia, M. (2014). Essential Gene Identification and Drug Target Prioritization in Leishmania Species. *Mol. Biosyst.* 10 (5), 1184–1195. doi:10.1039/c3mb70440h

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A Deep forest Model to Predict Anti-cancer Drug Response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

Su, W., Liu, M.-L., Yang, Y.-H., Wang, J.-S., Li, S.-H., Lv, H., et al. (2021). PPD: A Manually Curated Database for Experimentally Verified Prokaryotic Promoters. *J. Mol. Biol.* 433 (11), 166860. doi:10.1016/j.jmb.2021.166860

Sun, H., Yang, J., Zhang, T., Long, L. P., Jia, K., Yang, G., et al. (2013). Using Sequence Data to Infer the Antigenicity of Influenza Virus. *mBio* 44 (4), e00230–13. doi:10.1128/mBio.00230-13

Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD. *Comput. Math. Methods Med.* 2020, 8926750. doi:10.1155/2020/8926750

Tavasolian, F., Hosseini, A. Z., Soudi, S., and Naderi, M. (2020). miRNA-146a Improves Immunomodulatory Effects of MSC-Derived Exosomes in Rheumatoid Arthritis. *Curr. Gene Ther.* 20 (4), 297–312. doi:10.2174/1566523220666200916120708

Uddin, R., Masood, F., Azam, S. S., and Wadood, A. (2019). Identification of Putative Non-host Essential Genes and Novel Drug Targets against Acinetobacter Baumannii by In Silico Comparative Genome Analysis. *Microb. Pathogenesis* 128, 28–35. doi:10.1016/j.micpath.2018.12.015

Voshol, G. P., Meyer, V., and van den Hondel, C. A. (2015). GTP-binding Protein Era: a Novel Gene Target for Biofuel Production. *BMC Biotechnol.* 15 (1), 21. doi:10.1186/s12896-015-0132-1

Wang, J., Liu, X., Shen, S., Deng, L., and Liu, H. (2021). DeepDDS: Deep Graph Neural Network with Attention Mechanism to Predict Synergistic Drug Combinations. *Brief. Bioinform.*, bbab390. doi:10.1093/bib/bbab390

Wang, J., Shi, Y., Wang, X., and Chang, H. (2020). A Drug Target Interaction Prediction Based on LINE-RF Learning. *Curr. Bioinformatics* 15 (7), 750–757. doi:10.2174/1574893615666191227092453

Wang, N., Zhang, J., and Liu, B. (2021). IDRBP-PPCT: Identifying Nucleic Acid-Binding Proteins Based on Position-specific Score Matrix and Position-specific Frequency Matrix Cross Transformation. *Ieee/acm Trans. Comput. Biol. Bioinf.* (99), 1. doi:10.1109/TCBB.2021.3069263

Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., and Ma, Q. (2018). Protein-protein Interaction Sites Prediction by Ensemble Random Forests with Synthetic Minority Oversampling Technique. *Bioinformatics* 35 (14), 2395–2402. doi:10.1093/bioinformatics/bty995

Wang, Y., Yang, S., Zhao, J., Du, W., Liang, Y., Wang, C., et al. (2019). Using Machine Learning to Measure Relatedness between Genes: A Multi-Features Model. *Sci. Rep.* 9 (1), 4192. doi:10.1038/s41598-019-40780-7

Wei, L., Tang, J., and Zou, Q. (2017). SkipCPP-Pred: an Improved and Promising Sequence-Based Predictor for Predicting Cell-Penetrating Peptides. *Bmc Genomics* 18, 742. doi:10.1186/s12864-017-4128-1

Wei, L., He, W., Malik, A., Su, R., Cui, L., and Manavalan, B. (2020). Computational Prediction and Interpretation of Cell-specific Replication Origin Sites from Multiple Eukaryotes by Exploiting Stacking Framework. *Brief. Bioinform.* 22, bbaa275. doi:10.1093/bib/bbaa275

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. *Ieee/acm Trans. Comput. Biol. Bioinf.* 11 (1), 192–201. doi:10.1109/tcbb.2013.146

Xu, Z., Luo, M., Lin, W., Xue, G., Wang, P., Jin, X., et al. (2021). DLpTCR: an Ensemble Deep Learning Framework for Predicting Immunogenic Peptide Recognized by T Cell Receptor. *Brief Bioinform* 22, bbab335. doi:10.1093/bib/bbab335

Yan, K., Fang, X., Xu, Y., and Liu, B. (2019). Protein Fold Recognition Based on Multi-View Modeling. *Bioinformatics* 35 (17), 2982–2990. doi:10.1093/bioinformatics/btz040

Yan, K., Wen, J., Liu, J.-X., Xu, Y., and Liu, B. (2021). Protein Fold Recognition by Combining Support Vector Machines and Pairwise Sequence Similarity Scores. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18, 2008–2016. doi:10.1109/TCBB.2020.2966450

Yan, K., Wen, J., Xu, Y., and Liu, B. (2021).Protein Fold Recognition Based on Auto-Weighted Multi-View Graph Embedding Learning Model, Proceeding of the IEEE/ACM Trans. Comput. Biol. and Bioinf., April 2020. IEEE, 1. doi:10.1109/TCBB.2020.2991268

Yang, J., Huang, T., Huang, T., Petralia, F., Long, Q., Zhang, B., et al. (2015). Synchronized Age-Related Gene Expression Changes across Multiple Tissues in

Human and the Link to Complex Diseases. *Sci. Rep.* 5, 15145. doi:10.1038/srep15145

Yang, L., Wang, J., Wang, H., Lv, Y., Zuo, Y., Li, X., et al. (2014). Analysis and Identification of Essential Genes in Humans Using Topological Properties and Biological Information. *Gene* 551 (2), 138–151. doi:10.1016/j.gene.2014.08.046

Ye, Y. N., Hua, Z. G., Huang, J., Rao, N., and Guo, F. B. (2013). CEG: a Database of Essential Gene Clusters. *BMC Genomics* 14 (1), 1–10. doi:10.1186/1471-2164-14-769

Yu, B., Chen, C., Zhou, H., Liu, B., and Ma, Q. (2020). GTB-PPI: Predict Protein-Protein Interactions Based on L1-Regularized Logistic Regression and Gradient Tree Boosting. *Genomics, Proteomics & Bioinformatics* 18 (5), 582–592. doi:10.1016/j.gpb.2021.01.001

Yu, L., Xu, F., and Gao, L. (2020). Predict New Therapeutic Drugs for Hepatocellular Carcinoma Based on Gene Mutation and Expression. *Front. Bioeng. Biotechnol.* 8, 8. doi:10.3389/fbioe.2020.00008

Yu, L., Zhou, D., Gao, L., and Zha, Y. (2020). Prediction of Drug Response in Multilayer Networks Based on Fusion of Multiomics Data. *Methods* 192, 85–92. doi:10.1016/j.ymeth.2020.08.006

Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., Hou, Y., et al. (2020). Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning. *J. Proteome Res.* 19 (11), 4624–4636. doi:10.1021/acs.jproteome.0c00316

Zeng, X., Tu, X., Liu, Y., Fu, X., and Su, Y. (2022). Toward Better Drug Discovery with Knowledge Graph. *Curr. Opin. Struct. Biol.* 72, 114–126. doi:10.1016/j.sbi.2021.09.003

Zeng, X., Zhu, S., Hou, Y., Zhang, P., Li, L., Li, J., et al. (2020). Network-based Prediction of Drug-Target Interactions Using an Arbitrary-Order Proximity Embedded Deep forest. *Bioinformatics* 36 (9), 2805–2812. doi:10.1093/bioinformatics/btaa010

Zhai, Y., Chen, Y., Teng, Z., and Zhao, Y. (2020). Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions. *Front. Cel Dev. Biol.* 8, 591487. doi:10.3389/fcell.2020.591487

Zhang, F., Ma, A., Wang, Z., Ma, Q., Liu, B., Huang, L., et al. (2018). A Central Edge Selection Based Overlapping Community Detection Algorithm for the Detection of Overlapping Structures in Protein-Protein Interaction Networks. *Molecules* 23 (10), 2633. doi:10.3390/molecules23102633

Zhang, S., Su, M., Sun, Z., Lu, H., and Zhang, Y. (2020). The Signature of Pharmaceutical Sensitivity Based on ctDNA Mutation in Eleven Cancers. *Exp. Biol. Med. (Maywood)* 245 (8), 720–732. doi:10.1177/1535370220906518

Zhou, L.-Y., Qin, Z., Zhu, Y.-H., He, Z.-Y., and Xu, T. (2019). Current RNA-Based Therapeutics in Clinical Trials. *Curr. Gene Ther.* 19 (3), 172–196. doi:10.2174/1566523219666190719100526

Zhuang, J., Dai, S., Zhang, L., Gao, P., Han, Y., Tian, G., et al. (2020). Identifying Breast Cancer-Induced Gene Perturbations and its Application in Guiding Drug Repurposing. *Curr. Bioinformatics* 15 (9), 1075–1089. doi:10.2174/1574893615666200203104214

Zou, Q. (2019). Latest Machine Learning Techniques for Biomedicine and Bioinformatics. *Curr. Bioinformatics* 14 (3), 176–177. doi:10.2174/1574893614031902201128555

Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Sun, Z.-J., Dao, F.-Y., Yu, X.-L., et al. (2021). Identification of Cyclin Protein Using Gradient Boost Decision Tree Algorithm. *Comput. Struct. Biotechnol. J.* 19, 4123–4131. doi:10.1016/j.csbj.2021.07.013

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Guo, Ju, Chen and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 06 December 2021
doi: 10.3389/fcell.2021.800756

Exploration of Lung Cancer-Related Genetic Factors *via* Mendelian Randomization Method Based on Genomic and Transcriptomic Summarized Data

Nitao Cheng[1], *Xinran Cui*[2], *Chen Chen*[3], *Changsheng Li*[1] *and Jingyu Huang*[1]*

[1]Department of Thoracic Surgery, Zhongnan Hospital, Wuhan University, Wuhan, China, [2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [3]Department of Biological Repositories, Zhongnan Hospital, Wuhan University, Wuhan, China

Lung carcinoma is one of the most deadly malignant tumors in mankind. With the rising incidence of lung cancer, searching for the high effective cures become more and more imperative. There has been sufficient research evidence that living habits and situations such as smoking and air pollution are associated with an increased risk of lung cancer. Simultaneously, the influence of individual genetic susceptibility on lung carcinoma morbidity has been confirmed, and a growing body of evidence has been accumulated on the relationship between various risk factors and the risk of different pathological types of lung cancer. Additionally, the analyses from many large-scale cancer registries have shown a degree of familial aggregation of lung cancer. To explore lung cancer-related genetic factors, Genome-Wide Association Studies (GWAS) have been used to identify several lung cancer susceptibility sites and have been widely validated. However, the biological mechanism behind the impact of these site mutations on lung cancer remains unclear. Therefore, this study applied the Summary data-based Mendelian Randomization (SMR) model through the integration of two GWAS datasets and four expression Quantitative Trait Loci (eQTL) datasets to identify susceptibility genes. Using this strategy, we found ten of Single Nucleotide Polymorphisms (SNPs) sites that affect the occurrence and development of lung tumors by regulating the expression of seven genes. Further analysis of the signaling pathway about these genes not only provides important clues to explain the pathogenesis of lung cancer but also has critical significance for the diagnosis and treatment of lung cancer.

Keywords: lung cancer, susceptibility genes, GWAS, eQTL, SMR

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Yuansong Zhao,
University of Texas Health Science Center at Houston, United States
Weiyue Ding,
Harvard Medical School, United States

***Correspondence:**
Jingyu Huang
huangjingyu1009@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *24 October 2021*
Accepted: *22 November 2021*
Published: *06 December 2021*

Citation:
Cheng N, Cui X, Chen C, Li C and Huang J (2021) Exploration of Lung Cancer-Related Genetic Factors via Mendelian Randomization Method Based on Genomic and Transcriptomic Summarized Data. Front. Cell Dev. Biol. 9:800756. doi: 10.3389/fcell.2021.800756

INTRODUCTION

Pulmonary carcinoma is the leading cause of cancer deaths in the world. According to the Global Cancer Statistics report, there were about 2 million newly detected lung cancer cases worldwide in 2018, accounting for 11.6% of all tumors (Bray et al., 2018). Deaths from lung cancer worldwide were roughly 1.7 million, accounting for 18.4% of tumor-related deaths. Lung cancer is a malignant disease resulting from malignant changes of bronchial and alveolar epithelial cells. Lung cancer is divided into two

categories: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Of all diagnosed cases of lung cancer, NSCLC accounts for about 85% (Oser et al., 2015). Based on the characteristics of lung carcinoma, non-small cell lung cancer is also split into three: pulmonary adenocarcinoma, Lung squamous cell carcinoma, and large cell lung cancer. Although SCLC is a small proportion, it grows and spreads faster than NSCLC and tends to metastasize early in the disease (Zakaria et al., 2015). Bearing these facts in mind, the therapy of lung cancer faces a serious challenge.

It is immediately unclear how pulmonary carcinoma develops. As shown in many kinds of research, the occurrence of lung cancer can be driven by multiple factors including genetic and environmental factors. In particular, SCLC is generally related to lifestyle and environment (Rudina et al., 2021). Besides external environmental factors, genetic factors also play an important role in the occurrence of lung cancer. For example, the incidence of lung adenocarcinoma represents more than 55% of NSCLC, and it is more appropriate to many women and non-smokers (Barr Kumarakulasinghe et al., 2015). These people tend to carry gene mutations. The common mutated genes contain EGFR, ALK, ROS1 gene, and so on (Lamberti et al., 2020). In a word, exploring genetic factors has critical implications for understanding the pathogenic mechanism of pulmonary cancer.

In recent decades, many genetic studies have been utilized to determine the genetic factors of lung cancer. Genome-wide association studies (GWAS) as a hot research tool at present have been applied in lung cancer, and have successfully identified many genetic susceptibility sites (Landi et al., 2009). GWAS perform genotype on several hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) according to the comparison of a large number of lung cancer cases and healthy controls, and then bioinformatics techniques and statistical methods were employed to compare the frequency diversities of each SNP between the two groups (Ko and Urban, 2013). In the results of GWAS analysis, most important SNPs are located in non-coding regions, thus it is difficult to directly explore the regulatory mechanism of these loci (Rojano et al., 2016). Therefore, although we have identified many susceptibility sites associated with lung tumors, it remains unknown about the role of lung cancer pathogenesis-related genes or DNA elements.

Accordingly, this research adopted the method of multi-omics analysis to consolidate various sources of data in an attempt to identify genes related to lung cancer incidence. With the development of high-throughput omics platforms, multi-omics analysis is one of the most important and widely used in biomedical research (Tanaka et al., 2020; Tarwadi et al., 2020; Zhao et al., 2021; Zhao, 2021). As these are kind of hard to describe diseases with complex etiologies by a single omics analysis, the multi-omics research method is developed to provide more evidence for the disease pathogenesis and dig out candidate key factors in depth. Data from different omics sources such as genomics, transcriptomics, and proteomics can be integrated through machine learning or statistical methods for purpose of underlying biological mechanisms (Zhao et al., 2020; Reel, 2021). In this study, Summary data-based Mendelian Randomization (SMR) was used as a statistical method for multi-omics analysis.

SMR was conducted to investigate the pathogenic genes of lung cancer. SMR analysis is an improved model based on Mendelian randomization study that uses genetic variation as a tool to predict the effects of simulated exposure factor on disease using gene-disease causal inference (Zhu et al., 2016). Compared with the general instrumental variable models, the MR model can significantly reduce the estimation bias caused by measuring error. In the MR model, genetic variants as an instrumental variable have to satisfy three core assumptions. The three core assumptions consist of the relevance assumption, the independence assumption, and the exclusion restriction assumption (Hartwig and Davies, 2016). The relevance refers to a robust correlation between genetic variants and exposure factors. The independence assumption represents the independence of genetic variation and confounding factors affecting the expose-outcome relationship (Zhao et al., 2019). The exclusion restriction states that genetic variation can affect outcomes only through exposure factors and not through other pathways. The violation of any of the core assumptions may distort the statistical inference results of MR analysis, leading to incorrect inferences of causality.

However, the existence of linkage disequilibrium (LD) or pleiotropy of instrumental variables can be contrary to the exclusion restriction. Once two genes are not completely independent, they will show some degree of linkage, which is called linkage imbalance. Pleiotropy is the phenomenon that one gene can affect more than one trait. Due to the coordinate positions in the genome between SNPs and the complicated biological effects between genetic variants and traits, LD and pleiotropy are often unavoidable (Wang, 2021). Therefore, to soften the influence of the exclusivity hypothesis, a variety of modified MR methods, such as Mendelian Randomization-Egger regression (MR-Egger), Inverse Variance Weighted (IVW), Mode-Based Estimate (MBE) method, have been proposed in recent years. These methods only consider the analytical error because of pleiotropy. SMR method also takes into account the LD between instrumental variables to decrease the false positive rate. In addition, there are fewer requirements for the summary data of GWAS and eQTL, and the sample size in SMR analysis. As a result, we adopted this approach because of its wide applicability and high accuracy.

In this research, we integrated lung cancer GWAS data and eQTL data to recognize functional genes and regulatory elements in pulmonary carcinoma. Statistical analysis of the relationship between a single SNP and gene expression is referred to as eQTL analysis (Shabalin, 2012). Namely, the expression of a gene is affected by a single SNP, this variable site is considered to be an eQTL locus. Since the subjects of both GWAS and eQTL were SNPs, SMR analysis of overlapping mutants presented in GWAS and eQTL data could reveal lung cancer related-genes. The discovery of these genes could effectively support the design of targeted drugs and precise treatment of lung cancer.

MATERIALS AND METHODS

Acquisition of GWAS Summary Data

We downloaded two available GWAS summary data of lung cancer from the GWAS Catalog website. One of the GWAS

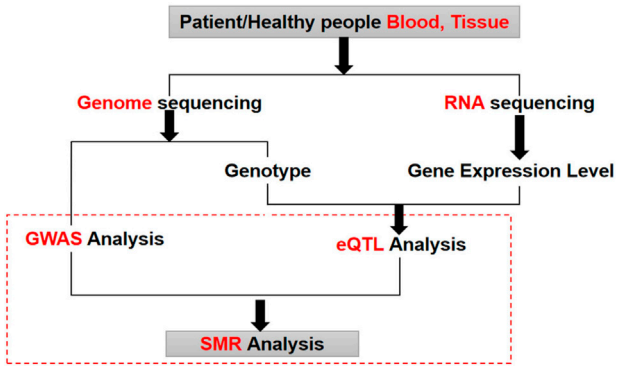


FIGURE 1 | The process of GWAS and eQTL conjoint analysis. GWAS analysis and eQTL analysis are two different types of omics analysis. Using SMR method can discover the biological mechanism behind lung cancer by combining these two different omics data.

datasets was obtained from the meta-analysis of the United Kingdom Biobank (UKB) and the Kaiser Permanente Genetic Epidemiology Research on Adult Health and Aging (GERA) (Rashkin et al., 2020). The UKB study was composed of 502,611 British people aged 40 to 69 at the time of recruitment. Study participants provided biological samples as well as detailed information on lifestyle and health-related factors. GERA participants were members of the Adult Kaiser Permanente Northern California (KPNC) Health Plan. The subjects of this study included 102,979 participants selected for their genotyping. Cancer cases in the UKB were identified on the basis of information provided by various national cancer registries that gathered data from hospitals, nursing homes, and so on. GERA cancer cases group were classified by using diagnoses from the KPNC cancer registry. The control group for the UKB and GERA studies was limited to individuals in the study who did not have any history of lung cancer in the relevant registries.

The other GWAS summary data was collected from the analysis of the BioBank Japan Project (BBJ) data (Sakaue, 2021). BBJ is a prospective biobank that collaboratively collects the DNA and serum samples of 200,000 Japanese participants from 12 medical institutions in Japan. All study participants were diagnosed with one or more of 47 target diseases by doctors. In this study, 178,726 participants of Japanese ancestry were selected from all participants. The researchers used a sample of the cohort without a given lung cancer diagnosis as a control group.

Collection of eQTL Summary Data

Four eQTL data containing three blood eQTL data and one lung eQTL data were applied in our research. Two of eQTL data were obtained by sequencing whole peripheral blood mRNA of 2,116 healthy adults from four Dutch cohorts (Zhernakova et al., 2017). The difference between the two datasets was that one described 23,060 eQTL effects of gene level, the other included 21,888 eQTL effects of exon level. The remaining two of eQTL data were downloaded from the GTEx website. Since eQTL varies in different tissues or cells, we selected one lung eQTL dataset and one blood eQTL dataset to perform the analysis. Lung and blood eQTL datasets stored eQTL information about 25,245 genes and 20,049 genes, respectively.

SMR Analysis Through Lung Cancer GWAS and eQTL Summary Data

The research adopted the SMR analysis method to perform the conjoint analysis of eQTL and GWAS (**Figure 1**). Prior to compute SMR, we first need to find overlapped SNPs of GWAS and eQTL and generate a new dataset containing all overlapping SNPs information. Then the GWAS and eQTL data of new datasets were standardized by Z Score. We can calculate the Z value based on the Beta (OR) value and one of the *p* values or the SE value. The core algorithms of eQTL and GWAS are based on linear regression. Therefore, the Beta/OR value of GWAS and eQTL represents the value of the regression coefficient. SE value refers to the standard error of regression coefficient. The *p* value means the value of whether the association is significant. We listed two calculation formulas of the Z value below (**Eq. 1**; **Eq. 2**).

The Z values of GWAS and eQTL were used to calculate the T_{SMR} values according to **Eq. 3** based on the Delta algorithm (**Figure 2**). The T_{SMR} values were output as P_{SMR} values by chi-square test. As False Discovery Rate (FDR) method can control false positive events by correcting the *p* values, we used FDR to adjust the P_{SMR} values, resulting in obtaining the P_{adj} values. The P_{adj} value of 0.05 was taken as the threshold of statistical significance, which means that the genes corresponding to P_{adj} less than 0.05 could be regarded as lung cancer-related pathogenic genes found in this study.

$$c = \mathrm{qnorm}(1 - p \div 2);\ \mathrm{Beta} > 0,\ c = Z\ \mathrm{Score},\ \mathrm{Beta} < 0;\ c = -Z\ \mathrm{Score} \quad (1)$$

$$Z\ \mathrm{Score} = \mathrm{Beta} \div \mathrm{SE} \quad (2)$$

$$T_{SMR} \approx \frac{Z_{GWAS}^2 \times Z_{eQTL}^2}{Z_{GWAS}^2 + Z_{eQTL}^2} \quad (3)$$

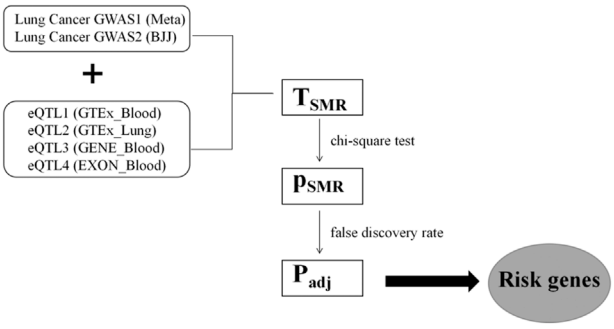


FIGURE 2 | The process of GWAS and eQTL conjoint analysis. The two GWAS datasets of lung cancer were cross-analyzed with four groups of eQTL datasets. Significance of gene expression was found after T_{SMR} calculation, chi-square test, and FDR correction.

TABLE 1 | The new dataset list for SMR analysis. Eight SMR datasets were created for this analysis.

GWAS	eQTL	SMR	GWAS	eQTL	SMR
Meta	GTEx_Blood	Meta_GTEx_Blood	BBJ	GTEx_Blood	BBJ_GTEx_Blood
	GTEx_Lung	Meta_GTEx_Lung		GTEx_Lung	BBJ_GTEx_Lung
	GENE_Blood	Meta_GENE_Blood		GENE_Blood	BBJ_GENE_Blood
	EXON_Blood	Meta_EXON_Blood		EXON_Blood	BBJ_EXON_Blood

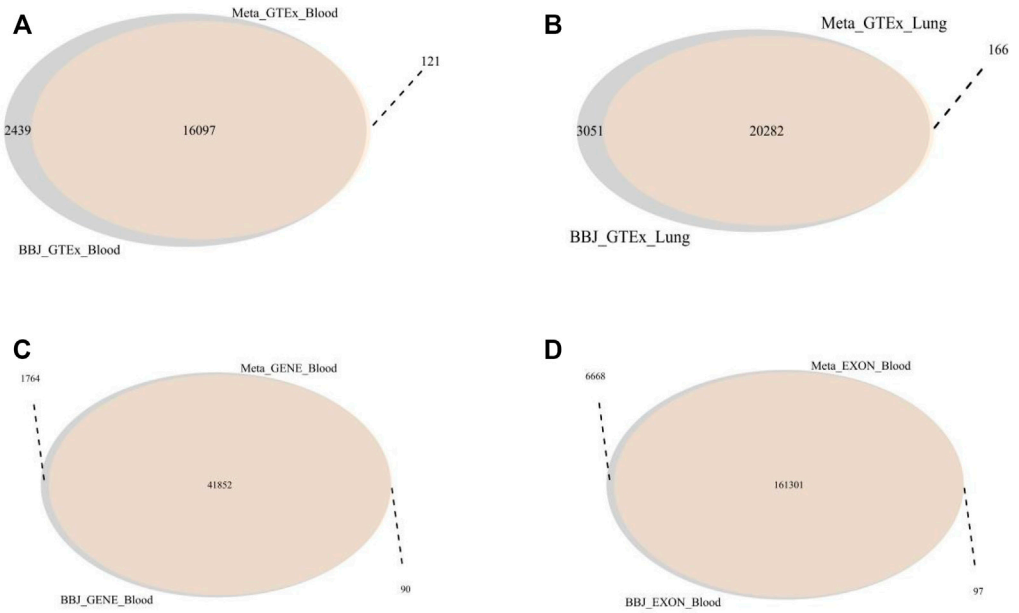


FIGURE 3 | The overlapping genes between two different SMR datasets. **(A)** The overlap genes of Meta_GTEx_Blood and BBJ_GTEx_Blood. 16,097 genes were overlapped. **(B)** The overlap genes of the Meta_GTEx_Lung and BBJ_GTEx_Lung. There were 20,282 overlapped genes. **(C)** The overlap genes of Meta_GENE_Blood and BBJ_GENE_Blood. These two datasets have 41,852 identical genes. **(D)** the overlap genes of Meta_EXON_Blood and BBJ_EXON_Blood. These two datasets have 161,301 identical genes.

RESULTS

Discovery of the Overlapping SNPs Between GWAS and eQTL

In this SMR analysis, the first step was to find the same SNPs between one of the GWAS datasets and one of the eQTL datasets and generate a new SMR dataset containing the details of all the overlapping SNPs (**Table 1**). We performed four SMR analyses by using GWAS summary data from a large meta-analysis to examine the association between lung cancer and gene expression probes from four eQTL data. The number of SNPs found in Meta_Blood_Exon data was higher than the other three data, with nearly a hundred thousand SNPs and more than twenty thousand genes found (**Figure 3**). The remaining three SMR datasets contained 10,000–20,000 SNPs and gene overlaps. In addition, GWAS data collected from the Japanese populations was also conducted four SMR analyses with different eQTL data. The overlap analysis results show that both BBJ_GTEx_Blood and BBJ_GTEx_Lung contained about 20,000 SNPs and genes. BBJ_GENE_Blood and BBJ_EXON_Blood included over 20,000 genes and thirty thousand SNPs and a hundred thousand SNPs, respectively.

Identification of Genes Associated With Lung Carcinoma

The Z values of GWAS and eQTL were calculated by the T_{SMR} formula. After the Chi-squared test, T_{SMR} values were transferred to P_{SMR} (**Figure 4**). Then we performed FDR on P_{SMR} to obtain P_{adj} values. Finally, we filtered P_{adj} values not less than 5%. As a result, two Padj values corresponding to different SNP susceptibility loci (rs8042849 and rs931794) were screened out (**Table 2**). These two SNP loci might regulate gene expression and hence cause the occurrence and development of cancer. Both loci were relevant to the expression of PSMA4 gene.

The analysis of four SMR datasets from Asian populations identified a total of 10 SNP significant sites, most of which were repeatedly identified in these SMR analyses. The expression of seven genes was influenced by these 10 SNPs. The seven genes were comprised of HLA-DQB1, CLPTM1L, RAD52, IREB2, PSMA4, HLA-DRB9, and HLA-DQB2. According to the

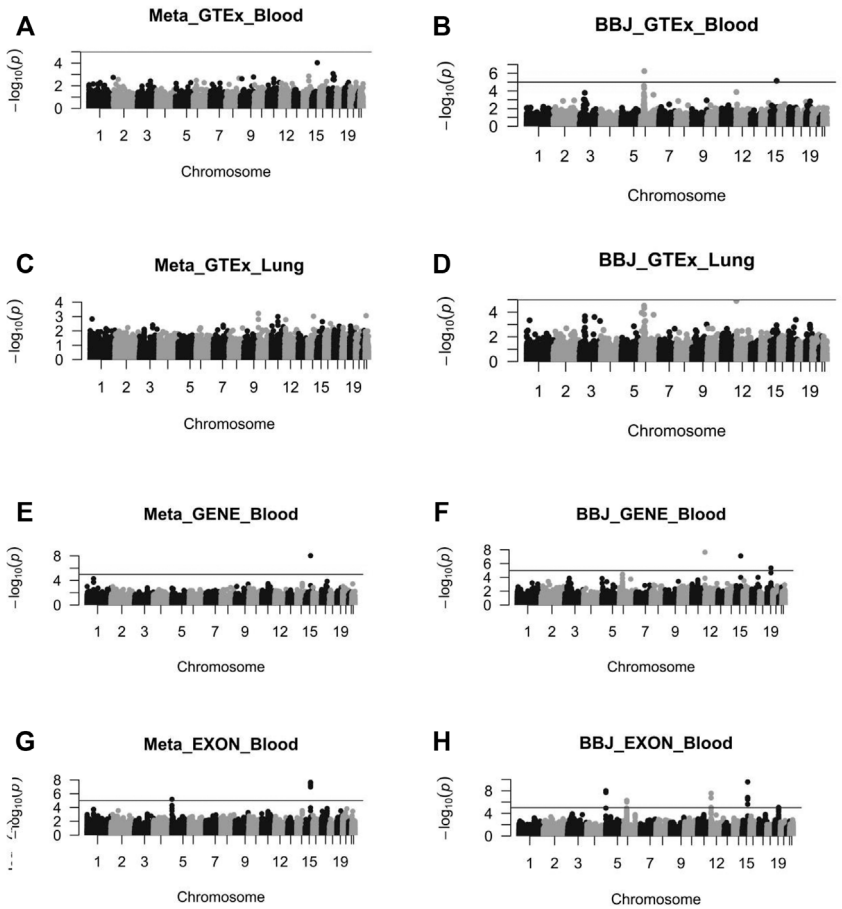


FIGURE 4 | The Manhattan diagram of these eight SMR analysis results. **(A)** the SMR analysis results of Meta_GTEx_Blood. **(B)** the SMR analysis results of BBJ_GTEx_Blood. **(C)** the SMR analysis results of Meta_GTEx_Lung. **(D)** the SMR analysis results of BBJ_GTEx_Lung. **(E)** the SMR analysis results of Meta_GENE_Blood. **(F)** the SMR analysis results of BBJ_GENE_Blood. **(G)** the SMR analysis results of Meta_EXON_Blood. **(H)** the SMR analysis results of BBJ_EXON_Blood.

TABLE 2 | The list of discovered lung cancer causative genes by the SMR analysis. This list showed the significant SNPs and genes found in different datasets, and the corresponding value of P_{SMR} and P_{adj} calculated.

SMR datasets	Significant SNPs	P_{SMR} value	P_{adj} value	Gene name
Meta_GENE_Blood	rs8042849	9.17×10^{-9}	0.0004	PSMA4
Meta_EXON_Blood	rs931794	1.06×10^{-7}	0.0029	PSMA4
	rs8042849	2.44×10^{-8}	0.0013	PSMA4
BBJ_GTEx_Blood	rs9274510	5.70×10^{-7}	0.0106	HLA-DQB1
BBJ_GENE_Blood	rs147560086	2.26×10^{-8}	0.0010	RAD52
	rs8042849	7.81×10^{-8}	0.0017	PSMA4
BBJ_EXON_Blood	rs31487	1.96×10^{-8}	0.0011	CLPTM1L
	rs147560086	2.77×10^{-8}	0.0012	RAD52
	rs12592111	2.40×10^{-6}	0.0288	IREB2
	rs8042849	1.66×10^{-7}	0.0040	PSMA4
	rs12822733	1.74×10^{-7}	0.0037	RAD52
	rs931794	2.70×10^{-10}	4.53×10^{-5}	PSMA4
	rs402710	9.79×10^{-9}	0.0008	CLPTM1L
	rs9274564	5.15×10^{-7}	0.0072	HLA-DQB1
	rs9274564	5.14×10^{-7}	0.0079	HLA-DRB9
	rs28755305	9.50×10^{-7}	0.0123	HLA-DQB2

results, PSMA4 was recognized again. Furthermore, no significant susceptibility sites were found in the association analysis of GWAS data and lung eQTL data.

Gene Function Analysis

To further explain the potential regulatory mechanism of gene expression on lung cancer susceptibility, we conducted the analysis of gene function and involved biological pathways on these seven genes on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. According to the analysis results, PSMA4 is an essential subunit of the proteasome protein. The proteasome is a large protein complex and is the key to regulating cell function cellular functions, such as cell differentiation, cell cycle (Voutsadakis, 2017). Many studies have shown that the proteasome plays an important role in the development and progression of various tumors, including lung cancer. In addition, three of the seven genes were the members of the Major Histocompatibility Complex (MHC) gene family. Human Leukocyte Antigen (HLA) stands for human MHC, The HLA gene family is divided into three subgroups, which encode three types of molecules: MHC class I molecules, MHC class II molecules, and MHC class III molecules. These three genes were of the MHC class II. When the external invaders are engulfed and processed by lysosome to be fragmented, the MHC class II molecules bind to these fragments and present them on the cell surface for T cells to recognize. The study indicates that HLA class II genes on tumor cells impact tumor immunogenicity (Chen et al., 2019).

Moreover, the CLPTM1L gene can encode cleft lip and palate transmembrane protein 1-Like protein. Recent research shows that the SNPs of CLPTM1L were associated with many human malignancies, especially in some highly aggressive and metastatic cancers (Shete et al., 2020). RAD52 function on the recombination repair of double strand breaks (DSB) is important for maintaining chromosome integrity. Based on the cause of cell death by DSB unrepair, cancer cells have a higher DSB burden, and hence the relationship between DSB and RAD52 can be exploited for cancer treatment (Trenner and Sartori, 2019). IREB2 gene can encode an iron-responsive element-binding protein that is an RNA-binding protein and can bind to iron response elements (IRES) for the regulation of the translation and stability of mRNAs in cells. Variants in the gene have been linked to lung cancer and chronic obstructive pulmonary disease (COPD). These results further proved this inference of our analysis. Accordingly, the analysis of the relationship between these genes and lung cancer can provide an important theoretical basis for revealing the genetic mechanism behind lung cancer.

DISCUSSION

We applied multi-omics analysis to integrate two of the GWAS datasets and four of eQTL datasets based on the SMR method, resulting in the functionally important genes related-lung cancer being found. SMR research can facilitate more accurate and effective identification of the potential causal relationship between exposure and outcome, for disease risk factors had a further understanding. Once risk factors have been successfully discovered, they can be prevented or targeted for cancer treatment. By analyzing the causal genes associated with lung cancer, we recognized seven genes associated with lung cancer, suggesting a possible mechanism for the development of lung cancer by altering the expression of risk genes.

With the analysis of these SMR datasets, the PSMA4 gene was repeatedly identified. It has been reported that down-regulation of proteasome subunit encoding gene PSMA4 reduces proteasome activity (Wang et al., 2007). The proteasome is a large protein complex that can degrade various cellular proteins rapidly and timely. The proteasome is responsible for regulating many cellular processes, including transcription, cell cycle progression, and apoptosis (Cardoso et al., 2004). Proteasome dysfunction leads to many diseases including cancer. Drugs that inhibit proteasome activity directly affect the susceptibility of lung cancer by regulating cell proliferation and apoptosis. Additionally, PSMA4 is involved in cancer cell proliferation, also its polymorphisms have been shown to increase the risk of lung cancer in the Chinese Han population (Wang et al., 2015). Based on the above information, we predicted that PSMA4 may govern cell proliferation and apoptosis, and hence can promote the proliferation of lung cancer cells.

Additionally, HLA-DQB1, HLA-DRB9, and HLA-DQB2 are all homologs of HLA class II molecules, which are essential for immune response. HLA plays a key role in the interaction between tumor cells and the human immune system. HLA II molecules are involved in anti-tumor immunity and exogenous antigen presentation by $CD4^+$ T cells. There is evidence that HLA II on tumor cells affects tumor migration and invasion, cancer progression, immune response and prognosis of a variety of malignancies *in vitro* and *in vivo*. For example, some studies have shown that tumor HLA-DQB1 level is associated with the recurrence free survival (RFS) of early Lung Adenocarcinoma, and the mechanism may be related to anti-tumor immunity (Zhang et al., 2019). Currently, there is no definitive study on the relationship between HLA-DRB9, HLA-DQB2 and lung cancer, and further verification will be needed.

CLPTM1L is a mitochondrial protein and is thought to be involved in lung cancer. Some studies have found that the expression of CLPTM1L is significantly increased in lung cancer tissues compared with normal tissues by immunohistochemistry, especially in lung adenocarcinoma (Ni et al., 2012). Also, the researchers found that increased copy number in the CLPTM1L region is the most common genetic event in the early stages of NSCLC (James et al., 2012). Some GWAS data illustrate that the region of CLPTM1L gene is associated with lung cancer. Therefore, the gene expression of CLPTM1L is closely related to lung tumor susceptibility.

Moreover, according to the GWAS results, the SNPs of RAD52 gene were associated with increased risk of lung cancer, particularly the development of lung squamous cell carcinoma (Shi et al., 2012). As the development of lung squamous cell carcinoma is closely associated with smoking, DNA repair is increased in the lung tissue of smokers (Lieberman et al., 2017). RAD52 encoded gene product can bind to the single stranded

DNA terminal and can be involved in homologous recombination and DNA repair in mammals. To sum up, the relationship between RAD52 gene expression and the occurrence of lung cancer may be due to its DNA repair function in lung tissue.

The iron-binding protein encoded by IREB2 is involved in maintaining iron homeostasis in human cells. IREB2 is located in the susceptibility locus of lung cancer. A previous study showed that iron levels in the lungs increase with age, with higher concentrations in the lungs of smokers. Therefore, abnormal IREB2 expression or function may lead to iron metabolism disorder. Iron load leads to cell proliferation in cancer cell lines, and cancer cell lines have a poor ability to regulate IREB2 expression (Khiroya et al., 2017). In addition, GWAS results illustrate that IREB2 has been shown to be associated COPD and the expression of its protein product IRP2 is altered in lung cancer patients carrying lung cancer (Fehringer et al., 2012). The role of IRP2 is to control iron levels in cells by regulating the input, output and storage of various iron proteins. Therefore, the misexpression of IREB2 gene may alter intracellular iron levels and hence lead to the occurrence of malignant tumors. If IRP2 gene can be targeted therapeutically, it may provide a new approach for lung cancer treatment.

In summary, our study integrated genomic and transcriptome information and found seven genes associated with lung cancer by the SMR method, revealing possible regulatory mechanisms that alter the expression of these genes and further pathogenicity. This provides a research direction for understanding the precise treatment and targeted drug design of lung cancer. Although the SMR method can avoid the interference of confounding effect and reverse causal association in the study, and can be not limited by the sample size, thus producing more reliable results, this method can only preliminarily judge that the gene expression may affect the development of lung cancer, and cannot show how gene expression regulates the onset and progression of this disease. It takes experiments and other analytical methods to confirm a causal link between genes and disease. In the following study, we will continue to use SMR analysis method to integrate GWAS and other types of QTL data to mine more genes affecting lung cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

NC wrote this manuscript. XC did experiments. CC collected data. CL helped revise this manuscript. JH designed the whole work.

REFERENCES

Barr Kumarakulasinghe, N., Zanwijk, N. v., and Soo, R. A. (2015). Molecular Targeted Therapy in the Treatment of Advanced Stage Non-small Cell Lung Cancer (NSCLC). *Respirology* 20 (3), 370–378. doi:10.1111/resp.12490

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a Cancer J. clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Cardoso, F., Ross, J. S., Piccart, M. J., Sotiriou, C., and Durbecq, V. (2004). Targeting the Ubiquitin-Proteasome Pathway in Breast Cancer. *Clin. Breast Cancer* 5 (2), 148–157. doi:10.3816/cbc.2004.n.020

Chen, B., Khodadoust, M. S., Olsson, N., Wagar, L. E., Fast, E., Liu, C. L., et al. (2019). Predicting HLA Class II Antigen Presentation through Integrated Deep Learning. *Nat. Biotechnol.* 37 (11), 1332–1343. doi:10.1038/s41587-019-0280-2

Fehringer, G., Liu, G., Pintilie, M., Sykes, J., Cheng, D., Liu, N., et al. (2012). Association of the 15q25 and 5p15 Lung Cancer Susceptibility Regions with Gene Expression in Lung Tumor Tissue. *Cancer Epidemiol. Biomarkers Prev.* 21 (7), 1097–1104. doi:10.1158/1055-9965.epi-11-1123-t

Hartwig, F. P., and Davies, N. M. (2016). Why Internal Weights Should Be Avoided (Not Only) in MR-Egger Regression. *Int. J. Epidemiol.* 45 (5), 1676–1678. doi:10.1093/ije/dyw240

James, M. A., Wen, W., Wang, Y., Byers, L. A., Heymach, J. V., Coombes, K. R., et al. (2012). Functional Characterization of CLPTM1L as a Lung Cancer Risk Candidate Gene in the 5p15.33 Locus. *PloS one* 7 (6), e36116. doi:10.1371/journal.pone.0036116

Khiroya, H., Moore, J. S., Ahmad, N., Kay, J., Woolnough, K., Langman, G., et al. (2017). IRP2 as a Potential Modulator of Cell Proliferation, Apoptosis and Prognosis in Nonsmall Cell Lung Cancer. *Eur. Respir. J.* 49 (4). doi:10.1183/13993003.00711-2016

Ko, D. C., and Urban, T. J. (2013). Understanding Human Variation in Infectious Disease Susceptibility through Clinical and Cellular GWAS. *Plos Pathog.* 9 (8), e1003424. doi:10.1371/journal.ppat.1003424

Lamberti, G., Andrini, E., Sisi, M., Rizzo, A., Parisi, C., Di Federico, A., et al. (2020). Beyond EGFR, ALK and ROS1: Current Evidence and Future Perspectives on Newly Targetable Oncogenic Drivers in Lung Adenocarcinoma. *Crit. Rev. Oncology/Hematology* 156, 103119. doi:10.1016/j.critrevonc.2020.103119

Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., et al. (2009). A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma. *Am. J. Hum. Genet.* 85 (5), 679–691. doi:10.1016/j.ajhg.2009.09.012

Lieberman, R., Pan, J., Zhang, Q., and You, M. (2017). Rad52 Deficiency Decreases Development of Lung Squamous Cell Carcinomas by Enhancing Immuno-Surveillance. *Oncotarget* 8 (21), 34032–34044. doi:10.18632/oncotarget.16371

Ni, Z., Tao, K., Chen, G., Chen, Q., Tang, J., Luo, X., et al. (2012). CLPTM1L Is Overexpressed in Lung Cancer and Associated with Apoptosis. *PloS one* 7 (12), e52598. doi:10.1371/journal.pone.0052598

Oser, M. G., Niederst, M. J., Sequist, L. V., and Engelman, J. A. (2015). Transformation from Non-small-cell Lung Cancer to Small-Cell Lung Cancer: Molecular Drivers and Cells of Origin. *Lancet Oncol.* 16 (4), e165–e172. doi:10.1016/s1470-2045(14)71180-5

Rashkin, S. R., Graff, R. E., Kachuri, L., Thai, K. K., Alexeeff, S. E., Blatchins, M. A., et al. (2020). Pan-cancer Study Detects Genetic Risk Variants and Shared Genetic Basis in Two Large Cohorts. *Nat. Commun.* 11 (1), 4423–4514. doi:10.1038/s41467-020-18246-6

Reel, P. S. (2021). Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnol. Adv.*. doi:10.1016/j.biotechadv.2021.107739

Rojano, E., Ranea, J. A., and Perkins, J. R. (2016). Characterisation of Non-coding Genetic Variation in Histamine Receptors Using AnNCR-SNP. *Amino acids* 48 (10), 2433–2442. doi:10.1007/s00726-016-2265-5

Rudin, C. M., Brambilla, E., Faivre-Finn, C., and Sage, J. (2021). Small-cell Lung Cancer. *Nat. Rev. Dis. Primers* 7 (1), 3–20. doi:10.1038/s41572-020-00235-0

Sakaue, S., Kanai, M., and Tanigawa, Y. (2021). *A Global Atlas of Genetic Associations of 220 Deep Phenotypes*. medRxiv. doi:10.1101/2020.10.23.20213652

Shabalin, A. A. (2012). Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations. *Bioinformatics* 28 (10), 1353–1358. doi:10.1093/bioinformatics/bts163

Shete, S., Liu, H., Wang, J., Yu, R., Sturgis, E. M., Li, G., et al. (2020). A Genome-wide Association Study Identifies Two Novel Susceptible Regions for Squamous Cell Carcinoma of the Head and Neck. *Cancer Res.* 80 (12), 2451–2460. doi:10.1158/0008-5472.can-19-2360

Shi, J., Chatterjee, N., Rotunno, M., Wang, Y., Pesatori, A. C., Consonni, D., et al. (2012). Inherited Variation at Chromosome 12p13.33, Including RAD52, Influences the Risk of Squamous Cell Lung Carcinoma. *Cancer Discov.* 2 (2), 131–139. doi:10.1158/2159-8290.cd-11-0246

Tanaka, E., Uchida, D., Shiraha, H., Kato, H., Ohyama, A., Iwamuro, M., et al. (2020). Promising Gene Therapy Using an Adenovirus Vector Carrying REIC/Dkk-3 Gene for the Treatment of Biliary Cancer. *Cgt* 20 (1), 64–70. doi:10.2174/1566523220666200309125709

Tarwadi., T., Jazayeri, J. A., Pambudi, S., Arbianto, A. D., Rachmawati, H., Kartasasmita, R. E., et al. (2020). In-silico Molecular Interaction of Short Synthetic Lipopeptide/Importin-Alpha and *In-Vitro* Evaluation of Transgene Expression Mediated by Liposome- Based Gene Carrier. *Cgt* 20, 383–394. doi:10.2174/1566523220666201005104224(

Trenner, A., and Sartori, A. A. (2019). Harnessing DNA Double-Strand Break Repair for Cancer Treatment. *Front. Oncol.* 9, 1388. doi:10.3389/fonc.2019.01388

Voutsadakis, I. A. (2017). Proteasome Expression and Activity in Cancer and Cancer Stem Cells. *Tumour Biol.* 39 (3), 1010428317692248. doi:10.1177/1010428317692248

Wang, T., Chen, T., Thakur, A., Liang, Y., Gao, L., Zhang, S., et al. (2015). Association of PSMA4 Polymorphisms with Lung Cancer Susceptibility and Response to Cisplatin-Based Chemotherapy in a Chinese Han Population. *Clin. Transl Oncol.* 17 (7), 564–569. doi:10.1007/s12094-015-1279-x

Wang, W., Spitz, M. R., Yang, H., Lu, C., Stewart, D. J., and Wu, X. (2007). Genetic Variants in Cell Cycle Control Pathway Confer Susceptibility to Lung Cancer. *Clin. Cancer Res.* 13 (19), 5974–5981. doi:10.1158/1078-0432.ccr-07-0113

Wang, Y. (2021). A Novel Method for Mendelian Randomization Analyses with Pleiotropy and Linkage Disequilibrium in Genetic Variants from Individual Data. *Front. Genet.* 12, 394. doi:10.3389/fgene.2021.634394

Zakaria, N., Yusoff, N. M., Zakaria, Z., Lim, M. N., Baharuddin, P. J., Fakiruddin, K. S., et al. (2015). Human Non-small Cell Lung Cancer Expresses Putative Cancer Stem Cell Markers and Exhibits the Transcriptomic Profile of Multipotent Cells. *BMC cancer* 15 (1), 84–16. doi:10.1186/s12885-015-1086-3

Zhang, L., Li, M., Deng, B., Dai, N., Feng, Y., Shan, J., et al. (2019). HLA-DQB1 Expression on Tumor Cells Is a Novel Favorable Prognostic Factor for Relapse in Early-Stage Lung Adenocarcinoma. *Cmar* Vol. 11, 2605–2616. doi:10.2147/cmar.s197855

Zhao, T., Hu, Y., and Cheng, L. (2021). Deep-DRM: a Computational Method for Identifying Disease-Related Metabolites Based on Graph Deep Learning Approaches. *Brief Bioinform* 22 (4), bbaa212. doi:10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2019). Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. *Front. Genet.* 10, 1021. doi:10.3389/fgene.2019.01021

Zhao, T. (2021). Prediction and Collection of Protein–Metabolite Interactions. *Brief. Bioinform.*.

Zhernakova, D. V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., et al. (2017). Identification of Context-dependent Expression Quantitative Trait Loci in Whole Blood. *Nat. Genet.* 49 (1), 139–145. doi:10.1038/ng.3737

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets. *Nat. Genet.* 48 (5), 481–487. doi:10.1038/ng.3538

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Cheng, Cui, Chen, Li and Huang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 08 December 2021
doi: 10.3389/fcell.2021.790878

An Apoptosis-Related Gene Prognostic Index for Colon Cancer

*Hanmin Tang, Jing Wang, Xuehui Luo, Qi Wang, Jie Chen, Xinyue Zhang, Qiuting Li, Chengyi Gao, Yuesen Li and Suxia Han**

Department of Radiation Oncology, First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

Purpose: To construct an apoptosis-related gene prognostic index (ARGPI) for colon cancer, and clarify the molecular and immune characteristics of the risk subgroup as defined by the prognostic index and the benefits of adjuvant chemotherapy. Integrating the prognostic index and clinicopathological risk factors to better evaluate the prognosis of patients with colon cancer.

Methods: Based on the colon adenocarcinoma data in the TCGA database, 20 apoptosis-related hub genes were screened by weighted gene co-expression network analysis (WGCNA). Five genes constituting the prognosis model were determined by Cox regression and verified by the Gene Expression Omnibus (GEO) dataset. Then the molecular and immune characteristics of risk subgroups defined by the prognostic index and the benefits of adjuvant chemotherapy were analyzed. Finally, nomograms integrating ARGPI and four clinicopathological risk factors were used to evaluate the prognosis of patients with colon cancer.

Results: The ARGPI was constructed based on the FAS, VWA5A, SPTBN2, PCK1, and TIMP1 genes. In the TCGA cohort, patients in the low-risk subgroup had a longer progression-free interval (PFI) than patients in the high-risk subgroup, which coincided with the results of the GEO cohort. The comprehensive results showed that the high-risk score was related to the enrichment of the cell cycle pathway, high mutation rate of TP53 and KRAS, high infiltration of T regulatory cells (Tregs), immunosuppressive state, and less chemotherapeutic benefit. However, low-risk scores are related to drug metabolism-related pathways, low TP53 and KRAS mutation rates, high infiltration of plasma cells, more resting CD4 memory cells and eosinophils, active immune function, and better chemotherapeutic benefits. Receiver operating characteristic curve of two-year progress prediction evaluation showed that the ARGPI had higher prognostic accuracy than TNM staging. Nomograms integrating ARGPI and clinicopathological risk factors can better evaluate the prognosis of patients with colon cancer.

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Tao Huang,
Shanghai Institute of Nutrition and Health (CAS), China
Yingjie Guo,
University of Electronic Science and Technology of China, China

****Correspondence:***
Suxia Han
shan87@mail.xjtu.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *07 October 2021*
Accepted: *11 November 2021*
Published: *08 December 2021*

Citation:
Tang H, Wang J, Luo X, Wang Q, Chen J, Zhang X, Li Q, Gao C, Li Y and Han S (2021) An Apoptosis-Related Gene Prognostic Index for Colon Cancer. Front. Cell Dev. Biol. 9:790878. doi: 10.3389/fcell.2021.790878

Abbreviations: ARGPI, apoptosis related gene prognostic index; COAD, colon adenocarcinoma; DEGs, differentially expressed genes; GEO, gene expression omnibus; GO, Gene Ontology; GSEA, gene set enrichment analysis; GSVA, gene set variation analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; PFI, progression free interval events; ROC, receiver operating characteristic curve; TCGA, the cancer genome atlas; TMB, tumor mutation burden.

Conclusions: The ARGPI is a promising biomarker for determining risk of colon cancer progression, molecular and immune characteristics, and chemotherapeutic benefit. This is a reliable method to predict the prognosis of colon cancer patients. It also can assist doctors in formulating more effective treatment strategies.

Keywords: apoptosis gene, prognostic index, colon cancer, tumor microenvironment, treatment, nomograms

INTRODUCTION

Globally, colorectal cancer is the third most common malignant tumor and second most common cause of death due to malignant tumors (Bray et al., 2018). At present, radical surgery remains the main treatment scheme for non-metastatic colon cancer. However, postoperative recurrence is a major problem. Approximately 80% of recurrence occurs in the first 3 years after treatment, often causing death (Sargent et al., 2005). For colon cancer with distant metastasis, even after receiving palliative treatment, the median survival time of patients is only 10–12 months (Labianca et al., 2010). Therefore, to better tumor control, it is essential to predict tumor recurrence and death due to colon cancer. At present, prognosis evaluation and treatment plan formulation of colon cancer patients largely depend on the TNM staging system (Kehoe and Khatri, 2006). However, due to tumor heterogeneity, even with the same tumor stage, there are still huge differences in disease progression and clinical results. This suggests the limitation of the TNM staging system in predicting the prognosis of patients with colon cancer. Therefore, it is necessary to develop new and reliable biomarkers to accurately predict the prognosis of patients with colon cancer.

Apoptosis is a programmed cell death regulated by genes and plays a very important role in maintaining human homeostasis and normal function (Rich et al., 1999). Apoptosis triggers cell death by activating cysteine protease, resulting in morphological changes in nuclear concentration and membrane vesiculation (Galluzzi et al., 2018). If DNA damaged cells do not receive the signal of apoptosis, then the balance between cell division and cell death may lead to cancer (Fan and Guo 2001). Studies have shown that inhibition of apoptosis is involved in the initial stage of carcinogenesis. For example, the combination of Fas and FasL can start the transduction of apoptosis signals and cause apoptosis. In the process of tumor development, Fas (CD95) is often downregulated, and tumor cells become anti-apoptotic (Fouqué et al., 2014). In addition, an important mechanism of chemoresistance is to inhibit tumor cell apoptosis. For example, Bcl-2 anti-apoptotic molecules are frequently upregulated in acquired drug-resistant cancer cells to block drug-induced apoptosis (Maji et al., 2017). However, there are few studies on the relationship between apoptosis and prognosis of colon cancer using high-throughput expression profiling.

In this study, we used weighted gene co-expression network analysis (WGCNA) and Cox regression to construct the apoptosis-related gene prognosis index (ARGPI), which improves risk stratification of colon cancer prognosis. The nomogram, based on progression-free interval (PFI), integrates the ARGPI and four clinicopathological risk factors, which are used to evaluate the clinical prognosis.

MATERIALS AND METHODS

Patients and Datasets

The colon adenocarcinoma-associated mRNA sequences and clinical data used in this study were from TCGA[1] and GEO (GSE39582) databases. Patients who met the following selection criteria were included: (A) the primary site of the tumor must be the colon, not colon metastasis; (B) The pathological type of colon cancer must be adenocarcinoma; (C) Available gene expression data; (D) Available survival information. The clinical data details of two train datasets used in this study are shown in **Supplementary Table S1**. The list of apoptosis-related genes was extracted from molecular signatures database[2] (Liberzon et al., 2015). The search strategy includes the following keywords: "apoptosis" and "*Homo sapiens*." Finally, 2,737 apoptosis-related genes were listed. PFI (progression-free interval), the most reliable prognostic index recommended in the TCGA database, was selected as the end point of disease progression. PFI is the period from the date of diagnosis until the date of the first occurrence of a new tumor event, which includes progression of the disease, locoregional recurrence, distant metastasis, new primary tumor, or death with tumor (Liu J. et al., 2018). **Supplementary Figure S1** showed the flowchart for the present research.

Identification of Apoptosis Related Hub Gene

The differentially expressed genes (DEGs) between colon cancer samples (379 cases) and normal samples (39 cases) in the TCGA dataset were identified using the limma package in R (p-value < 0.05, $|\log2FC| > 1$) (Ritchie et al., 2015). The apoptosis-related genes obtained from the molecular signatures database were intersected with the DEGs of colon cancer to obtain apoptosis-related DEGs. The DEGs related to apoptosis were analyzed by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) using Metascape[3] to determine the potential molecular mechanism of gene expression profile (Zhou et al., 2019).

Weighted gene co-expression network analysis (WGCNA) is a bioinformatics analysis method used to describe the gene association patterns between different samples (Langfelder and Horvath, 2008). Clustering of genes with similar expression patterns promotes web-based hub gene screening, which can

[1]https://portal.gdc.cancer.gov/

[2]http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp

[3]https://metascape.org

be used to identify candidate biomarkers (Tang J. et al., 2018; Chen et al., 2021). The DEGs related to apoptosis were analyzed by the WGCNA package in R. First, the Pearson correlation was calculated from the expression data, and the relationship matrix was established. Then, the correlation matrix was transformed into an adjacency matrix, and the power exponential weighting was introduced to construct the scale-free network. Next, based on the adjacency matrix, the Tom matrix was established to calculate the Tom difference degree (1-TOM) between genes, and then the gene module was established. Finally, the results of the module were visualized, and the target module was selected for analysis. In this study, 127 genes of the most significantly related module (Turquoise module) were selected for further exploration. Twenty apoptosis-related hub genes significantly related to progression were screened by the R package ("survival" and "surviviner") for further analysis.

Construction and Validation of Apoptosis-Related Gene Prognostic Index

From 20 apoptosis-related hub genes, the genes independently related to the prognosis of colon cancer were screened by multivariate Cox regression analysis. The prognostic index of apoptosis related genes was determined using the following formula: ARGPI = Expression gene 1 * Coefficient + Expression gene 2 * Coefficient + + Expression gene * coefficient. The value of the ARGPI was defined as the risk score of each patient. According to the median risk score, the patients in the two databases were divided into the high-risk group and low-risk group for follow-up evaluation. The prognostic evaluation ability of the ARGPI was evaluated by plotting the K-M survival curve of TCGA and GEO cohort ($p < 0.05$, log rank test).

Analysis of Molecular Characteristics of Different ARGPI Subgroups

Gene set enrichment analysis (GSEA) is a method to evaluate microarray data at the level of gene set (Subramanian et al., 2005). In the GSEA software, the gene set (c2. cp.kegg.v7.4. symbols.gmt) was used as the reference gene set to explore the important signal pathways enriched in high-risk subgroups and low-risk subgroups ($p < 0.05$ and FDR <0.25). In the analysis of gene mutation, information on mutations related to colon cancer samples was obtained from the TCGA database. Somatic variants in tumor patients in the high-risk and low-risk subgroups were comprehensively analyzed using the Maftools package in R (Mayakonda et al., 2018).

Analysis of Immune Characteristics of Different ARGPI Subsets

CIBERSORT[4] is a tool for deconvoluting the expression matrix of immune cell subtypes based on the principle of linear support vector regression. This method utilized expression profile data of colon cancer to estimate the infiltration of 22 types of immune cells in different risk subgroups (Newman et al., 2015). ssGSEA (single-sample gene set enrichment analysis) is used to study the enrichment level of immune-related functions in tissue expression profiles (Barbie et al., 2009). The enrichment levels of 29 immune-related functional features in different risk subgroups were analyzed using the GSVA package in R.

Construction of Nomograms

To improve the prognostic risk stratification of colon cancer and assist in clinical diagnosis and treatment, a nomogram model integrating the ARGPI and clinicopathological features was constructed. It is used as a quantitative tool to predict the prognosis of patients with colon cancer. The effectiveness of the model was evaluated using a calibration curve.

Statistical Analysis

Differences among variables were analyzed using an independent *t*-test and a chi-squared test. Univariate survival was assessed using Kaplan–Meier survival analysis and log rank test. Univariate Cox regression and multivariate Cox regression were used to analyze the effects of the ARGPI and clinical characteristics on prognosis. The data were statistically analyzed by R (version 4.1.1) and SPSS software (version 25.0). Bilateral $p < 0.05$ was considered statistically significant.

RESULTS

The Apoptosis Related Hub Gene was Identified by WGCNA Analysis

A total of 7,859 DEGs were obtained by comparing colon adenocarcinoma samples with normal tissue samples, of which 5,432 genes were upregulated and 2,427 genes were downregulated (**Figure 1A**). Among these DEGs, 655 genes were related to apoptosis. Compared with normal samples, 392 genes were upregulated and 263 genes were downregulated in tumor samples (**Figure 1B**). Functional enrichment analysis identified GO terms and KEGG pathways that were enriched by 655 apoptosis-related DEGs (**Figures 1C,D**).

To identify apoptosis-related hub genes, WGCNA was used for candidate genes (n = 655). As shown in the left figure of **Figure 2A**, the horizontal axis is the soft threshold and the vertical axis is the evaluation parameters of scale-free network. The higher the value of evaluation parameters, the more consistent the network is with the characteristics of scale-free network. In the left graph, the horizontal line indicates that the threshold value is 0.90. As shown in the right figure of **Figure 2A**, the relationship between soft threshold (power) and mean connectivity shows that the optimal soft-thresholding power was seven based on the scale-free network. Finally, six modules are identified by hierarchical clustering and optimal soft threshold capability (**Figure 2B**). As shown in **Figure 2C**, the turquoise module with the highest absolute correlation with colon cancer was selected as the candidate module for further analysis, which contains 127 genes. Significantly enriched GO terms and KEGG pathways of turquoise modular genes are shown in **Supplementary Figure S2**. The expression profiles and clinical prognosis information of

[4]http://cibersort.stanford.edu/

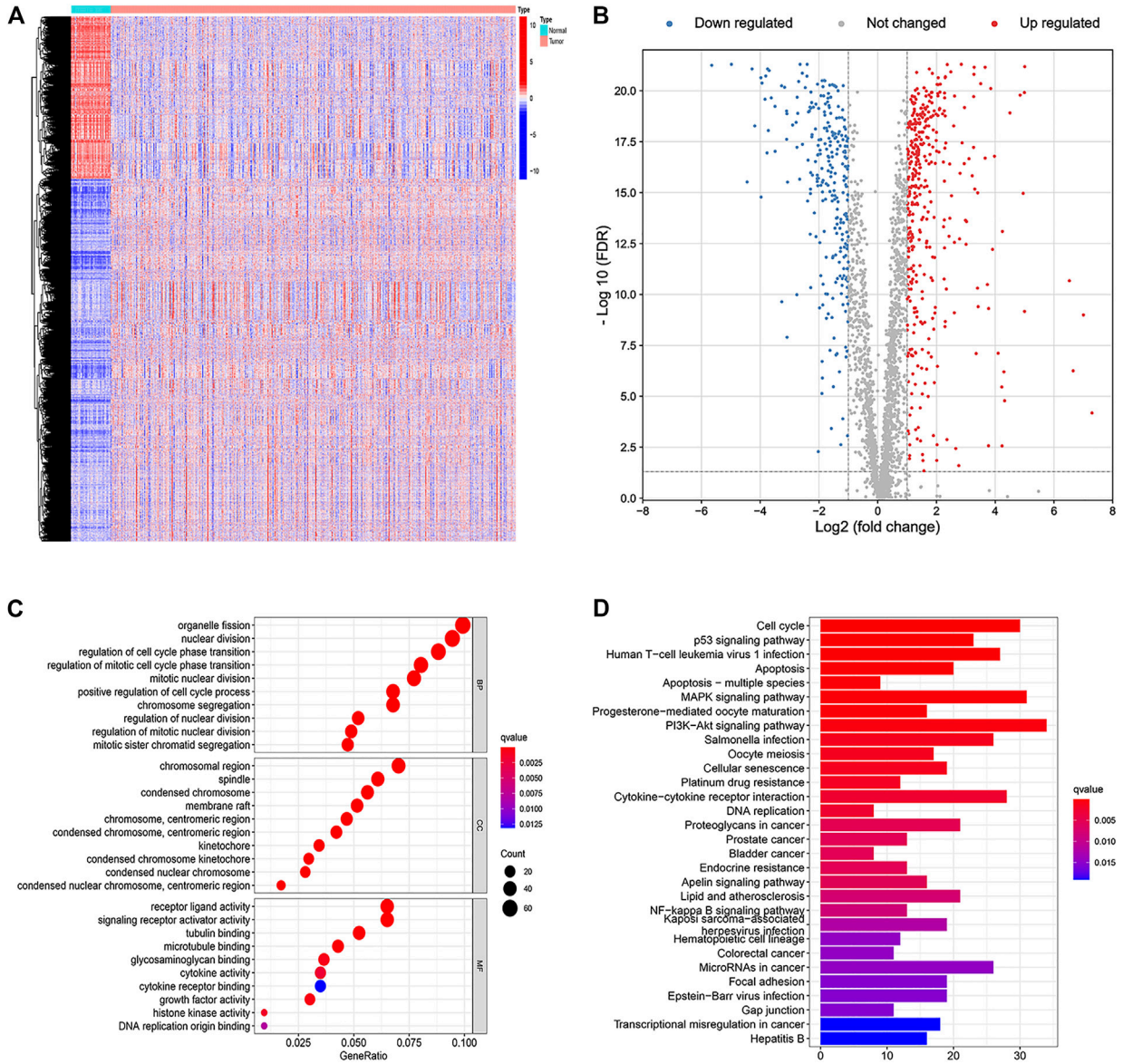


FIGURE 1 | Identification and functional enrichment analyses of apoptosis-related differentially expressed genes. **(A)** Heat map of differentially expressed genes in colon cancer. **(B)** Volcano plot of apoptosis-related DEGs. **(C)** Gene Ontology (GO) enrichment analysis of the apoptosis-related DEGs. **(D)** Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of the apoptosis-related DEGs.

127 genes in the turquoise module were obtained. Through K-M survival analysis, the expression of 20 apoptosis-related hub genes was closely related to the progression of colon cancer patients (**Supplementary Figure S3**; $p < 0.05$, log rank test).

Prognostic Value of ARGPI in Colon Cancer

Multivariate Cox regression analysis was performed on 20 apoptosis-related hub genes to determine independent prognostic genes. As shown in **Figures 3A,B**, five genes (FAS, VWA5A, SPTBN2, PCK1, and TIMP1) significantly affect the prognosis of patients with colon cancer. Then, we constructed the ARGPI for all colon cancer patients. Through the Risk formula = Expression level of FAS * (−0.2326) + Expression level of VWA5A * (−0.3434) + Expression level of SPTBN2 * 0.3991 + Expression level of PCK1 * (−0.1833) + Expression level of TIMP1 * 0.3263. Taking the median risk value as the critical value, low-risk patients had better prognosis than high-risk patients ($p < 0.001$, log rank test) (**Figure 3C**). Then, the role of the prognostic model was verified by using the GSE39582 (n = 517) colon cancer data set. As shown in **Figure 3D**, the prognosis of patients in the low-risk group was significantly better than that in the high-risk group, which was concordant to the results of the TCGA dataset ($p = 0.018$, log rank test).

The distribution of clinicopathological features of 379 colon cancer patients in the TCGA cohort in different risk subgroups is

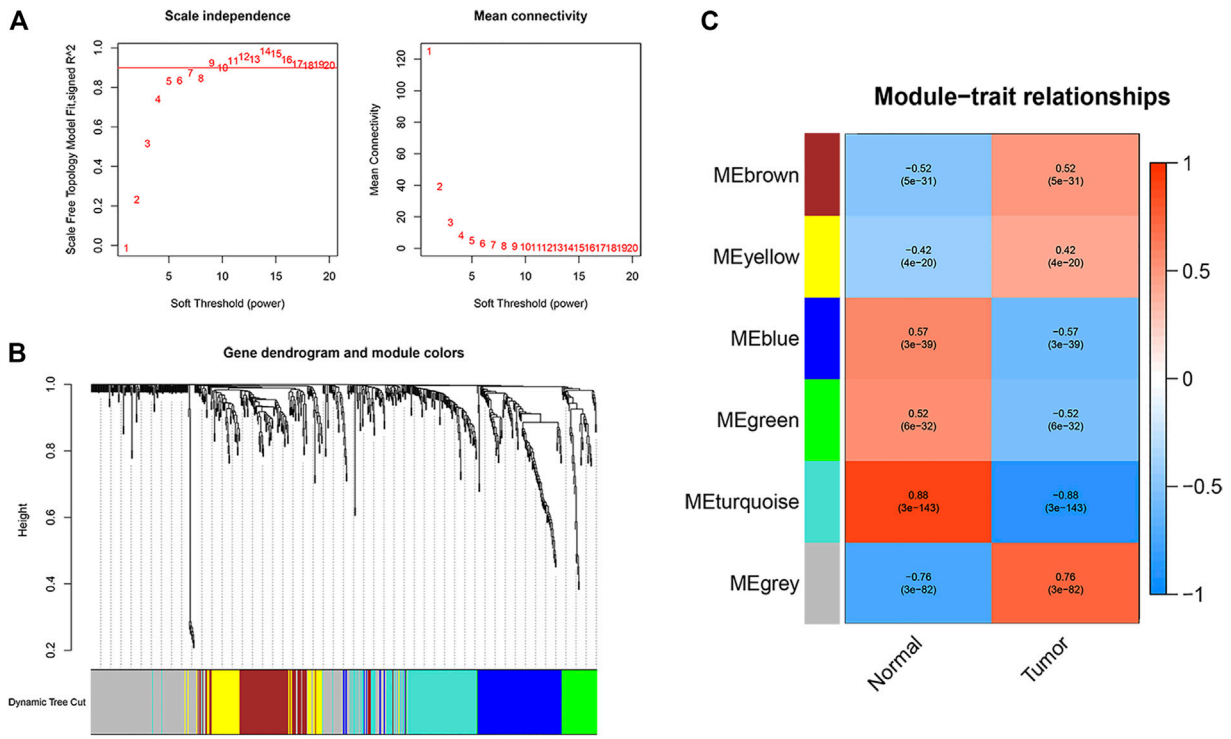


FIGURE 2 | Identification of apoptosis related hub genes. **(A)** Determination of the soft-thresholding power in the WGCNA analysis. **(B)** Gene dendrogram and module colors. **(C)** Gene modules related to COAD obtained by WGCNA.

shown in **Figure 4A**. In **Figure 4B** shows significant differences in the distribution of tumor stages between the two risk subgroups. Univariate Cox regression analysis showed that stage and risk score were significantly correlated to colon cancer prognosis (**Figure 4C**). Multivariate Cox regression analysis confirmed that risk score was an independent prognostic factor after adjusting for other clinicopathological factors (**Figure 4D**).

Molecular Characteristics of Different ARGPI Subgroups

GSEA analysis was performed to identify pathways enriched in different risk subgroups. High-risk subgroup enriched CELL_CYCLE and FOCAL_ADHESION and other related pathways (**Figure 5A**). The low-risk subgroup was enriched in DRUG_ METABOLISM_ CYTOCHROME_ P450, RETINOL_ Metabolism and other related pathways (**Figure 5B**). **Supplementary Table S2** lists the detailed results of GSEA, in which apoptosis related pathways are also enriched.

Next, we analyzed gene mutations to gain further biological insight into the nature of subgroup differences. We found that the mutation count in the low-risk subgroup was higher than that in the high-risk subgroup. Missense mutations were the most common mutation type, followed by nonsense mutations and frameshift deletions. We then identified the top 20 genes with the highest mutation rate in the risk subgroup (**Figures 5C,D**). The mutation rates of APC, TP53, TTN, KRAS, SYNE1, PIK3CA, MUC16 and FAT4 in the two groups were >20%.

Immune Microenvironment of Different ARGPI Subgroups

To analyze the tumor microenvironment status in different risk subgroups, we first used CIBERSORT algorithm to estimate the proportion of 22 types of invasive immune cells in different risk subgroups. We found that macrophages M0, macrophages M1 and T cells regulatory (Tregs) were higher in the high-risk subgroup, while B cells naïve, plasma cells, T cells CD4 memory resting and eosinophils were more abundant in the low-risk subgroup (**Figure 6A**). We then applied the ssGSEA score to quantify the enrichment level of 29 immune features in each colon cancer sample. The results showed that there were more macrophages in the high-risk subgroup and more IDCs and NK_cells correlation signals in the low-risk subgroup (**Figure 6B**). We further studied the effects of immune and molecular functional differences between different risk subgroups on survival. As shown in **Figures 6C,D**, patients with more IDCs and NK_cells have better prognosis.

Therapeutic Benefits of Different ARGPI Subgroups

To explore the relationship between apoptosis characteristics and treatment, we analyzed the chemotherapy data and molecular characteristics of colon cancer in the GEO (GSE39582) databases. As shown in **Figure 7A**, the high-risk group is more likely to progress after adjuvant chemotherapy than the low-risk group. Therefore, we observed the effect of the expression of five core genes on the progression after chemotherapy. As shown in **Figures 7B,C**, when

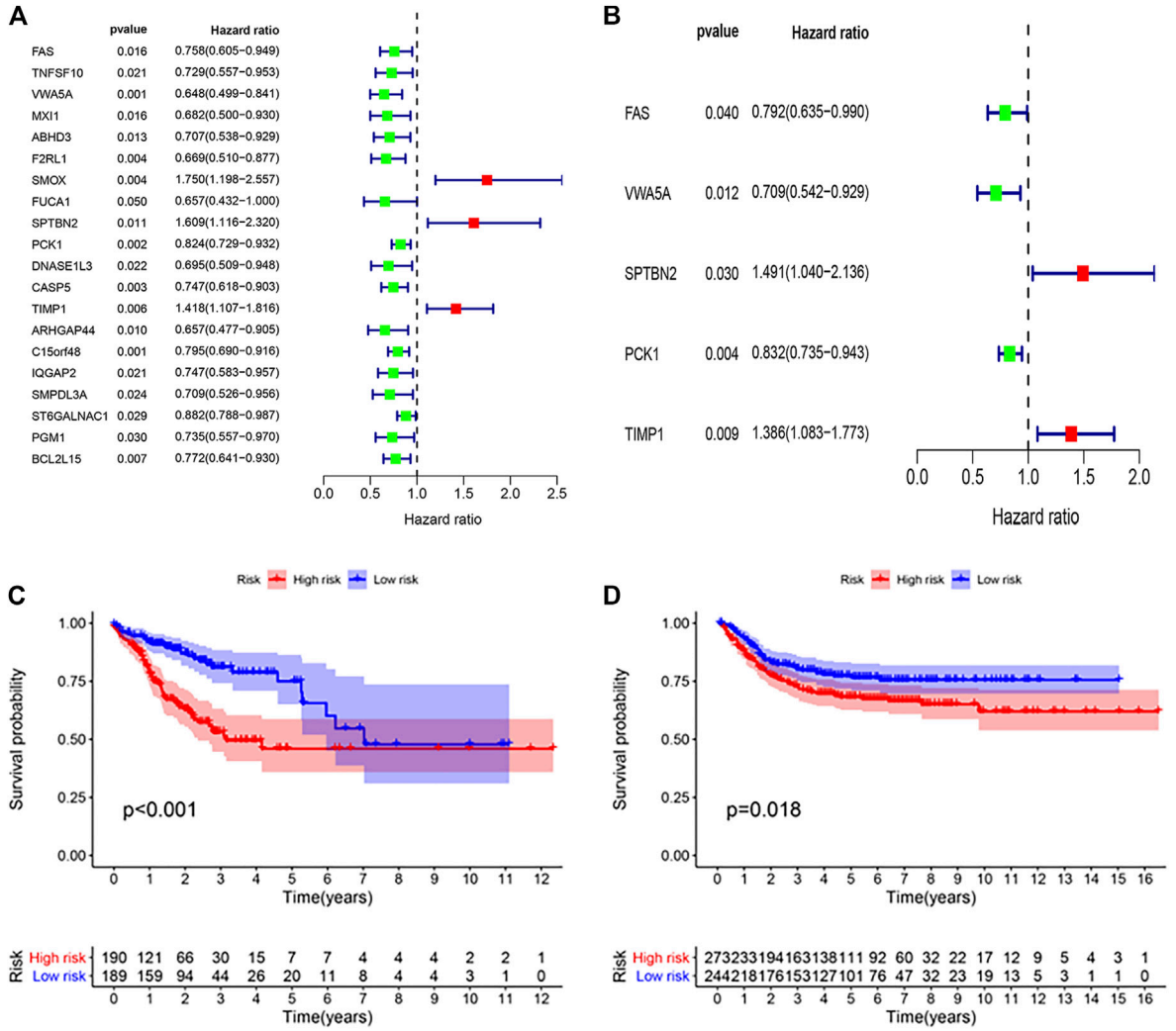


FIGURE 3 | Prognostic analysis of different ARGPI subgroups. **(A)** Univariate Cox analysis of 20 apoptosis-related hub genes. **(B)** Apoptosis-related hub gene is independently associated with the prognosis of colon cancer. **(C)** Kaplan–Meier survival analysis of the ARGPI subgroups in the TCGA cohort. **(D)** Kaplan–Meier survival analysis of the ARGPI subgroups in the GEO cohort.

the VWA5A gene is highly expressed, tumor patients make less progress after chemotherapy, suggesting that the upregulation of this gene promote chemotherapy sensitivity. When the TIMP1 gene is highly expressed, tumor patients progress more after chemotherapy, suggesting that the upregulation of TIMP1 gene reduce chemotherapy sensitivity. **Figure 7D** again shows that when the tumor shows upregulation of the VWA5A gene and downregulation of the TIMP1 gene, the tumor has better chemosensitivity. Stratified analysis showed that the apoptosis-related prognostic model was still a clinically and statistically significant prognostic model in patients with pMMR status, BRAF wild type, and KRAS mutation (**Figures 7E–H**).

A Nomogram for Predicting the Prognosis of Patients with Colon Cancer

Receiver operating characteristic curve predicted by two-year progression showed that the ARGPI had higher prognostic accuracy than other clinicopathological features (**Figure 8A**). Then, to provide clinicians with a quantitative method to predict the possible risk of cancer progression, we constructed a nomogram based PFI, which integrates the ARGPI and four clinicopathological risk factors (**Figure 8B**). The prediction accuracy of the nomogram is as follows. The calibration curve of the nomogram shows that it does not deviate from the reference line and does not need to be recalibrated [**Figure 8C** (1 year), **Figure 8D** (3 years), **Figure 8E** (5 years)].

DISCUSSION

Colorectal cancer is the second most common cause of cancer death in the United States. According to the cancer statistics of

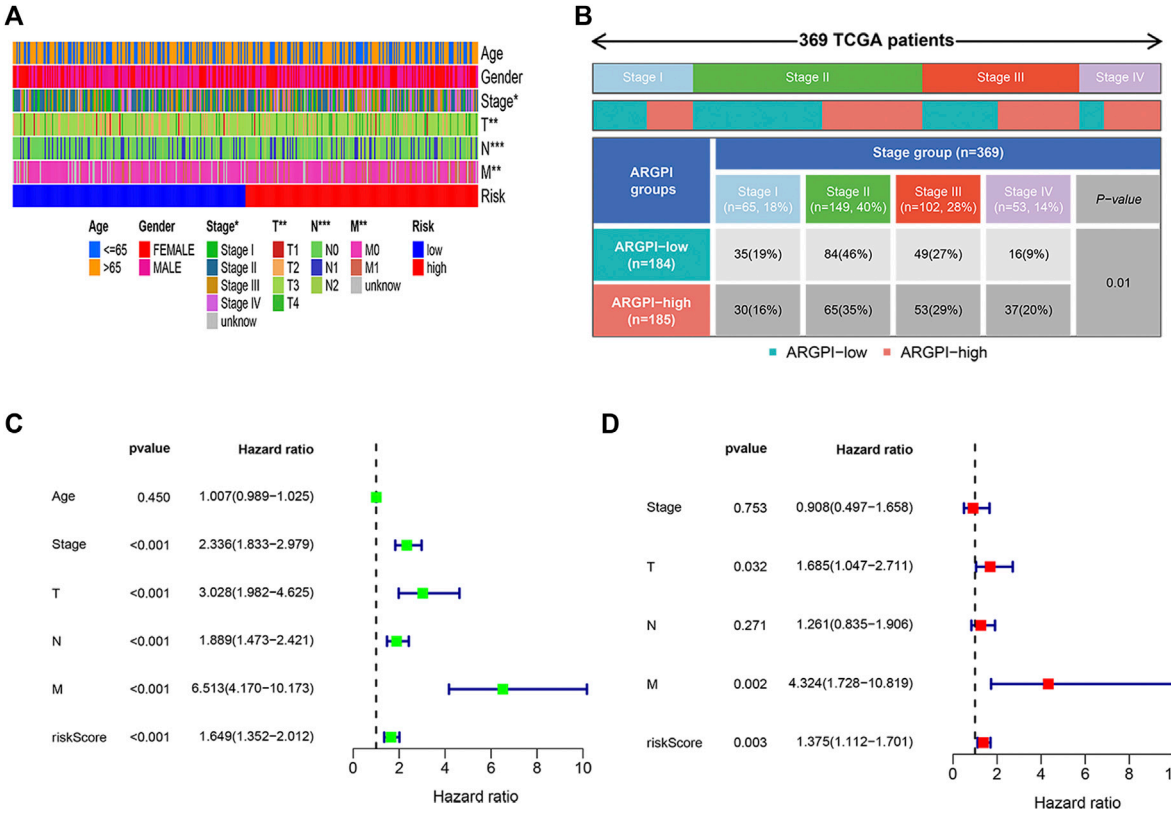


FIGURE 4 | Clinical characteristics and independent prognosis of different ARGPI subgroups. **(A)** Distribution of clinical characteristics of ARGPI subgroup in TCGA cohort. **(B)** Differential distribution of tumor stage in ARGPI subgroup in TCGA cohort ($p = 0.01$). **(C)** Univariate Cox analysis of clinicopathological factors and the ARGPI score. **(D)** Independent prognostic analysis included clinicopathological factors and ARGPI score.

the United States in 2021, it is estimated that there are about 149,500 patients with new colorectal cancer and about 52,980 patients who die from colorectal cancer (Siegel et al., 2021). Colon cancer is a malignant tumor that is easy to relapse after treatment, tumor progression is closely related to prognosis. At present, the depth of tumor invasion, regional lymph node invasion and distant metastasis (TNM stage) are still the most accurate prognostic tools. However, with the application of next-generation sequencing, molecular markers can be independent of TNM staging. Through risk stratification of colorectal cancer patients, the prediction of survival results has reached a high level of reliability (Yu et al., 2015). Although TNM staging remains the most convincing tool for us to evaluate the prognosis of patients and guide further treatment, the continuous expansion of our understanding of the molecular level in the occurrence and development of colon cancer allows improvement of staging and the prognosis system.

Apoptosis is an autonomous and orderly cell death controlled by genes to maintain the stability of the internal environment (Tower, 2015). Apoptosis is a complex process involving a large number of genes. The changes of these apoptotic genes play an important role in the formation, development, metastasis and drug resistance of malignant tumors (Fan and Guo, 2001; Wong, 2011). The role of apoptosis in the occurrence and development of colon cancer has been fully valued and studied. For example, changes in important molecules of key protein signaling pathways inducing apoptosis such as p53, APC, and RAS accumulate genetic errors, increasing the risk of colon cancer progression (Watson, 2006). In addition, changes in apoptosis pathways play an important role in anti-tumor therapy (such as radiotherapy and chemotherapy) (Potten and Grant 1998; Johnstone et al., 2002). These research results have prompted us to investigate apoptosis-related genes, their prospects as prognostic markers, and further provide new ideas for clinical treatment.

By combining the high-throughput expression profile of apoptosis-related genes and WGCNA analysis method that can effectively identify candidate biomarkers, we analyzed the prediction efficiency of apoptosis-related gene characteristics in predicting the prognosis of colon cancer patients and the potential impact on chemotherapy. Finally, the ARGPI was constructed based on five apoptosis genes (FAS, VWA5A, SPTBN2, PCK1, and TIMP1), which can effectively improve the prediction of colon cancer progression. In this model, both TCGA and GEO cohorts demonstrated that patient progression rate is higher in the high-risk subgroup and

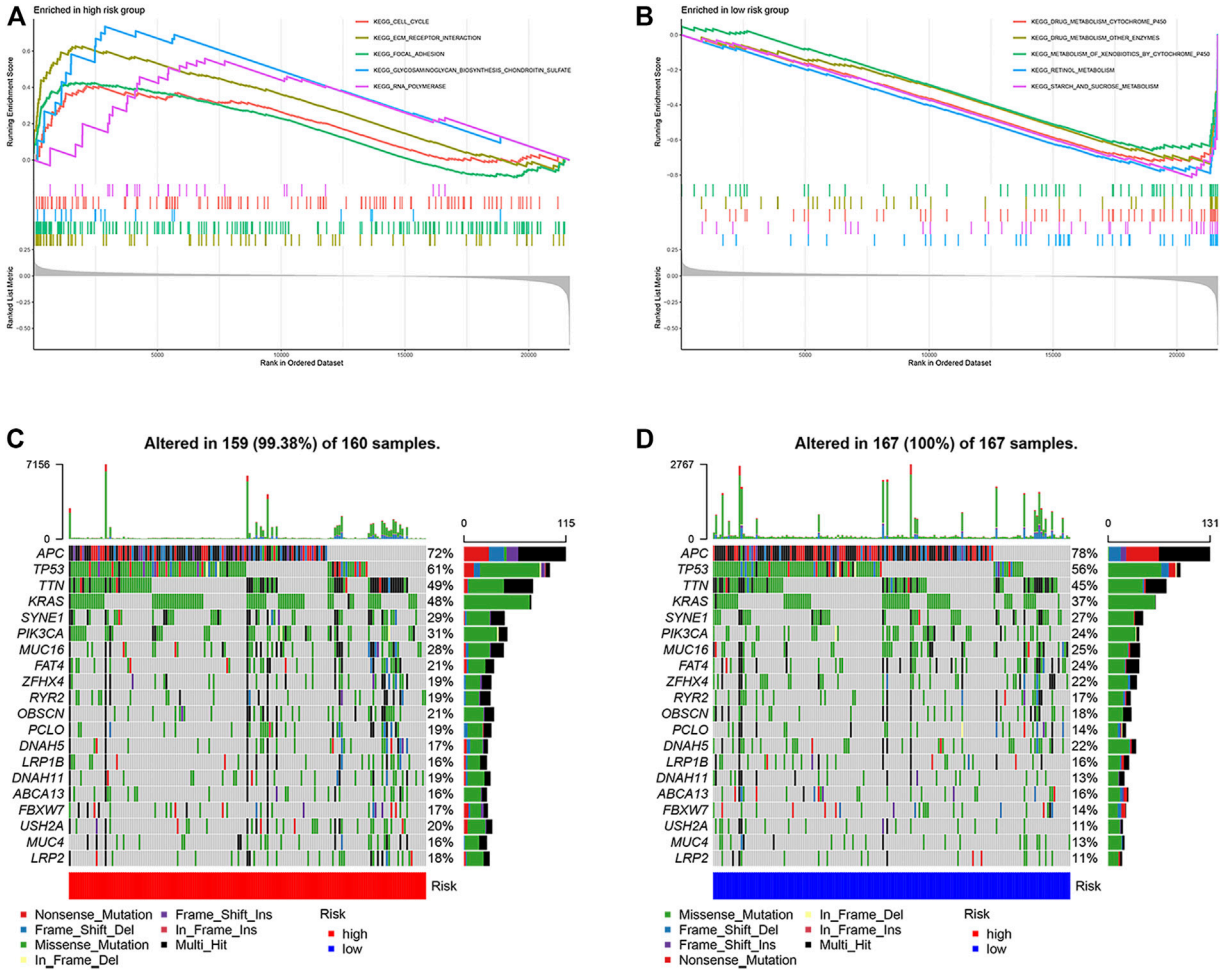


FIGURE 5 | Molecular characteristics of different ARGPI subgroups. **(A)** Gene sets enriched in ARGPI-high subgroup. **(B)** Gene sets enriched in ARGPI-low subgroup. **(C)** Significantly mutated genes in the mutated COAD samples in ARGPI-high subgroup. **(D)** Significantly mutated genes in mutated COAD samples in ARGPI-low subgroup.

lower in the low-risk subgroup. In terms of its ability to predict progression, the predictive value of two-year progression is better than the classical TNM staging, which means that apoptosis signals can be used to improve the current prognostic evaluation system. Therefore, we combined the ARGPI and four clinicopathological risk variables (T, N, M, and age) to construct a nomogram to predict the probability of progression at one, three, and 5 years. This will help to effectively identify high-risk patients with progression and guide the formulation of more reasonable treatment schemes for colon cancer, which is effective for clinical application.

The ARGPI was comprised of five genes, FAS, VWA5A, SPTBN2, PCK1, and TIMP1. FAS (also known as TNFRSF6 or CD95) is a transmembrane protein. It is a cell surface receptor of a member of the tumor necrosis factor receptor superfamily. Its binding with FasL can start the transduction of apoptotic signals and cause apoptosis (Xiao et al., 2019). The expression of Fas in colorectal tissues of patients with colorectal cancer is often lower than in normal colorectal tissues. Studies have shown that the decrease in Fas expression may be the reason for the reduction in apoptosis in colon cancer cells (Mo et al., 2015). Silencing Fas expression promotes colon cancer immune escape and 5-fluorouracil resistance (Paschall et al., 2015). VWA5A (also known as BCSC-1 or LOH11CR2A), located on chromosome 11q23, contains two conserved domains, one of which is between the N-terminal vault proteins α-trypsin inhibitor domain, a C-terminal von Willebrand factor type a domain. Because it participates in the process of apoptosis induced by p53, it is considered as a potential tumor suppressor genes (Gentile et al., 2001). In breast cancer, BCSC-1 inhibits NF-κB signal transduction that disrupts breast cancer metastasis (Di et al., 2018). In nasopharyngeal carcinoma, the decrease of Wnt signal transduction may be involved in the tumor inhibition mechanism of BCSC-1 (Zhou et al., 2009). The main function of SPTBN2 (also known as GTRAP41 or SCAR14) is to regulate glutamate signaling pathway by stabilizing

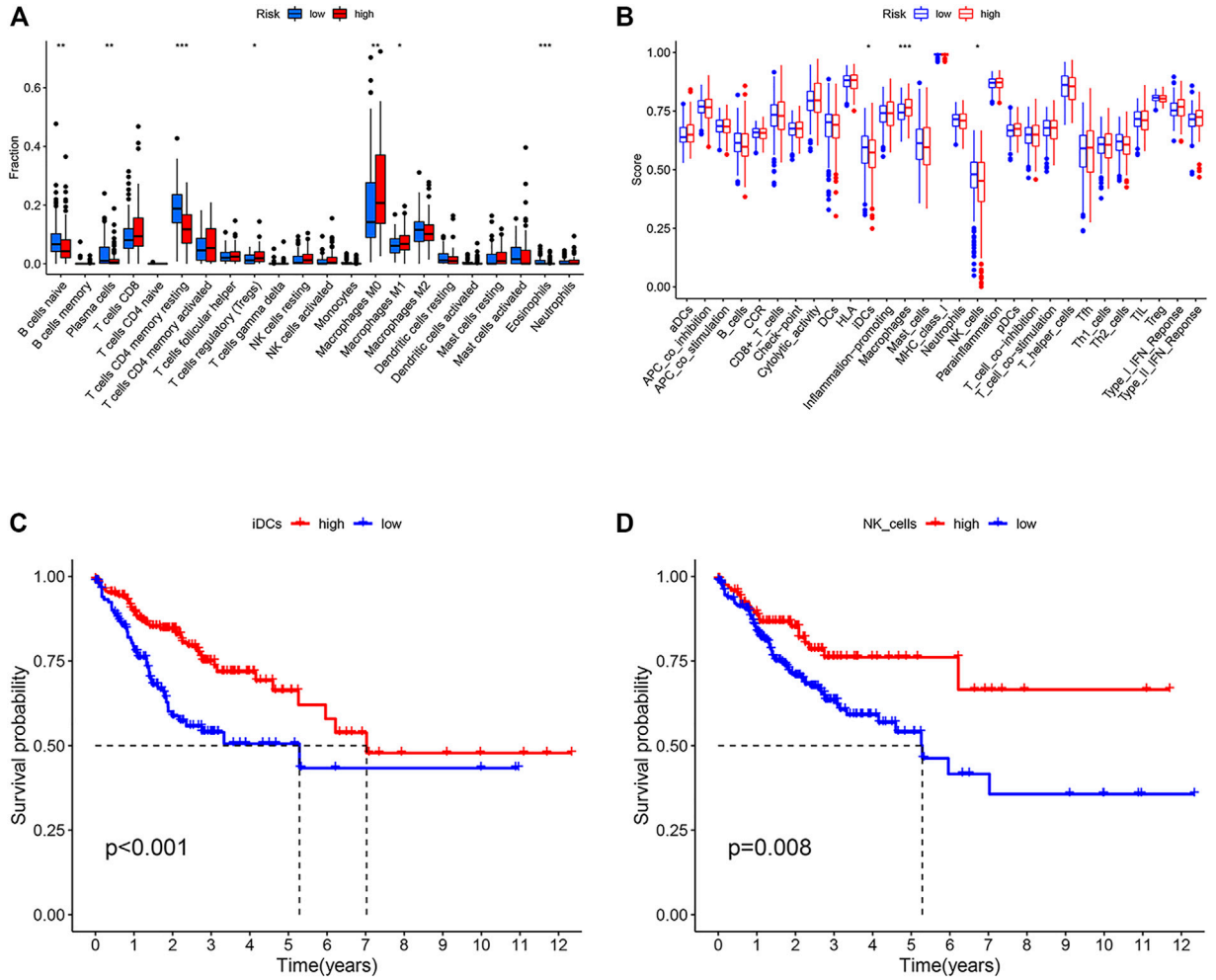


FIGURE 6 | Immune characteristics of different ARGPI subgroups. **(A)** The proportions of the 22 immune cells in different ARGPI subgroups. **(B)** Immune related functions in different ARGPI subgroups. **(C)** Relationship between immune function and survival of interdigitating dendritic cells (IDCs). **(D)** Relationship between immune function and survival of NK cells.

glutamate transporter EAAT4 on the plasma membrane surface. This gene mutation can lead to spinocerebellar ataxia (Yıldız et al., 2017). In ovarian cancer, the upregulation of SPTBN2 has been associated with poor prognosis (Feng et al., 2021). TIMP1, tissue inhibitor matrix metalloproteinase-1, belongs to the tissue inhibitor of metalloproteinases family. This protein can inhibit the proteolytic activity of matrix metalloproteinases (MMPs) and regulate the balance of matrix remodeling during extracellular matrix degradation (Batra et al., 2012). Studies have shown that TIMP1 is highly expressed in colon cancer and leads to tumor proliferation, metastasis, and anti-apoptosis through the FAK-PI3K/AKT and MAPK pathways (Song et al., 2016). A randomized controlled study showed that in patients with metastatic colorectal cancer receiving first-line combined chemotherapy, the upregulation of plasma TIMP-1 indicates that the objective effective rate of chemotherapy is very low (Sørensen et al., 2007). Studies have shown that the overexpression of TIMP-1 leads to reduced sensitivity to chemotherapy (Fu et al., 2011; Aaberg-Jessen et al., 2019). PCK1 (phosphoenolpyruvate carboxykinase 1) is a ladder-limiting enzyme of gluconeogenesis. Studies have shown that the expression level of PCK1 in liver cancer specimens and liver cancer cells is lower than that in normal samples. Overexpression of PCK1 can reduce the survival rate of liver cancer cell lines, induce apoptosis, inhibit cell migration, and activate the expression of PCK1 and thus may be potentially utilized as a therapeutic strategy for liver cancer (Liu MX. et al., 2018; Tang Y. et al., 2018).

Next, we analyzed gene mutations in different ARGPI subsets. Missense mutations, nonsense mutations, and frameshift deletion are the most common mutations in colon cancer, which is concordant to the reported tumor mutation load in colon cancer (Lee et al., 2017). In our study, the tumor mutation burden (TMB) was higher in the

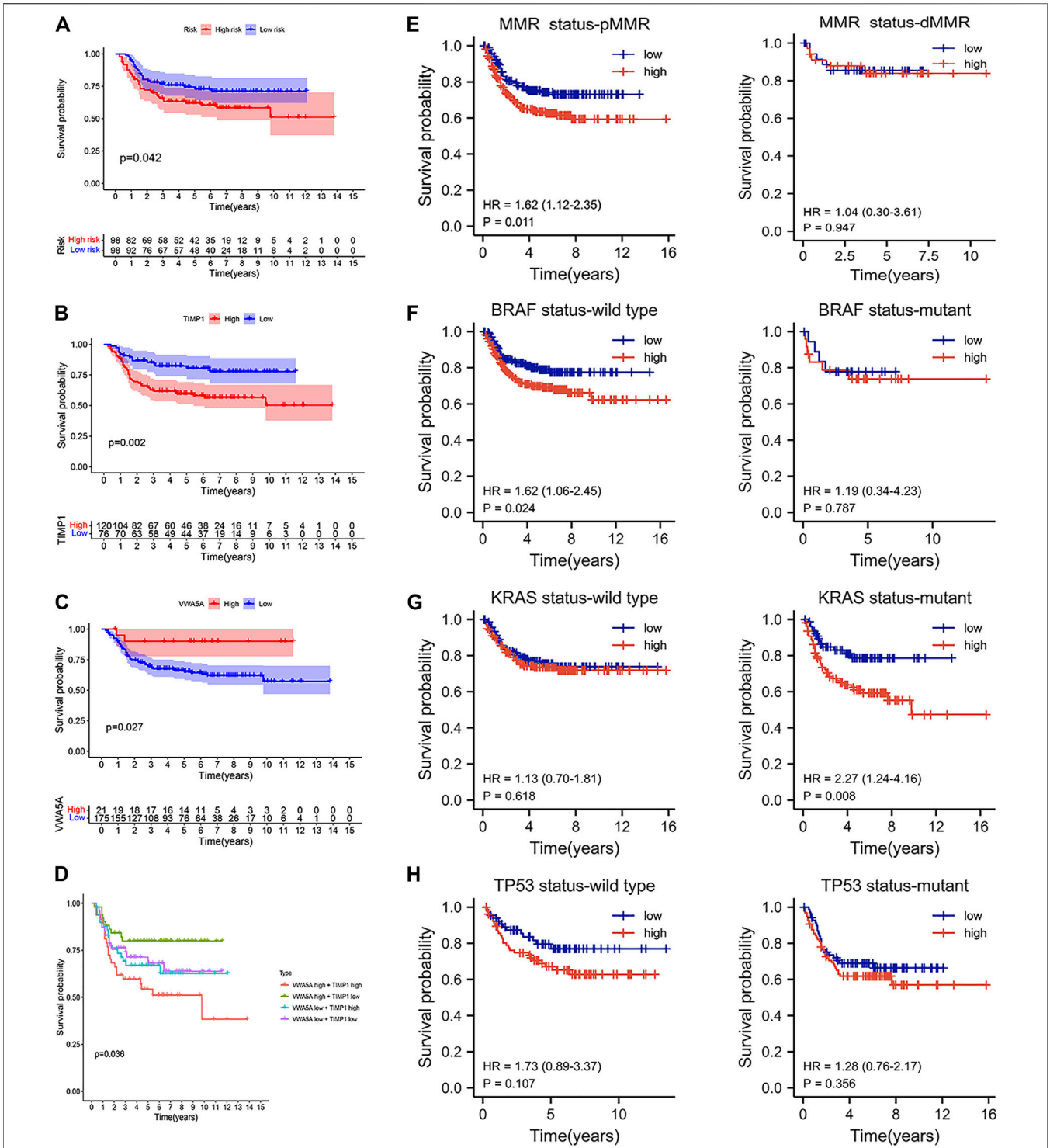


FIGURE 7 | The benefit of therapy in different ARGPI subgroups. **(A)** Prognostic value of ARGPI in adjuvant chemotherapy of colon cancer. **(B)** Relationship between TIMP1 gene expression and the effect of adjuvant chemotherapy. **(C)** Relationship between VWA5A gene expression and the effect of adjuvant chemotherapy. **(D)** Double gene expression and the effect of adjuvant chemotherapy. **(E–H)** Effect of MMR/KRAS/BRAF/TP53 status on survival of different ARGPI subgroups.

low-risk subgroup than the high-risk subgroup. High-TMB has been associated with better prognosis in colorectal cancer patients undergoing postoperative adjuvant chemotherapy (Lee et al., 2019). This may be one of the reasons for the observed lower recurrence rate after adjuvant chemotherapy in the low-risk and high-risk subgroup. The largest mutation

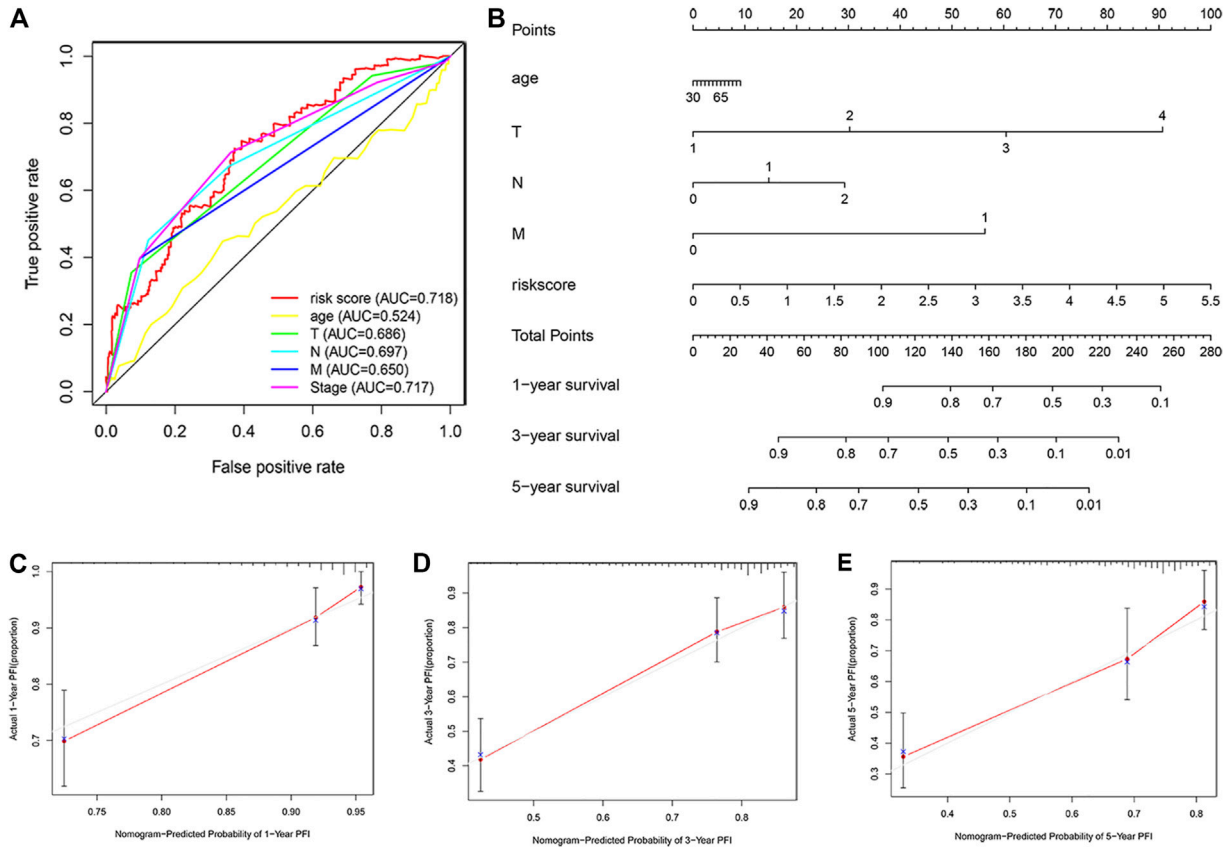


FIGURE 8 | Nomograms for predicting colon cancer progression. **(A)** The receiver operating characteristic curve evaluates the accuracy of various indicators in predicting cancer progression within 2 years. **(B)** A nomogram predicts the risk of progression in patients with colon cancer by combining ARGPI and four clinicopathological features. **(C–E)** The calibration curve is used to evaluate the accuracy of one-, three-, and five-year progress forecasts of nomograms.

difference between the two groups was observed in the KRAS, which was more common in high-risk samples than in low-risk subgroups (48 *vs.* 37%). Studies have shown that KRAS mutations may lead to poor prognosis by enhancing tumor cell proliferation (McCuaig et al., 2020). Clinical controlled studies have shown that the recurrence risk of KRAS mutant tumors is significantly higher than KRAS wild-type tumors (Hutchins et al., 2011). In our model, among patients with KRAS mutation, the low-risk subgroup made less progress than the high-risk subgroup.

Finally, we analyzed the tumor microenvironment of different ARGPI subsets. The composition of some immune cells in different ARGPI subsets varied. Macrophages M0 and M1 and T cells regulation (Tregs) were more abundant in the high-risk subgroup, while B cells naïve, plasma cells, T cells CD4 memory resetting and eosinophils were more common in the low-risk subgroup. Regulatory T cells (Tregs) play an important role in maintaining immune dynamic balance. However, studies have shown that inhibition of Treg function enhances antitumor effects (Yan et al., 2019). In addition, the activation of Wnt/-catenin signal in Tregs is related to the promotion of colon cancer (Keerthivasan et al., 2014). A large number of studies have shown that intensive infiltration of T cells, especially resting CD4 memory T cells, predicts good prognoses (Jin and Qin, 2020; Zhang et al., 2020). GSEA analysis showed that the high-risk subgroup was enriched with DEGs related to CELL CYCLE pathways, while low risk subgroups were enriched with DEGs associated with DRUG METABOLISM CYTOCHROME P450 pathways. This suggests that high-risk subgroups are characterized by tumor progression (Evan and Vousden, 2001). Most tumor chemotherapeutic drugs need to be achieved through CYP450 enzyme metabolism, and the low-risk subgroup may be more chemosensitive (Scripture et al., 2005).

CONCLUSION

The ARGPI is a promising apoptosis related prognostic biomarker. The ARGPI, combined with TNM staging, may help distinguish high-risk groups, more accurately evaluate the prognosis of patients, and guide the formulation of more scientific treatment strategies. The ARGPI may also potentially

used as a prognostic indicator of chemotherapy, although further research is warranted.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

HT: research design and drafting the manuscript. JW, XL, CG, and XZ: data analysis. QW, JC, QL, and YL: the existing literature was searched and read. SH: review, revision and writing guidance of manuscripts. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Innovation Capacity Support Plan of Shaanxi Province (under Grant # 2018TD-002).

ACKNOWLEDGMENTS

The author would like to thank Xi'an Jiaotong University for its bioinformatics related training and support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.790878/full#supplementary-material

REFERENCES

Aaberg-Jessen, C., Fogh, L., Sørensen, M. D., Halle, B., Brünner, N., and Kristensen, B. W. (2019). Overexpression of TIMP-1 and Sensitivity to Topoisomerase Inhibitors in Glioblastoma Cell Lines. *Pathol. Oncol. Res.* 25, 59–69. doi:10.1007/s12253-017-0312-5

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA Interference Reveals that Oncogenic KRAS-Driven Cancers Require TBK1. *Nature* 462, 108–112. doi:10.1038/nature08460

Batra, J., Robinson, J., Soares, A. S., Fields, A. P., Radisky, D. C., and Radisky, E. S. (2012). Matrix Metalloproteinase-10 (MMP-10) Interaction with Tissue Inhibitors of Metalloproteinases TIMP-1 and TIMP-2: Binding Studies and crystal Structure. *J. Biol. Chem.* 287, 15935–15946. doi:10.1074/jbc.M112.341156

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J. Clinicians* 68, 394–424. doi:10.3322/caac.21492

Chen, Y., Li, Z.-Y., Zhou, G.-Q., and Sun, Y. (2021). An Immune-Related Gene Prognostic Index for Head and Neck Squamous Cell Carcinoma. *Clin. Cancer Res.* 27, 330–341. doi:10.1158/1078-0432.ccr-20-2166

Di, D., Chen, L., Guo, Y., Wang, L., Zhao, C., and Ju, J. (2018). BCSC-1 Suppresses Human Breast Cancer Metastasis by Inhibiting NF-Kb Signaling. *Int. J. Oncol.* 52, 1674–1684. doi:10.3892/ijo.2018.4309

Evan, G. I., and Vousden, K. H. (2001). Proliferation, Cell Cycle and Apoptosis in Cancer. *Nature* 411, 342–348. doi:10.1038/35077213

Fan, X. Q., and Guo, Y. J. (2001). Apoptosis in Oncology. *Cell Res* 11, 1–7. doi:10.1038/sj.cr.7290060

Feng, P., Ge, Z., Guo, Z., Lin, L., and Yu, Q. (2021). A Comprehensive Analysis of the Downregulation of miRNA-1827 and its Prognostic Significance by Targeting SPTBN2 and BCL2L1 in Ovarian Cancer. *Front. Mol. Biosci.* 8, 687576. doi:10.3389/fmolb.2021.687576

Fouqué, A., Debure, L., and Legembre, P. (2014). The CD95/CD95L Signaling Pathway: a Role in Carcinogenesis. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1846, 130–141. doi:10.1016/j.bbcan.2014.04.007

Fu, Z. Y., Lv, J. H., Ma, C. Y., Yang, D. P., and Wang, T. (2011). Tissue Inhibitor of Metalloproteinase-1 Decreased Chemosensitivity of MDA-435 Breast Cancer Cells to Chemotherapeutic Drugs through the PI3K/AKT/NF-κB Pathway. *Biomed. Pharmacother.* 65, 163–167. doi:10.1016/j.biopha.2011.02.004

Galluzzi, L., Vitale, I., Aaronson, S. A., Abrams, J. M., Adam, D., Agostinis, P., et al. (2018). Molecular Mechanisms of Cell Death: Recommendations of the Nomenclature Committee on Cell Death 2018. *Cell Death Differ* 25, 486–541. doi:10.1038/s41418-017-0012-4

Gentile, M., Ahnström, M., Schön, F., and Wingren, S. (2001). Candidate Tumour Suppressor Genes at 11q23-Q24 in Breast Cancer: Evidence of Alterations in PIG8, a Gene Involved in P53-Induced Apoptosis. *Oncogene* 20, 7753–7760. doi:10.1038/sj.onc.1204993

Hutchins, G., Southward, K., Handley, K., Magill, L., Beaumont, C., Stahlschmidt, J., et al. (2011). Value of Mismatch Repair, KRAS, and BRAF Mutations in Predicting Recurrence and Benefits from Chemotherapy in Colorectal Cancer. *J. Clin. Oncol.* 29, 1261–1270. doi:10.1200/JCO.2010.30.1366

Jin, Y., and Qin, X. (2020). Profiles of Immune Cell Infiltration and Their Clinical Significance in Head and Neck Squamous Cell Carcinoma. *Int. Immunopharmacol* 82, 106364. doi:10.1016/j.intimp.2020.106364

Johnstone, R. W., Ruefli, A. A., and Lowe, S. W. (2002). Apoptosis: a Link between Cancer Genetics and Chemotherapy. *Cell* 108, 153–164. doi:10.1016/s0092-8674(02)00625-6

Keerthivasan, S., Aghajani, K., Dose, M., Molinero, L., Khan, M. W., Venkateswaran, V., et al. (2014). β-Catenin Promotes Colitis and colon Cancer through Imprinting of Proinflammatory Properties in T Cells. *Sci. Transl Med.* 6, 225–228. doi:10.1126/scitranslmed.3007607

Kehoe, J., and Khatri, V. P. (2006). Staging and Prognosis of colon Cancer. *Surg. Oncol. Clin. North America* 15, 129–146. doi:10.1016/j.soc.2005.08.006

Labianca, R., Beretta, G. D., Kildani, B., Milesi, L., Merlin, F., Mosconi, S., et al. (2010). Colon Cancer. *Crit. Rev. Oncology/Hematology* 74, 106–133. doi:10.1016/j.critrevonc.2010.01.010

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Lee, D. W., Han, S. W., Bae, J. M., Jang, H., Han, H., Kim, H., et al. (2019). Tumor Mutation Burden and Prognosis in Patients with Colorectal Cancer Treated with Adjuvant Fluoropyrimidine and Oxaliplatin. *Clin. Cancer Res.* 25, 6141–6147. doi:10.1158/1078-0432

Lee, S. H., Jung, S. H., Kim, T. M., Rhee, J. K., Park, H. C., Kim, M. S., et al. (2017). Whole-exome Sequencing Identified Mutational Profiles of High-Grade colon Adenomas. *Oncotarget* 8, 6579–6588. doi:10.18632/oncotarget.14172

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cel Syst.* 1, 417–425. doi:10.1016/j.cels.2015.12.004

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018a). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* 173, 400–e11. doi:10.1016/j.cell.2018.02.052

Liu, M. X., Jin, L., Sun, S. J., Liu, P., Feng, X., Cheng, Z. L., et al. (2018b). Metabolic Reprogramming by PCK1 Promotes TCA Cataplerosis, Oxidative Stress and Apoptosis in Liver Cancer Cells and Suppresses Hepatocellular Carcinoma. *Oncogene* 37, 1637–1653. doi:10.1038/s41388-017-0070-6

Maji, S., Panda, S., Samal, S. K., Shriwas, O., Rath, R., Pellecchia, M., et al. (2018). Bcl-2 Antiapoptotic Family Proteins and Chemoresistance in Cancer. *Adv. Cancer Res.* 137, 37–75. doi:10.1016/bs.acr.2017.11.001

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer. *Genome Res.* 28, 1747–1756. doi:10.1101/gr.239244.118

McCuaig, S., Barras, D., Mann, E. H., Friedrich, M., Bullers, S. J., Janney, A., et al. (2020). The Interleukin 22 Pathway Interacts with Mutant KRAS to Promote Poor Prognosis in Colon Cancer. *Clin. Cancer Res.* 26, 4313–4325. doi:10.1158/1078-0432

Mo, J. S., Alam, K. J., Kang, I. H., Park, W. C., Seo, G. S., Choi, S. C., et al. (2015). MicroRNA 196B Regulates FAS-Mediated Apoptosis in Colorectal Cancer Cells. *Oncotarget* 6, 2843–2855. doi:10.18632/oncotarget.3066

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Paschall, A. V., Yang, D., Lu, C., Choi, J. H., Li, X., Liu, F., et al. (2015). H3K9 Trimethylation Silences Fas Expression to Confer Colon Carcinoma Immune Escape and 5-Fluorouracil Chemoresistance. *J. Immunol.* 195, 1868–1882. doi:10.4049/jimmunol.1402243

Potten, C. S., and Grant, H. K. (1998). The Relationship between Ionizing Radiation-Induced Apoptosis and Stem Cells in the Small and Large Intestine. *Br. J. Cancer* 78, 993–1003. doi:10.1038/bjc.1998.618

Rich, T., Watson, C. J., and Wyllie, A. (1999). Apoptosis: the Germs of Death. *Nat. Cel Biol* 1, E69–E71. doi:10.1038/11038

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Sargent, D. J., Wieand, H. S., Haller, D. G., Gray, R., Benedetti, J. K., Buyse, M., et al. (2005). Disease-free Survival versus Overall Survival as a Primary End point for Adjuvant colon Cancer Studies: Individual Patient Data from 20,898 Patients on 18 Randomized Trials. *Jco* 23, 8664–8670. doi:10.1200/JCO.2005.01.6071

Scripture, C. D., Sparreboom, A., and Figg, W. D. (2005). Modulation of Cytochrome P450 Activity: Implications for Cancer Therapy. *Lancet Oncol.* 6, 780–789. doi:10.1016/S1470-2045(05)70388-0

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer Statistics, 2021. *CA A. Cancer J. Clin.* 71, 7–33. doi:10.3322/caac.21654

Song, G., Xu, S., Zhang, H., Wang, Y., Xiao, C., Jiang, T., et al. (2016). TIMP1 Is a Prognostic Marker for the Progression and Metastasis of colon Cancer through FAK-Pi3k/AKT and MAPK Pathway. *J. Exp. Clin. Cancer Res.* 35, 148. doi:10.1186/s13046-016-0427-7

Sørensen, N. M., Byström, P., Christensen, I. J., Berglund, A., Nielsen, H. J., Brünner, N., et al. (2007). TIMP-1 Is Significantly Associated with Objective Response and Survival in Metastatic Colorectal Cancer Patients Receiving Combination of Irinotecan, 5-fluorouracil, and Folinic Acid. *Clin. Cancer Res.* 13, 4117–4122. doi:10.1158/1078-0432

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Tang, J., Kong, D., Cui, Q., Wang, K., Zhang, D., Gong, Y., et al. (2018a). Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis. *Front. Oncol.* 8, 374. doi:10.3389/fonc.2018.00374

Tang, Y., Zhang, Y., Wang, C., Sun, Z., Li, L., Cheng, S., et al. (2018b). Overexpression of PCK1 Gene Antagonizes Hepatocellular Carcinoma through the Activation of Gluconeogenesis and Suppression of Glycolysis Pathways. *Cell Physiol Biochem* 47, 344–355. doi:10.1159/000489811

Tower, J. (2015). Programmed Cell Death in Aging. *Ageing Res. Rev.* 23, 90–100. doi:10.1016/j.arr.2015.04.002

Watson, A. J. (2006). An Overview of Apoptosis and the Prevention of Colorectal Cancer. *Crit. Rev. Oncol. Hematol.* 57, 107–121. doi:10.1016/j.critrevonc.2005.06.005

Wong, R. S. (2011). Apoptosis in Cancer: from Pathogenesis to Treatment. *J. Exp. Clin. Cancer Res.* 30, 87. doi:10.1186/1756-9966-30-87

Xiao, W., Ibrahim, M. L., Redd, P. S., Klement, J. D., Lu, C., Yang, D., et al. (2019). Loss of Fas Expression and Function Is Coupled with Colon Cancer Resistance to Immune Checkpoint Inhibitor Immunotherapy. *Mol. Cancer Res.* 17, 420–430. doi:10.1158/1541-7786.MCR-18-0455

Yan, S., Zhang, Y., and Sun, B. (2019). The Function and Potential Drug Targets of Tumour-Associated Tregs for Cancer Immunotherapy. *Sci. China Life Sci.* 62, 179–186. doi:10.1007/s11427-018-9428-9

Yıldız, B. E., Afzal, M., Mumtaz, S., Ahmad, N., Malik, S., and Tolun, A. (2017). Progressive SCAR14 with Unclear Speech, Developmental Delay, Tremor, and Behavioral Problems Caused by a Homozygous Deletion of the SPTBN2 Pleckstrin Homology Domain. *Am. J. Med. Genet. A.* 173, 2494–2499. doi:10.1002/ajmg.a.38332

Yu, J., Wu, W. K. K., Li, X., He, J., Li, X.-X., Ng, S. S. M., et al. (2015). Novel Recurrently Mutated Genes and a Prognostic Mutation Signature in Colorectal Cancer. *Gut* 64, 636–645. doi:10.1136/gutjnl-2013-306620

Zhang, X., Quan, F., Xu, J., Xiao, Y., Li, X., and Li, Y. (2020). Combination of Multiple Tumor-Infiltrating Immune Cells Predicts Clinical Outcome in colon Cancer. *Clin. Immunol.* 215, 108412. doi:10.1016/j.clim.2020.108412

Zhou, Y. Q., Chen, S. L., Ju, J. Y., Shen, L., Liu, Y., Zhen, S., et al. (2009). Tumor Suppressor Function of BCSC-1 in Nasopharyngeal Carcinoma. *Cancer Sci.* 100, 1817–1822. doi:10.1111/j.1349-7006.2009.01261.x

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nat. Commun.* 10, 1523. doi:10.1038/s41467-019-09234-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Tang, Wang, Luo, Wang, Chen, Zhang, Li, Gao, Li and Han. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 15 December 2021
doi: 10.3389/fcell.2021.759542

Association of *IL33*, *IL1RL1*, *IL1RAP* Polymorphisms and Asthma in Chinese Han Children

Maolan Wu[1], *Xiangrong Zheng*[1]*, *Juan Huang*[2] and *Xiaolei Hu*[3]

[1]Department of Pediatrics, Xiangya Hospital, Central South University, Changsha, China, [2]Department of Pediatrics, The First Hospital of Changsha, Changsha, China, [3]National Institution of Drug Clinical Trial, Xiangya Hospital, Central South University, Changsha, China

Background: Genome-wide association studies have identified interleukin 33 (*IL33*), interleukin 1 receptor-like 1 (*IL1RL1*), interleukin 1 receptor accessory protein (*IL1RAP*) as asthma susceptibility loci in Europeans. IL33, IL1RL1, and IL1RAP constitute a ligand-receptor complex.

Objective: We analyzed associations of asthma susceptibility, eosinophilic airway inflammation, and response to inhaled corticosteroid (ICS) with single nucleotide polymorphisms (SNPs) of 3 genes encoding IL33, IL1RL1, and its coreceptor IL1RAP in Chinese Han nationality children.

Methods: A total of 153 non-asthmatic children and 265 asthmatic children who visited the Xiangya Hospital between September 2015 and August 2019 were recruited for this study. Pulmonary function tests, peripheral blood eosinophil counts (PBEC), and fractional exhaled nitric oxide (FeNO) tests were performed before treatment, and 3 months after treatment. Each participant's DNA was extracted from the peripheral blood, and a Mass ARRAY system was used to genotype the SNPs.

Results: The T allele of rs4742170 in *IL33* was associated with a risk of higher FeNO at baseline, and no improvement in FeNO and airway hyperresponsiveness was found after ICS treatment. The A allele of rs10208293 and C allele of rs13424006 in *IL1RL1* both were associated with lower susceptibility to asthma and lower FeNO. The TT genotype of rs1420101 and AA genotype of rs4142132 in *IL1RL1* were associated with a greater probability of improvement in PBEC after ICS treatment.

Conclusion: IL33-IL1RL1-IL1RAP complex polymorphisms are associated with childhood asthma susceptibility, eosinophilic airway inflammation, and ICS response in Chinese Han children in Hunan. We speculate that IL33-IL1RL1-IL1RAP complex polymorphisms affect the development of asthma, airway inflammation, and subsequent ICS response in childhood.

Keywords: IL33, IL1RL1, IL1RAP, SNPs, childhood asthma, eosinophilic airway inflammation, ICS, Chinese Han nationality

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Qihong Deng,
Zhengzhou University, China
Huaping Rao,
Hunan Children's Hospital, China

***Correspondence:**
Xiangrong Zheng
xrzheng@csu.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *16 August 2021*
Accepted: *22 November 2021*
Published: *15 December 2021*

Citation:
Wu M, Zheng X, Huang J and Hu X (2021) Association of IL33, IL1RL1, IL1RAP *Polymorphisms and Asthma in Chinese Han Children. Front. Cell Dev. Biol. 9:759542. doi: 10.3389/fcell.2021.759542*

Abbreviations: FeNO, fractional exhaled nitric oxide; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; GR, glucocorticoid receptor; ICS, inhaled corticosteroid; IL1RL1, interleukin-1 receptor-like 1 protein; IL1RAP, interleukin-1 receptor accessory protein; IL33, interleukin 33; MMEF, maximal mid-expiratory flow; PBEC, peripheral blood eosinophil count; PD20, provocative dose of methacholine causing a 20% drop in FEV1; SNPs, single nucleotide polymorphisms.

INTRODUCTION

Asthma is a heterogeneous disease characterized by chronic airway inflammation and is the most common chronic disease in childhood (El-Husseini et al., 2020; di Palmo et al., 2021). The development of asthma results from a complex interaction between genetic predisposition and environmental triggers (Papi et al., 2018; Wang et al., 2020). In the complex pathogenesis of asthma, estimates of heritability range from 40 to 85%, and are higher for childhood-onset asthma than it is for adult asthma (Morales and Duffy, 2019). Given the extensive heterogeneity of asthma pathogenesis, a different strategy, starting from specific phenotype observations that are more likely to be driven by genetic mechanisms, might help to increase the power of genetic studies (di Palmo et al., 2021).

It is believed that typical eosinophilic airway inflammation, formed through a series of unclear pathophysiological mechanisms, is the dominant inflammation in most asthmatic children and 50% of asthmatic adults (2020). Eosinophilic airway inflammation is a manifestation of type 2 inflammation. IL33 is known as an alarmin cytokine from the IL-1 family of cytokines and has significant roles in type 2 immunity, and diseases such as allergic diseases and refractory asthma (Cayrol and Girard, 2018, 33; Chan et al., 2019, 33; Murdaca et al., 2019, 33; Zheng et al., 2020). It is mainly secreted by the cells of barrier tissues, such as endothelial cells, fibroblasts, and epithelial cells of the skin, gastrointestinal tract, and lungs that are exposed to the environment (Cayrol and Girard, 2018, 33; Chan et al., 2019, 33; Gabryelska et al., 2019). IL33 is an important immunomodulator and is released as an alarm signal upon cell necrosis (Cayrol and Girard, 2018, 33). Major targets of IL33 are mast cells, group 2 innate lymphoid cells (ILC2s), and tissue regulatory T cells (Tregs), that constitutively express membrane receptor ST2, which is encoded by the IIL1RL1 gene (Molofsky et al., 2015; Liew et al., 2016, 33). It has been shown in many studies that IL33 is a critical regulator of these tissue-resident immune cells. IL33 binds to its receptor, ST2, to interact with the IL-1 receptor accessory protein (IL1RAP), a co-receptor made by a central five-stranded sheet surrounded by five helices placed on the cytosolic end of the protein (Martin, 2013, 33). The IL33/ST2/IL1RAP complex then induces target cells, such as mast cells and ILC2s, to release type 2 inflammatory cytokines, promote the differentiation of B cells into IgE-producing cells, and induce eosinophil proliferation, thus forming an internal environment for a type 2 inflammatory reaction in the lung or other tissues (Gasiuniene et al., 2019; Johansson and McSorley, 2019).

Clinically, type 2 airway inflammation is usually characterized by eosinophil increase or fractional exhaled nitric oxide (FeNO) elevation, while type 1 inflammation is usually characterized by the neutrophil increase (2020). ICS is the first-line medicine for long-term control of asthma. Generally, type 2 airway inflammation, which is classified as mild or moderate asthma inflammation, is rapidly improved after regular and correct use of (GINA, 2020). However, the efficacy of ICS has obvious individual differences. Patients who respond poorly to ICS treatment or who are particularly sensitive to ICS, account for a significant proportion of the total population. The wide variability of response to ICS treatment, and the extensive similarity of response among relatives, suggest a genetic basis for the differences in ICS efficacy (Chung et al., 2014). Non-response to ICS treatment is often associated with loss of work, increased family and socioeconomic burden, and so on. The number of patients with type 2 airway inflammation could be as high as 50% of patients with a poor response to ICS. The emerging immune biological agents targeting type 2 provide great hope for the control of ICS-unresponsive asthma (Papi et al., 2018; Chan et al., 2019, 33).

Even though large-scale, genome-wide association studies have demonstrated that the IL33, IL1RL1, and IL1RAP genes are associated with a susceptibility to asthma in Europeans (Savenije et al., 2014; El-Husseini et al., 2020), there is still a lack of studies into the association of gene polymorphisms of the IL33/ST2/IL1RAP complex for other population groups. Therefore, we analyzed the association of asthma susceptibility, eosinophilic airway inflammation, and response to inhaled corticosteroid (ICS), with single nucleotide polymorphisms (SNPs) of IL33, IL1RL1, and IL1RAP in Chinese Han children, to investigate the effects of those three gene polymorphisms on the risk of childhood asthma, types of airway inflammation, and ICS response in asthmatic Chinese children.

METHODS

Study Population and Design

This case-control study was done in Hunan province, China. Two hundred sixty-five asthmatic children aged 3–14 years and 153 controls were recruited between September 2015 to August 2019 at Xiangya Hospital, Central South University. Diagnosis and treatment are based on Chinese childhood asthma guidelines 2016 (Association Editorial Board Chinese Journal of Pediatrics., 2016), which closely follows the Global Initiative for Asthma (GINA, 2015) (2015). Children with a long history of exposure to second-hand smoke, respiratory infections, and systemic infections within 1 month, diagnosed with congenital lung malformations, airway obstruction, extraluminal oppression, congenital heart disease, active tuberculosis, bronchiectasis, or severe systemic diseases were excluded. In addition to not having an asthma diagnosis, children in the control group did not have a diagnosis of bronchial asthma, bronchiolitis, allergic diseases, severe systemic diseases, or a family history of allergic diseases. Basic demographic information was collected at the inception of the study.

Measurement of PBEC, FeNO, and Pulmonary Function

PBEC, FeNO, and pulmonary function were measured at the first visit, and again after 3 months of ICS treatment. A PBEC higher than 300 /μL is considered to be elevated (2020). For asthmatic children over 6 years old, Pulmonary function was

TABLE 1 | of children with (case) and without (control) doctor-diagnosed asthma (n = 418).

Characteristics	Case(n = 265)		Control (n = 153)		p-value
	N	(%)	N	(%)	
Sex					0.106
Boys	185	(69.81)	95	(62.09)	
Girls	80	(30.19)	58	(37.91)	
Age (years)					0.467
<6	60	(22.64)	30	(19.61)	
≥6	205	(77.36)	123	(80.39)	
Residence					0.271
Rural	73	(27.55)	49	(32.03)	
Urban	192	(72.45)	104	(67.97)	
Gestational age (weeks)					0.235
<37	16	(6.04)	11	(7.19)	
≥37	249	(93.96)	142	(92.81)	
Mode of delivery					0.336
Natural labor	166	(62.64)	103	(67.32)	
Caesarean birth	99	(37.36)	50	(32.68)	
Atopy					<**0.001**
Yes	90	(33.96)	0	(0.00)	
No	175	(66.04)	153	(100.00)	
Parental atopy					<**0.001**
Yes	84	(31.70)	0	(0.00)	
No	181	(68.30)	153	(100.00)	

The values p ≤ *0.05 are in bold.*

performed using the Jaeger Masterscope spirometry system (Jaeger, Wuerzburg, Germany). A provocative dose of methacholine causing a 20% drop in FEV1 (PD20) was used to represent airway hyperresponsiveness. Forced expiratory volume in 1s/Forced vital capacity (FEV1/FVC), and percentage of the predicted value of maximal mid-expiratory flow (MMEF/pre), were used to evaluate pulmonary function. A Nakoulon breath machine (Sunvou, China) was used to measure FeNO. A FeNO measurement greater than 25ppb is considered to be elevated (Pulmonary Function Cooperation Group of Respiratory Group Pediatrics Society of Chinese Medical Association Editorial Board of the Chinese Journal of Clinical Applied Pediatrics, 2017).

SNP Selection and Genotyping

In this study, 10 SNPs in three genes (rs4742170, rs2381416, rs928413, and rs992969 in IL33; rs10208293, rs13424006, rs1420101 and rs4142132 in IL1RL1; rs9290936 and rs10513854 in IL1RAP) were chosen for genotyping, based on their potential functionality and their reported association with asthma (Savenije et al., 2014; Morales and Duffy, 2019; Schoettler et al., 2019). SNPs were excluded if the minor allele frequency in CHB was less than 0.1, or the r^2 value was less than 0.8.

Genotyping was performed using the iPLEX Mass ARRAY genotyping platform (Sequenom, Inc., San Diego, CA). DNA was extracted from 2 ml of the collected blood using a DNA extraction kit (SQ Blood DNA KitII, Omega, United States). The primers were designed by AssayDesigner4.0.

Statistical Analysis

SPSS24.0 (SPSS Inc., Tokyo, Japan) was used for statistical analysis. In the analysis of the difference between the genotype frequency and the expected genotype frequency, $p > 0.05$ was considered to be consistent with the Hardy-Weinberg equilibrium (HWE). T-test, Chi-square test, and logistic regression were used to calculate the significance of differences, $p < 0.05$ was considered to be statistically significant.

RESULTS

Population Characteristics

From September 2015 to August 2019, 265 asthmatic children (185 males, 80 females, mean age 7.62 ± 2.93 years) and 153 healthy controls (95 males, 58 females, mean age 7.17 ± 2.66 years) were recruited in the outpatient or emergency department of pediatrics and the pediatric ward of Xiangya Hospital, Central South University. There was no significant difference in sex, age, permanent residence, and birth history between the two groups. The detailed baseline demographics of subjects are listed in **Table 1**.

Of the265 children with asthma, 200 received standardized treatment for 3 months, 117 children were selected for examination and re-examination, 103 children were over 6 years old. **Table 2** presents a comparison of indicators from before and after treatment. The mean values of PBEC and FeNO decreased after treatment, while the mean values of FEV1/FVC and MMEF/PRE, increased.

Association of IL33-IL1RL1-IL1RAP Complex SNPs With Childhood Asthma Susceptibility

All the SNPs involved in our study were consistent with HWE. The allele and genotype frequencies of the 10 SNPs in asthmatics and controls are listed in **Table 3**. **Table 4** shows the significant results for the susceptibility of children to asthma. We observed that rs10208293 and rs13424006 in IL1RL1 were significantly different between asthmatic children and controls in allele frequencies (p = 0.015 and p = 0.017, respectively; **Table 3**). Children with an AA or AG genotype of rs10208293 had a decreased risk for asthma, compared with the other genotypes (corrected OR = 0.577, 95%CI: 0.359-0.927, p = 0.023). Children with a CT or CC genotype of rs13424006 (corrected OR = 0.584, 95%CI: 0.362-0.941, p = 0.027).

TABLE 2 | Indicators at the first visit and after 3 months of ICS treatment.

Indicator	Baseline	After ICS treatment	p-value
PBEC (/μL)	463.64 ± 369.69	382.64 ± 290.12	**0.002**
FeNO(bbp)	46.91 ± 30.66	20.19 ± 15.64	<**0.001**
FEV1/FVC(%)	93.06 ± 11.90	99.13 ± 16.83	**0.002**
MMEF/pre (%)	63.60 ± 24.12	76.24 ± 22.16	<**0.001**
PD20	0.81 ± 0.78	1.17 ± 0.88	**0.035**

*ICS: inhaled corticosteroid; PBEC: peripheral blood eosinophil count; FeNO: fractional exhaled nitric oxide; FEV1: forced expiratory volume in 1 s; FVC: forced vital capacity; MMEF: maximal mid-expiratory flow; PD20: provocative dose of methacholine causing a 20% drop in FEV1.*p*-values < 0.05 are in bold.*

TABLE 3 | The allele and genotype frequency of 10 SNPs in children with (case) and without (control) doctor-diagnosed asthma.

Gene	SNP	Genotype/Allele	Case (N = 265)		Control (N = 153)		*p*-value	[a]DOM model *p*-value	[b]REC model *p*-value
			N	(%)	N	(%)			
IL33	rs4742170	Total	262	(100.00)	146	(100.00)	0.271	0.174	0.207
		CC	48	(18.32)	35	(23.97)			
		TC	134	(51.15)	75	(51.37)			
		TT	80	(30.53)	36	(24.66)			
		C	230	(43.89)	145	(49.66)	0.113		
		T	294	(56.11)	147	(50.34)			
	rs2381416	Total	261	(100.00)	150	(100.00)	0.812	0.535	0.666
		AA	235	(90.04)	137	(91.33)			
		CA	24	(9.20)	13	(8.67)			
		CC	2	(0.77)	0	(0.00)			
		A	494	(94.64)	287	(95.67)	0.513		
		C	28	(5.36)	13	(4.33)			
	rs928413	Total	262	(100.00)	151	(100.00)	0.504	0.302	0.460
		AA	224	(85.5)	133	(88.08)			
		GA	35	(13.36)	18	(11.92)			
		GG	3	(1.15)	0	(0.00)			
		A	483	(92.18)	284	(94.04)	0.316		
		G	41	(7.82)	18	(5.96)			
	rs992969	Total	264	(100.00)	152	(100.00)	0.555	0.557	0.512
		AA	3	(1.14)	0	(0.00)			
		AG	21	(7.95)	11	(7.24)			
		GG	240	(90.91)	141	(92.76)			
		A	27	(5.11)	11	(3.62)	0.320		
		G	501	(94.89)	293	(96.38)			
IL1RL1	rs10208293	Total	263	(100.00)	152	(100.00)	**0.015**	**0.010**	1.000
		AA	1	(0.38)	1	(0.66)			
		AG	48	(18.25)	44	(28.95)			
		GG	214	(81.37)	107	(70.39)			
		A	50	(9.51)	46	(15.13)	**0.015**		
		G	476	(90.49)	258	(84.87)			
	rs13424006	Total	263	(100.00)	150	(100.00)	**0.025**	**0.012**	1.000
		CC	1	(0.38)	1	(0.67)			
		CT	48	(18.25)	43	(28.67)			
		TT	214	(81.37)	106	(70.67)			
		C	50	(9.51)	45	(15.00)	**0.017**		
		T	476	(90.49)	255	(85.00)			
	rs1420101	Total	262	(100.00)	149	(100.00)	0.396	0.605	0.303
		CC	82	(31.30)	43	(28.86)			
		TC	127	(48.47)	82	(55.03)			
		TT	53	(20.23)	24	(16.11)			
		C	291	(55.53)	168	(56.38)	0.815		
		T	233	(44.47)	130	(43.62)			
	rs4142132	Total	262	(100.00)	148	(100.00)	0.482	0.416	0.266
		AA	63	(24.05)	43	(29.05)			
		GA	133	(50.76)	73	(49.32)			
		GG	66	(25.19)	32	(21.62)			
		A	259	(49.43)	159	(53.72)	0.238		
		G	265	(50.57)	137	(46.28)			
IL1RAP	rs9290936	Total	263	(100.00)	150	(100.00)	0.887	0.637	0.994
		GG	83	(31.56)	44	(29.33)			
		GT	138	(52.47)	82	(54.67)			
		TT	42	(15.97)	24	(16.00)			
		G	304	(57.79)	170	(56.67)	0.753		
		T	222	(42.21)	130	(43.33)			
	rs10513854	Total	254	(100.00)	130	(100.00)	0.742	1.000	0.537
		TT	5	(1.97)	3	(2.31)			
		TC	65	(25.59)	29	(22.31)			
		CC	184	(72.44)	98	(75.38)			
		T	75	(14.76)	35	(13.46)	0.626		
		C	433	(85.24)	225	(86.54)			

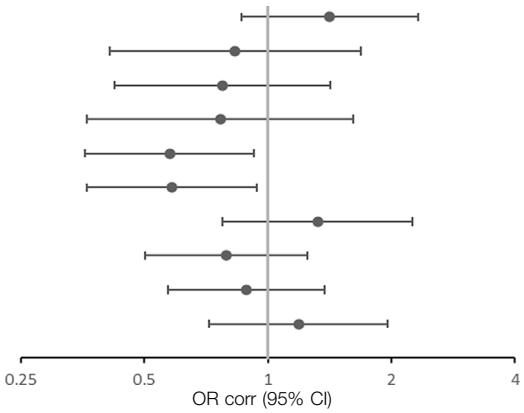
If the total number of cases is not 265 or the number of controls is not 135, it is due to missing data. p-*values < 0.05 are in bold.*

[a]*DOM, means AA vs (Aa + aa).*

[b]*REC, means (AA+ Aa) vs aa; "A" is the major allele and "a" is the minor allele.*

TABLE 4 | Association of gene SNPs with asthma of children susceptibility.

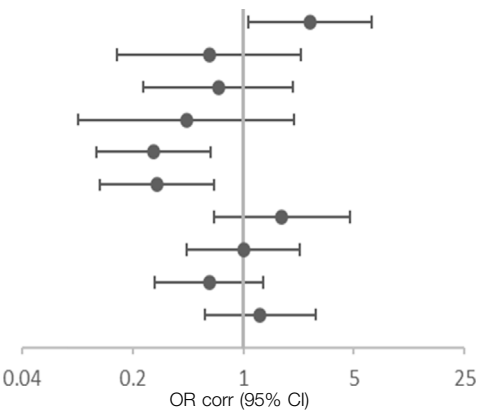
SNP	OR (95%CI)	OR corr[a]
rs4742170 DOM[b]	1.406 (0.859–2.300)	1.416 (0.863–2.325)
rs2381416 REC[c]	0.858 (0.427–1.724)	0.833 (0.413–1.682)
rs928413 REC[c]	0.798 (0.438–1.454)	0.775 (0.424–1.418)
rs992969 REC[c]	0.780 (0.371–1.641)	0.766 (0.363–1.616)
rs10208293 DOM[b]	**0.544 (0.341–0.868)**	**0.577 (0.359–0.927)**
rs13424006 DOM[b]	**0.552 (0.345–0.882)**	**0.584 (0.362–0.941)**
rs1420101 REC[c]	1.321 (0.777–2.245)	1.323 (0.776–2.254)
rs4142132 REC[c]	0.773 (0.491–1.217)	0.792 (0.502–1.251)
rs9290936 DOM[b]	0.900 (0.581–1.394)	0.887 (0.572–1.378)
rs10513854 DOM[b]	1.165 (0.718–1.892)	1.188 (0.719–1.962)



[a]*OR, corr: the* p *value after adjusting age and gender as covariates; OR: Odds ratio (reference group designated with an OR, of 1.0).* p*-values < 0.05 are in bold.*
[b]*DOM, means AA vs (Aa + aa).*
[c]*REC, means (AA+ Aa) vs aa; "A" is the major allele and "a" is the minor allele.*

TABLE 5 | Association of gene SNPs with risk of high FeNO at baseline.

SNP	OR (95%CI)	Adjusted OR (95% CI)[a]
rs4742170 DOM[b]	**2.583(1.063–6.277)**	**2.629(1.073–6.445)**
rs2381416 REC[c]	0.545 (0.145–2.052)	0.609 (0.159–2.327)
rs928413 REC[c]	0.618 (0.212–1.797)	0.695 (0.235–2.059)
rs992969 REC[c]	0.404 (0.085–1.920)	0.439 (0.091–2.109)
rs10208293 DOM[b]	**0.304(0.136–0.680)**	**0.271(0.118–0.621)**
rs13424006 DOM[b]	**0.317(0.142–0.708)**	**0.286(0.125–0.652)**
rs1420101 REC[c]	1.813 (0.681–4.827)	1.756 (0.654–4.717)
rs4142132 REC[c]	1.093 (0.487–2.452)	1.000 (0.439–2.275)
rs9290936 DOM[b]	0.604 (0.276–1.326)	0.608 (0.275–1.341)
rs10513854 DOM[b]	1.156 (0.526–2.540)	1.281 (0.573–2.865)



[a]*OR, corr: the* p *value after adjusting age and gender as covariates; OR: Odds ratio (reference group designated with an OR, of 1.0).* p*-values < 0.05 are in bold.*
[b]*DOM, means AA vs (Aa + aa).*
[c]*REC, means (AA+ Aa) vs aa; "A" is the major allele and "a" is the minor allele.*

Association of IL33-IL1RL1-IL1RAP Complex Snps With Inflammation-Type in Asthmatic Children

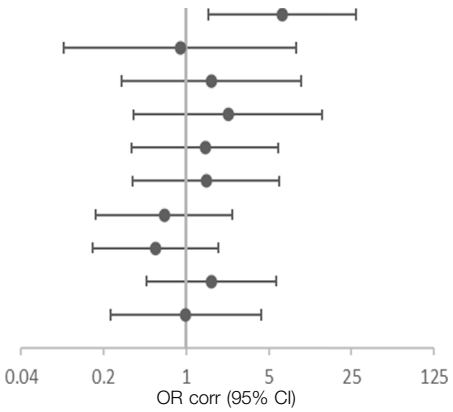
Asthmatic children were divided into FeNO increased group (FeNO ≥25 ppb) and FeNO normal group (FeNO<25 ppb). The allele and genotype frequencies of the 10 SNPs in these two groups is presented in **Supplementary eTable S1** in the Supplement. **Table 5** shows the association between genotypes and increased risk of FeNO in childhood asthma. Children with a CT or TT genotype of rs4742170 in the IL33 gene had a higher risk of FeNO elevation (corrected OR = 2.629, 95%CI: 1.073–6.445, $p = 0.035$). Children with a GA or AA genotype of rs10208293 in IL1RL1 had a lower risk of increased FeNO (corrected OR = 0.271, 95%CI: 0.118–0.621, $p = 0.002$), and both TC and CC genotypes of rs13424006 had a lower risk of increased FeNO in children with asthma (corrected OR = 0.286, 95%CI: 0.125–0.652, $p = 0.003$). However, when asthmatic children were divided into a PBEC increased group (PBEC >300 /μL) and a PBEC normal group (PBEC ≤300 /μL), there was no statistically significant difference in allele or genotype frequencies of these groups (**Supplementary eTable S2** in Supplement).

Association of IL33-IL1RL1-IL1RAP Complex SNPs With the Efficacy of ICS

The efficacy of ICS was evaluated by comparing indicators, including PBEC, FeNO, FEV1/FVC, MMEF/pre, and PD20, before and after treatment with ICS for 3 months.

TABLE 6 | Association of gene SNPs with risk of no response to ICS in terms of FeNO in asthmatic children after treatment with ICS for 3 months.

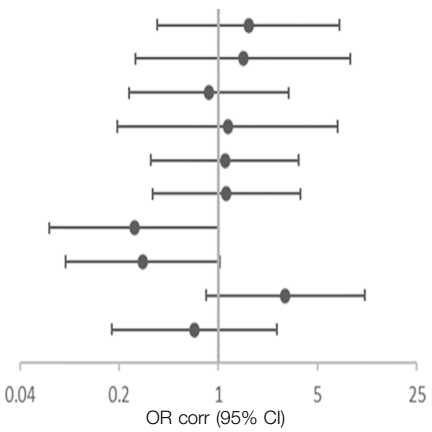
SNP	OR (95%CI)	Adjusted OR (95% CI)[a]
rs4742170 DOM[b]	6.771 (1.677–27.338)	**6.510(1.540–27.527)**
rs2381416 REC[c]	1.182 (0.126–11.102)	0.889 (0.092–8.570)
rs928413 REC[c]	1.909 (0.341–10.701)	1.634 (0.284–9.400)
rs992969 REC[c]	3.045 (0.497–18.669)	2.253 (0.360–14.109)
rs10208293 DOM[b]	1.132 (0.280–4.575)	1.437 (0.344–5.997)
rs13424006 DOM[b]	1.154 (0.285–4.666)	1.472 (0.353–6.144)
rs1420101 REC[c]	0.614 (0.166–2.272)	0.647 (0.170–2.456)
rs4142132 REC[c]	0.467 (0.140–1.561)	0.549 (0.160–1.879)
rs9290936 DOM[b]	1.594 (0.465–5.459)	1.636 (0.464–5.770)
rs10513854 DOM[b]	1.154 (0.285–4.666)	0.989 (0.229–4.276)



[a]*OR, corr: the* p *value after adjusting age and gender as covariates; OR: Odds ratio (reference group designated with an OR, of 1.0).* p*-values < 0.05 are in bold.*
[b]*DOM, means AA vs (Aa + aa).*
[c]*REC, means (AA+ Aa) vs aa; "A" is the major allele and "a" is the minor allele.*

TABLE 7 | Association of gene SNPs with risk of no response to ICS in terms of PBEC in asthmatic children after treatment with ICS for 3 months.

SNP	OR (95%CI)	Adjusted OR (95% CI)[c]
rs4742170 DOM[a]	1.784 (0.420–7.574)	1.634 (0.373–7.154)
rs2381416 REC[b]	1.806 (0.325–10.039)	1.490 (0.262–8.454)
rs928413 REC[b]	0.947 (0.268–3.351)	0.854 (0.234–3.118)
rs992969 REC[b]	1.459 (0.249–8.554)	1.160 (0.193–6.959)
rs10208293 DOM[a]	1.273 (0.403–4.021)	1.109 (0.335–3.673)
rs13424006 DOM[a]	1.312 (0.416–4.141)	1.137 (0.343–3.766)
rs1420101 REC[b]	**0.231(0.059–0.902)**	0.255 (0.064–1.011)
rs4142132 RECb	**0.260(0.076–0.891)**	0.295 (0.084–1.028)
rs9290936 DOM[a]	3.241 (0.940–11.168)	2.956 (0.817–10.700)
rs10513854 DOM[a]	0.504 (0.140–1.818)	0.676 (0.177–2.588)



[a]*DOM, means AA vs (Aa + aa).*
[b]*REC, means (AA+ Aa) vs aa; "A" is the major allele and "a" is the minor allele.*
[c]*OR, corr: the* p *value after adjusting age and gender as covariates; OR: Odds ratio (reference group designated with an OR, of 1.0).* p*-values < 0.05 are in bold.*

The risk of no improvement in FeNO in children with CT and TT genotypes of rs4742170 in the IL33 gene, was significantly higher than those children with CC genotypes (corrected OR = 6.510, 95%CI: 1.540–27.527, p = 0.011), suggesting that eosinophilic airway inflammation in children with a T allele at rs4742170 would respond poorly to ICS therapy (**Table 6**). The risk of no improvement in PBEC after treatment for children with a C allele at rs1420101 in IL1RL1 was lower than the other two genotypes (OR = 0.231, 95%CI: 0.059–0.902, p = 0.035), suggesting that the TT phenotype of rs1420101 indicated a higher risk of poor response to ICS. Children with a G allele at rs4142132 had a lower risk of no improvement in PBEC after treatment (OR = 0.260, 95%CI: 0.076–0.891, p = 0.032), suggesting that children with an AA genotype of rs4142132 had a higher risk of no improvement in PBEC after treatment (**Table 7**).

TABLE 8 | Association of the IL33 rs4742170 genotype with no improvement in PD20 after ICS treatment for 3 months.

Genotype	OR	95%CI	*p*-value
CC vs (CT + TT)	8.679	(1.029, 73.215)	0.047
(CC + CT) vs TT	4.183	(1.422, 12.304)	0.009

p*-values < 0.05 are in bold.*

The analysis of 10 SNPs and lung function indicators showed that the risk of no improvement in PD20 was significantly higher in children with CT and TT genotypes of rs4742170 in IL33 after

treatment (OR = 8.679, 95%CI: 1.029–73.215, $p = 0.047$), suggesting that asthmatic children with a T allele at rs4742170 had a significantly increased risk of poor response to ICS treatment for airway hyperresponsiveness (**Table 8**). However, similar associations were not found for FEV1/FVC or MMEF/pre.

DISCUSSION

As is well known, genetic factors play an important role in susceptibility to asthma. However, studies that focus on the association between genetic variations and asthma susceptibility in Chinese children are still limited. To the best of our knowledge, this is the first study that confirmed that SNPs of IL33/IL1RL1/IL1RAP are associated with Han Chinese childhood asthma. This study provides evidence for the association of asthma susceptibility, eosinophilic airway inflammation, and efficacy of ICS with distinct IL33/IL1RL1 ligand-receptor complex polymorphisms.

SNPs in IL33 previously associated with asthma susceptibility in European populations were not significantly associated with asthma susceptibility in Han children in our study. But we found that children with a T allele at rs4742170 in IL33 had a greater risk of increased levels of the eosinophilic airway inflammation marker FeNO (OR = 2.583, 95%CI: 1.063–6.277, $p = 0.036$), and a worse response to ICS (no improvement in FeNO after ICS treatment: OR = 6.771, 95%CI: 1.677–27.338, $p = 0.007$, no improvement in PD20 after ICS treatment: OR = 4.183, 95% CI: 1.422–12.304, $p = 0.009$). rs4742170 is located on the second intron of the IL33 gene, 27 KB distant from the IL33 promoter (Rosenbloom et al., 2013). The glucocorticoid receptor (GR) is expressed in lung epithelial cells. GR is activated by glucocorticoids and transferred to the nucleus to regulate the expression level of genes involved in stress and steroid responses in inflammation, by binding to nuclear glucocorticoid response elements (Surjit et al., 2011; Gorbacheva et al., 2018). The therapeutic effect of ICS is mainly to inhibit the expression of inflammatory genes through various DNA-binding mechanisms. At present, the specific GR-DNA of GR target genes is unclear. The five prime ends, the first intron and the second intron regions of the IL33 gene have some typical characteristics of regulatory elements (Gorbacheva et al., 2018).

Once the C allele at the rs4742170 site in lung cancer cells mutated to become a T allele, it was observed that the GR binding site was destroyed and the binding force between the GR and the IL33 gene was reduced (Gorbacheva et al., 2018). The activity of the IL33 promoter was enhanced, resulting in increased expression of IL33. The variation of rs4742170 may explain some of the pathophysiological mechanisms of asthma with ICS non-response.

IL1RL1 encodes three protein subtypes:1) IL1RL1-a (also known as soluble ST2), which can be detected in serum; 2) Transmembrane receptor protein IL1RL1-b (S2TL); and 3) IL1RL1-c (ST2V) (Takatori et al., 2018). IL1RL1-a, IL1RL1-b, and IL1RL1-c are all expressed in the lung. IL33 binds with IL1RL1-b, which is widely expressed on the membrane of type 2 immune cells, and then combines with IL1RAP to form a ternary complex, which activates immune cells to release type 2 cytokines such as IL4, IL5, and IL13 to promote airway inflammation. IL1RL1-a is considered to be a decoy receptor that blocks the IL33-IL1RL1 pathway by binding free IL33 and thus downregulates its regulatory ability on immune cells (Hayakawa et al., 2007).

The concentration of IL1RL1-a is greatly influenced by genetic factors. The heritability of IL1RL1-a was found to be 0.45, which means that 45% of the clinically unexplained differences in IL1RL1-a levels are due to genetic factors. In addition, no correlation was found between environmental factors and IL1RL1-a levels ($p = 0.25$) (Ho et al., 2013). The GWAS study reported in 2018 included some loci of the IL1RL1 gene region in the list of asthma highly related SNPs in Europeans (El-Husseini et al., 2020).

rs10208293 and rs13424006 are located in the 10th intron of IL1RL1. Two European birth cohorts and one paternity cohort study showed that these two sites were associated with late-onset wheezing in children (OR = 0.74, 95%CI: 0.62–0.87, $p < 0.01$) (Savenije et al., 2014). In our study, these two sites were associated with susceptibility to childhood asthma and eosinophilic airway inflammation in Han Chinese children. We speculate that these two SNPs may be involved in the regulation of IL1RL1-a.

rs1420101, located in the fifth intron of IL1RL1, has been widely validated in multiple GWAS studies (Demenais et al., 2018; Dahlin et al., 2019). A Dutch cohort and three European birth cohorts showed that SNPs in the IL1RL1 gene region were significantly correlated with gene methylation, and effectively regulated serum IL1RL1-a levels (Dijk et al., 2018). Studies have found that the T allele of rs1420101 in Europeans is not only a risk allele for asthma, but is also associated with higher eosinophil counts and higher IgE levels (Gudbjartsson et al., 2009; Grotenboer et al., 2013). Further studies have found that the T allele at rs1420101 was associated with lower levels of DNA methylation, low expression of IL1RL1 mRNA in airway epithelial cells and lung tissue, and reduced levels of IL1RL1-a in serum and alveolar lavage fluid (Akhabir and Sandford, 2010; Gordon et al., 2016; Dijk et al., 2018). The decrease inIL1RL1-a levels resulted in the upregulation of IL33 and increased IL1RL1-b binding force and the enhancement of type 2 inflammatory response. In our study, rs1420101 was not significantly correlated with asthma susceptibility or PBEC in Han children in the Hunan region. However, it was observed that children with asthma who had the TT genotype of rs1420101 or AA genotype ofrs4142132 in the first intron of IL1RL1, responded better to ICS treatment in PBEC. Therefore, we speculate that ICS can effectively control asthma with rs1420101 and rs4142132 mutations, through inhibiting the expression of upstream IL33.

There are few studies examining SNPs in the IL1RAP gene. IL1RAP is an essential co-receptor in the interleukin-1 family (Fields et al., 2019). Recent studies have found that, besides the membrane receptor, L1RAcP encodes a kind of soluble protein that inhibits the inflammatory response through enhancing the binding of Il1RL1-a with IL33 (Palmer et al., 2008, 33). rs9290936 in the first intron of the IL1RAP gene was associated with persistent wheezing in children in Dutch and UK populations (OR = 0.72, 95%CI: 0.61–0.85, $p < 0.01$), and rs10513854 in the second intron was associated with intermediate-onset wheezing (OR = 0.67, 95%CI: 0.55-0.83, $p < 0.01$) (Savenije et al., 2014). However, we did not find any association between these two SNPs and childhood asthma in Han children.

CONCLUSION

In conclusion, we found significant associations between IL1RL1 rs10208293 and rs13424006 polymorphisms and the risk of childhood asthma. In addition, we found that they are associated with type 2 airway inflammation, which is like rs4742170 in IL33. When it comes to the efficacy of ICS, rs4742170 in IL33, rs1420101, and rs4142132 in IL1RL1 polymorphisms were significantly associated in our study. No association between asthma and either SNP in IL1RAP was observed. Our results indicated that rs10208293 and rs13424006 could be asthma markers in children, while rs4742170, rs1420101, and rs4142132 might be used to predict the ICS response in childhood asthma.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Xiangya Hospital, Central South University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

Conceptualization: MW, XZ, JH, XH. Recruitment and specimens collection: JH, MW. Analysis and manuscript drafting: MW. Critical revision of the manuscript: XZ. All authors read and approved the manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (81873851, 82170019) and the Hunan Provincial Natural Science Foundation (2018JJ6141).

ACKNOWLEDGMENTS

We thank Wang Peng, Bairu Wu, and Yating Chen for writing assistance and Jian Kuang for data collection assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.759542/full#supplementary-material

REFERENCES

Akhabir, L., and Sandford, A. (2010). Genetics of Interleukin 1 Receptor-like 1 in Immune and Inflammatory Diseases. *Cg* 11, 591–606. doi:10.2174/138920210793360907

Association Editorial Board Chinese Journal of Pediatrics (2016), Zhonghua Er Ke Za Zhi. Guideline for the Diagnosis and Optimal Management of Asthma in Children. *Chin. J. Pediatr.*, 54, 167–181. doi:10.3760/cma.j.issn.0578-1310.2016.03.003

Cayrol, C., and Girard, J.-P. (2018). Interleukin-33 (IL-33): A Nuclear Cytokine from the IL-1 Family. *Immunol. Rev.* 281, 154–168. doi:10.1111/imr.12619

Chan, B. C. L., Lam, C. W. K., Tam, L.-S., and Wong, C. K. (2019). IL33: Roles in Allergic Inflammation and Therapeutic Perspectives. *Front. Immunol.* 10, 364. doi:10.3389/fimmu.2019.00364

Chung, K. F., Wenzel, S. E., Brozek, J. L., Bush, A., Castro, M., Sterk, P. J., et al. (2014). International ERS/ATS Guidelines on Definition, Evaluation and Treatment of Severe Asthma. *Eur. Respir. J.* 43, 343–373. doi:10.1183/09031936.00202013

Dahlin, A., Sordillo, J. E., Ziniti, J., Iribarren, C., Lu, M., Weiss, S. T., et al. (2019). Large-scale, Multiethnic Genome-wide Association Study Identifies Novel Loci Contributing to Asthma Susceptibility in Adults. *J. Allergy Clin. Immunol.* 143, 1633–1635. doi:10.1016/j.jaci.2018.11.037

Demenais, F., Margaritte-Jeannin, P., Barnes, K. C., Cookson, W. O. C., Altmüller, J., Ang, W., et al. (2018). Multiancestry Association Study Identifies New Asthma Risk Loci that Colocalize with Immune-Cell Enhancer marks. *Nat. Genet.* 50, 42–53. doi:10.1038/s41588-017-0014-7

di Palmo, E., Cantarelli, E., Catelli, A., Ricci, G., Gallucci, M., Miniaci, A., et al. (2021). The Predictive Role of Biomarkers and Genetics in Childhood Asthma Exacerbations. *Ijms* 22, 4651. doi:10.3390/ijms22094651

Dijk, F. N., Xu, C., Melén, E., Carsin, A.-E., Kumar, A., Nolte, I. M., et al. (2018). Genetic Regulation of IL1RL1 Methylation and IL1RL1-a Protein Levels in Asthma. *Eur. Respir. J.* 51, 1701377. doi:10.1183/13993003.01377-2017

El-Husseini, Z. W., Gosens, R., Dekker, F., and Koppelman, G. H. (2020). The Genetics of Asthma and the Promise of Genomics-Guided Drug Target Discovery. *Lancet Respir. Med.* 8, 1045–1056. doi:10.1016/S2213-2600(20)30363-5

Fields, J. K., Günther, S., and Sundberg, E. J. (2019). Structural Basis of IL-1 Family Cytokine Signaling. *Front. Immunol.* 10, 1412. doi:10.3389/fimmu.2019.01412

Gabryelska, A., Kuna, P., Antczak, A., Białasiewicz, P., and Panek, M. (2019). IL-33 Mediated Inflammation in Chronic Respiratory Diseases-Understanding the Role of the Member of IL-1 Superfamily. *Front. Immunol.* 10, 692. doi:10.3389/fimmu.2019.00692

Gasiuniene, E., Janulaityte, I., Zemeckiene, Z., Barkauskiene, D., and Sitkauskiene, B. (2019). Elevated Levels of Interleukin-33 Are Associated with Allergic and Eosinophilic Asthma. *Scand. J. Immunol.* 89, e12724. doi:10.1111/sji.12724

GINA (2015). *Global Strategy for Asthma Management and Prevention*. Fontana, WI, USA: GINA.

GINA (2020). *Global Strategy for Asthma Management and Prevention*. Fontana, WI, USA: GINA.

Gorbacheva, A., Kuprash, D., and Mitkin, N. (2018). Glucocorticoid Receptor Binding Inhibits an Intronic IL33 Enhancer and Is Disrupted by Rs4742170 (T) Allele Associated with Specific Wheezing Phenotype in Early Childhood. *IJMS* 19, 3956. doi:10.3390/ijms19123956

Gordon, E. D., Palandra, J., Wesolowska-Andersen, A., Ringel, L., Rios, C. L., Lachowicz-Scroggins, M. E., et al. (2016). IL1RL1 Asthma Risk Variants Regulate Airway Type 2 Inflammation. *JCI Insight* 1, e87871. doi:10.1172/jci.insight.87871

Grotenboer, N. S., Ketelaar, M. E., Koppelman, G. H., and Nawijn, M. C. (2013). Decoding Asthma: Translating Genetic Variation in IL33 and IL1RL1 into Disease Pathophysiology. *J. Allergy Clin. Immunol.* 131, 856–865. doi:10.1016/j.jaci.2012.11.028

Gudbjartsson, D. F., Bjornsdottir, U. S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G. M., et al. (2009). Sequence Variants Affecting Eosinophil Numbers Associate with Asthma and Myocardial Infarction. *Nat. Genet.* 41, 342–347. doi:10.1038/ng.323

Hayakawa, H., Hayakawa, M., Kume, A., and Tominaga, S.-i. (2007). Soluble ST2 Blocks Interleukin-33 Signaling in Allergic Airway Inflammation. *J. Biol. Chem.* 282, 26369–26380. doi:10.1074/jbc.M704916200

Ho, J. E., Chen, W.-Y., Chen, M.-H., Larson, M. G., McCabe, E. L., Cheng, S., et al. (2013). Common Genetic Variation at the IL1RL1 Locus Regulates IL-33/ST2 Signaling. *J. Clin. Invest.* 123, 4208–4218. doi:10.1172/JCI67119

Johansson, K., and McSorley, H. J. (2019). Interleukin-33 in the Developing Lung-Roles in Asthma and Infection. *Pediatr. Allergy Immunol.*, 13040. doi:10.1111/pai.13040

Liew, F. Y., Girard, J.-P., and Turnquist, H. R. (2016). Interleukin-33 in Health and Disease. *Nat. Rev. Immunol.* 16, 676–689. doi:10.1038/nri.2016.95

Martin, M. U. (2013). Special Aspects of Interleukin-33 and the IL-33 Receptor Complex. *Semin. Immunol.* 25, 449–457. doi:10.1016/j.smim.2013.10.006

Molofsky, A. B., Savage, A. K., and Locksley, R. M. (2015). Interleukin-33 in Tissue Homeostasis, Injury, and Inflammation. *Immunity* 42, 1005–1019. doi:10.1016/j.immuni.2015.06.006

Morales, E., and Duffy, D. (2019). Genetics and Gene-Environment Interactions in Childhood and Adult Onset Asthma. *Front. Pediatr.* 7, 499. doi:10.3389/fped.2019.00499

Murdaca, G., Greco, M., Tonacci, A., Negrini, S., Borro, M., Puppo, F., et al. (2019). IL-33/IL-31 Axis in Immune-Mediated and Allergic Diseases. *Ijms* 20, 5856. doi:10.3390/ijms20235856

Palmer, G., Lipsky, B. P., Smithgall, M. D., Meininger, D., Siu, S., Talabot-Ayer, D., et al. (2008). The IL-1 Receptor Accessory Protein (AcP) Is Required for IL-33 Signaling and Soluble AcP Enhances the Ability of Soluble ST2 to Inhibit IL-33. *Cytokine* 42, 358–364. doi:10.1016/j.cyto.2008.03.008

Papi, A., Brightling, C., Pedersen, S. E., and Reddel, H. K. (2018). Asthma. *The Lancet* 391, 783–800. doi:10.1016/S0140-6736(17)33311-1

Pulmonary Function Cooperation Group of Respiratory Group Pediatrics Society of Chinese Medical Association Editorial Board of the Chinese Journal of Clinical Applied Pediatrics (2017). Series Guide for Indicators of Lung Function and Non-traumatic Airway Inflammation in Children (7): Monitoring of Exhaled Nitric Oxide. *Chin. J. Pract. Pediatr.* 32, 1622–1627.

Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., et al. (2013). ENCODE Data in the UCSC Genome Browser: Year 5 Update. *Nucleic Acids Res.* 41, D56–D63. doi:10.1093/nar/gks1172

Savenije, O. E., Mahachie John, J. M., Granell, R., Kerkhof, M., Dijk, F. N., de Jongste, J. C., et al. (2014). Association of IL33-IL-1 Receptor-like 1 (IL1RL1) Pathway Polymorphisms with Wheezing Phenotypes and Asthma in Childhood. *J. Allergy Clin. Immunol.* 134, 170–177. doi:10.1016/j.jaci.2013.12.1080

Schoettler, N., Rodríguez, E., Weidinger, S., and Ober, C. (2019). Advances in Asthma and Allergic Disease Genetics: Is Bigger Always Better? *J. Allergy Clin. Immunol.* 144, 1495–1506. doi:10.1016/j.jaci.2019.10.023

Surjit, M., Ganti, K. P., Mukherji, A., Ye, T., Hua, G., Metzger, D., et al. (2011). Widespread Negative Response Elements Mediate Direct Repression by Agonist-Liganded Glucocorticoid Receptor. *Cell* 145, 224–241. doi:10.1016/j.cell.2011.03.027

Takatori, H., Makita, S., Ito, T., Matsuki, A., and Nakajima, H. (2018). Regulatory Mechanisms of IL-33-ST2-Mediated Allergic Inflammation. *Front. Immunol.* 9, 2004. doi:10.3389/fimmu.2018.02004

Wang, B., Chen, H., Chan, Y. L., Wang, G., and Oliver, B. G. (2020). Why Do Intrauterine Exposure to Air Pollution and Cigarette Smoke Increase the Risk of Asthma? *Front. Cel Dev. Biol.* 8, 38. doi:10.3389/fcell.2020.00038

Zheng, R., Chen, Y., Shi, J., Wang, K., Huang, X., Sun, Y., et al. (2020). Combinatorial IL-17RB, ST2, and TSLPR Signaling in Dendritic Cells of Patients with Allergic Rhinitis. *Front. Cel Dev. Biol.* 8, 207. doi:10.3389/fcell.2020.00207

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Wu, Zheng, Huang and Hu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 16 December 2021
doi: 10.3389/fcell.2021.801113

Gene-Based Testing of Interactions Using XGBoost in Genome-Wide Association Studies

Yingjie Guo[1,2], Chenxi Wu[3], Zhian Yuan[4], Yansu Wang[1,2], Zhen Liang[5], Yang Wang[2], Yi Zhang[6] and Lei Xu[2]**

[1]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [2]School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China, [3]Department of Mathematics, University of Wisconsin-Madison, Madison, WI, United States, [4]Research Institute of Big Data Science and Industry, Shanxi University, Taiyuan, China, [5]School of Life Science, Shanxi University, Taiyuan, China, [6]Beidahuang Industry Group General Hospital, Harbin, China

Among the myriad of statistical methods that identify gene–gene interactions in the realm of qualitative genome-wide association studies, gene-based interactions are not only powerful statistically, but also they are interpretable biologically. However, they have limited statistical detection by making assumptions on the association between traits and single nucleotide polymorphisms. Thus, a gene-based method (GGInt-XGBoost) originated from XGBoost is proposed in this article. Assuming that log odds ratio of disease traits satisfies the additive relationship if the pair of genes had no interactions, the difference in error between the XGBoost model with and without additive constraint could indicate gene–gene interaction; we then used a permutation-based statistical test to assess this difference and to provide a statistical *p*-value to represent the significance of the interaction. Experimental results on both simulation and real data showed that our approach had superior performance than previous experiments to detect gene–gene interactions.

Keywords: genome-wide association studies, gene–gene interactions, XGBoost, additive model, gene-based testing

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Quan Zou,
University of Electronic Science and Technology of China, China

****Correspondence:***
Yi Zhang
y_zhang1024@126.com
Lei Xu
csleixu@szpt.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *24 October 2021*
Accepted: *23 November 2021*
Published: *16 December 2021*

Citation:
Guo Y, Wu C, Yuan Z, Wang Y, Liang Z, Wang Y, Zhang Y and Xu L (2021) Gene-Based Testing of Interactions Using XGBoost in Genome-Wide Association Studies. Front. Cell Dev. Biol. 9:801113. doi: 10.3389/fcell.2021.801113

1 INTRODUCTION

Genome-wide association study (GWAS) is a collection of successful methods for identifying genetic loci associated with complex traits. More than 71,000 specific single nucleotide polymorphisms (SNPs) associated with diseases or traits have been identified (Hindorff et al., 2009; Yang et al., 2015; Liu et al., 2018a; Guo et al., 2018; Buniello et al., 2019; Loos, 2020; Lyu et al., 2020; Hu et al., 2021). Previous GWAS schemes relied mainly on a single locus model that verified the independent association of individual markers to particular phenotypes. Despite the successful recognition of many regions of disease susceptibility, most SNPs captured by this kind of method may have a small effect size that does not explain the heritability of complex traits fully. It is believed that genetic interactions that are engaged significantly in the genetic basis of complex traits and diseases (Cordell, 2009; Moore et al., 2010; Liu et al., 2018b; He et al., 2020; Luo et al., 2020; Shao and Liu, 2021) may be a potential solution to the problem of "missing heritability" (Manolio et al., 2009; Fang et al., 2019; Young, 2019). The solution may be partial, but it could enlighten the construction of new topologies for gene pathways.

Genetic interaction was first studied at the SNP level, and SNP–SNP interactions (i.e., epistasis) were detected by applying several methods (Li et al., 2015a; Ritchie and Van Steen, 2018), such as statistics based on entropy (Dong et al., 2008), logistic regression (Lin et al., 2016), and odds ratio (Emily, 2012); other techniques include multifactor dimensionality reduction (MDR) (Ritchie et al., 2003), BOOST (Wan et al., 2010), RRIntCC (Zhang et al., 2019), GenEpi (Chang et al., 2020), and some accelerate method (Nobre et al., 2021). One of the general challenges encountered by these SNP-based approaches is the statistical weakness of the higher-order or pairwise tests that result from massive multiple testing corrections over all the groups or pairs of SNPs. Instead, we investigated every possible SNP from two genes in single, gene-based interaction detection.

The success of gene-based approaches in marginal association studies of GWAS could extend to the analysis of gene–gene interactions (GGIs) (Emily, 2018; Emily et al., 2020). This approach has several potential advantages. First, it typically has far fewer genes than SNPs, reducing the number of pairwise tests drastically. For example, $\sim 2 \times 10^8$ tests are required to detect genetic interactions in pairs of 20,000 genes. However, over 5×10^{12} tests are required for 3 million SNPs in a marker-based interaction. Second, because a gene contains more information than a single SNP and genes interact diversely, gene-based methods are more powerful statistically, which applies to gene-based studies on the main effects as well (Liu et al., 2010; Li et al., 2011; Jiang et al., 2017; Wang et al., 2020; Wang et al., 2021). Additionally, biological prior knowledge (e.g., information about the known association of genes within protein–protein interactions (PPIs) or pathways) can be introduced easily. Finally, gene-based results are characterized by having better interpretability and important biological consequences.

Peng et al. (Peng et al., 2010) discovered a canonical correlation of a pair of genes in a case group and in a control group by applying a canonical correlation analysis–based U statistic (CCU), which measured the difference in the correlation of the gene pair. The difference then indicated the incidence of a GGI. In the analysis, however, only linear relationships were taken into consideration. Afterward, CCU was extended to kernelized CCU (KCCU) (Yuan et al., 2012; Larson et al., 2013), where a non-linear relationship was detected under the kernel. Recently, Emily (Emily, 2016) presented a method called AGGrGATOr that combined *p*-values interaction tests at the marker level to gauge how a pair of genes interacted, which was a strategy used by Ma et al. (Ma et al., 2013) earlier to detect interactions under quantitative phenotypes. Li et al. (Li et al., 2015b) proposed an entropy-based and nonparametric method called GBIGM.

At present, the new approach GGInt-XGBoost is proposed for identifying gene–gene interactions of complex phenotypes at the gene level in case-control studies by leveraging the eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), which is applied in co-expressed gene detection and to explore genetic associations in the field of bioinformatics (Jiang et al., 2013; Babajide Mustapha and Saeed, 2016; Liu et al., 2016; Liu and Jiang, 2016; Mrozek et al., 2016; Wei et al., 2017a; Wei et al., 2017b; Liu et al., 2017; Chen et al., 2018; Wei et al., 2018; Jiang et al., 2019; Liu et al., 2019; Yu et al., 2020a; Yu et al., 2020b; Lv et al., 2020; Li et al., 2021; Liu et al., 2021). A built-in mechanism of XGBoost is that one can impose constraints on the trained model to make it additive, which we assume characterizes the lack of interaction between two genes. Our method exhibited an outstanding performance for detecting the underlying gene–gene interactions at the gene level under various settings based on the experiments using a semi-empirical dataset. Its application using real datasets showed accurate identification of gene–gene interactions.

2 MATERIALS AND METHODS

In this section, we detailed the statistical workflow for GGInt-XGBoost. To evaluate the power to detect GGIs and type-I error, we present the different parameter settings for simulation studies based on empirical data. Then, we adopted a real-world rheumatoid arthritis dataset from the WTCCC (Wellcome Trust Case–Control Consortium) database to assess the performance of our method under a real situation.

2.1 GGInt-XGBoost

2.1.1 Preliminaries and Notation

Here, we take genes, a couple of SNPs, as the basic unit. Suppose that we have n random samples:

$$\left(G_{1,i}, G_{2,i}\right) \in \mathcal{R}^{p+q}, i = 1, 2, \ldots, n, \tag{1}$$

where

$$G_{1,i} = \left(g_{1,i,1}, g_{1,i,2}, \ldots, g_{1,i,p}\right), G_{2,i} = \left(g_{2,i,1}, g_{2,i,2}, \ldots, g_{2,i,q}\right),\ i = 1, 2, \ldots, n$$

and G_1 and G_2 represent two genes each with p and q SNPs, independently. In the case–control studies, $y_i \in \{0, 1\}$ is a categorical label, where 0 is a control subject and 1 is a case subject. $g_{k,i,j} \in \{0, 1, 2\}$ represents the copy number of the minor alleles of SNP j in the gene k for the sample i.

In this work, we created a statistic based on the XGBoost to quantify GGI intensity in order to see if there is a statistical interaction between two genes in a qualitative phenotype. To estimate the distribution of the statistic, we used a permutation resampling strategy. Our method was based on the assumption that if there was no interaction between two genes, adding a constraint to limit interactions between SNPs to only occurring in the same gene would not have a significant negative impact on XGBoost performance. The XGBoost's build-in mechanism for adding interaction constraints enables us to generate an additive model and use prior knowledge about the gene structure during model construction.

2.1.2 Definition of Total Additivity for Gene–Gene Interaction

We defined GGIs using the concept of additive models. Additive models were proposed by Friedman and Stuetzle (Friedman and

Stuetzle, 1981) and further developed and popularized by Stone (Stone, 1985), Hastie, and Tibshirani (Hastie and Tibshirani, 1990). Consider the regression problem where the feature lies in $\mathcal{R}^d$ and the objective function has a real value. Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_l$ be a disjoint partition of the index set $\{1, \ldots, d\}$ and denotes the elements of $\boldsymbol{s}_i$ to be $j_{i1}, \ldots, j_{id_i}$ and $\pi_i(x) = (x_{j_{i1}}, .., x_{j_{id_i}}) \in \mathcal{R}^{d_i}$. Now, real-valued function F on $\mathcal{R}^d$ is said to be additive for partition $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_l\}$ if there exists F_i: $\mathcal{R}^{d_i} \to \mathcal{R}$ such that

$$F(x) = \sum_{i=1}^{l} F_i(\pi_i(x)). \tag{2}$$

In our setting, the samples are elements in $\{0, 1, 2\}^{p+q} \subset \mathcal{R}^{p+q}$, and we let $\boldsymbol{s}_1 = \{1, 2, \ldots, p\}$ and $\boldsymbol{s}_2 = \{1, 2, \ldots, q\}$. We defined the absence of interaction between the two genes as the log odds ratio being additive with respect to the partition $\{\boldsymbol{s}_1, \boldsymbol{s}_2\}$. In other words, our null hypothesis is

$$H_0: \ \exists F_1, F_2 \text{ such that } P(y = 1 | G_1, G_2) = \frac{\exp(F_1(G_1) + F_2(G_2))}{1 + \exp(F_1(G_1) + F_2(G_2))}. \tag{3}$$

2.1.3 eXtreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) is a scalable machine-learning system for tree boosting, which researchers apply to bioactive molecular prediction (Babajide Mustapha and Saeed, 2016), protein submitochondrial localization prediction (Yu et al., 2020a), miRNA-disease association prediction (Chen et al., 2018), and in the bioinformatics field (Shao et al., 2021).

For a given dataset with n samples, $D = \{(x_i, y_i)\}$, $x_i \in \mathcal{R}^m$, $y_i \in \mathcal{R}$, and the XGBoost objective function is defined as:

$$obj(\theta) = \sum_i^n l\left(y_i, \hat{y}_i\right) + \sum_{t=1}^{T} \Omega(f_t), \tag{4}$$

where l is the loss function and Ω is the regularizer on the regression tree f_t, $\theta = (f_1, \ldots, f_T)$, and $\hat{y}_i = \sum f_t(x_i)$. The $t-th$ tree f_t was obtained iteratively by gradient boosting, that is,

$$f_t \approx \arg\min \sum_i \left(\partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) f_t(x_i) + \frac{1}{2} \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) f_t^2(x_i)\right) + \Omega(f_t). \tag{5}$$

In our setting,

$$l\left(y_i, \hat{y}_i\right) = \begin{cases} -\log\left(\dfrac{exp\left(\hat{y}_i\right)}{1 + exp\left(\hat{y}_i\right)}\right) & y = 1 \\ -\log\left(\dfrac{1}{1 + \exp\left(\hat{y}_i\right)}\right) & y = 0 \end{cases} \tag{6}$$

When running XGBoost, an essential step is to optimize its general parameters, booster parameters, and learning parameters.

2.1.4 XGBoost With the Additive Constraint

The base learner f_t used in XGBoost is a regression tree, and we considered features that appear in a path on the tree that starts at the root and ends at one of the leaves as features that interact with one another. XGBoost allows specification of feature interaction constraints in the form of lists of features where only the features in the same list are allowed to interact with one another. It is evident that when the lists are disjointed, and if every feature is included in one of the lists, the feature interaction constraint is equivalent to forcing each f_t to include only features in a single list, which implies that the regression model $\sum f_i$ must be additive concerning the partition specified by the lists^i. With the constraint [[0,1] (Hindorff et al., 2009; Liu et al., 2018a; Loos, 2020),], for example, the tree in **Figure 1A** violates the first constraint [0,1], thus so would not be in the boosting tree system, but the tree in **Figure 1B** complies with both the first and second constraints.

2.1.5 Illustration of the GGInt-XGBoost Workflow

Assume there are n samples in a case–control study for a pair of genes such that G_1 has p SNPs and G_2 has q SNPs. We can then apply XGBoost using the logistic regression loss function with and without constraints on additivity, to the dataset to estimate the performance in error using 10-fold cross-validation. We denote the error in the unconstrained model as err^0_{orig} and in the constrained model as err^0_{cons}. The improvement of the performance of the unconstrained model over the constrained model is $\Delta err^0 = \frac{err^0_{cons} - err^0_{orig}}{err^0_{orig}}$, which according to our assumption should be a statistic that characterizes the strength of interaction between these two genes. A positive Δerr^0 indicated that the unconstrained model performed better, and a larger positive Δerr^0 means there was a stronger interaction between the two genes.

To get a p-value, we needed to estimate the distribution of Δerr^0 under the null hypothesis. Here, we used a non-parametric strategy based on permutation: we shuffled the label y randomly m times, calculated Δerr using the exact same aforementioned procedure, and used the resulting empirical distribution as an estimate for the distribution of Δerr^0 under the null hypothesis. Let the result of these m permutations be $\Delta err^1, \ldots, \Delta err^m$, then an estimated p-value for the null hypothesis is

$$p = \frac{\left|\{i: \ \Delta err^i \geq \Delta err^0\}\right|}{m}. \tag{7}$$

For XGBoost, if we have n samples, K trees, a maximum depth of d per tree, and $||s||$ as the number of non-missing entries in the training data, the training time complexity is $O(Kd||s||logn)$. Prediction for a new sample takes $O(kd)$. We employed parallel programming to minimize the execution time of permutation resampling.

We summarized the process of GGInt-XGBoost in the algorithm below (Algorithm 1) and presented the overall workflow (**Figure 2**).

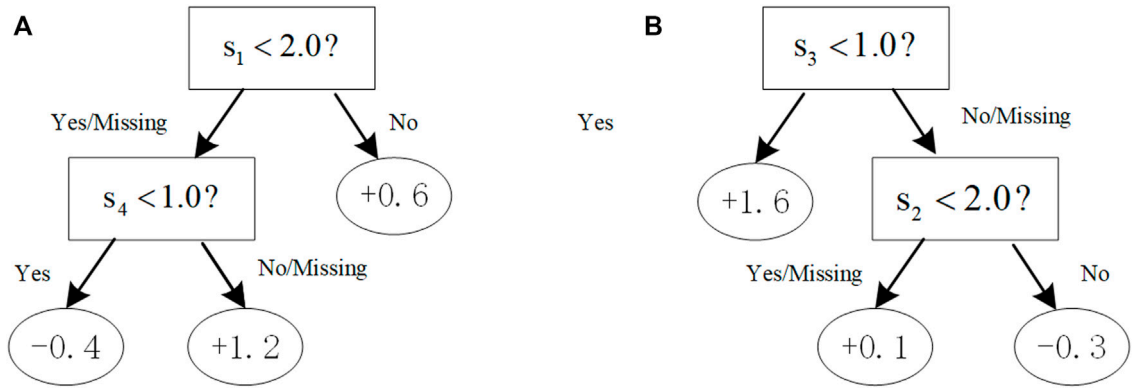


FIGURE 1 | Illustration of trees with the additive constraint. With the constraint [[0,1] (Hindorff et al., 2009; Liu et al., 2018a; Loos, 2020),], **(A)** a tree violates the first constraint (0,1) that would not be in the boosting tree system; **(B)** a tree complies with both the first and second constraints.

Algorithm 1: **GGInt-XGBoost**

Data: Genotype G_1, G_2, Phenotype y, permutation times m
Result: significance p-value for interaction between G_1, G_2
1 Apply unconstrained and constrained XGBoost with logistic loss for (G_1, G_2) to calculate err_{orig} and err_{cons} by 10-folds cross-validation;
2 Calculate the difference Δerr^0 between err_{orig} and err_{cons};
3 **for** $i = 1$ to m **do**
4 Randomly permute label y, and generate the new data set;
5 Repeat Steps 1 and 2;
6 **end**
7 Estimated p-value of Δerr^0 is the number of Δerr^i, $i = 1, \ldots, m$, larger than Δerr^0, divided by m.

2.2 Simulation Study

To assess the performance of GGInt-XGBoost to control type I error and to detect GGIs, we compared GGInt-XGBoost with KCCA (Larson et al., 2013), GBIGM (Li et al., 2015b), and AGGrEGATOr (Emily, 2016).

2.2.1 Simulation With Haplotype Data

gs2.0 (Li and Chen, 2008) is a semi-empirical simulation data generator that employs haplotype data as input and produces high-density SNP genotype data for qualitative samples. The generated dataset shares the same local linkage disequilibrium (LD) structure as that of human populations. We selected HapMap3 (a resident of Utah, the United States with Northern and Western European ancestry from https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html) to mimic the actual LD structure of the human population. The Central European (CEU) dataset with 90 haplotypes was used as the template haplotype data. In this research, we randomly picked one pair of gene loci (i.e., GNPDA2 from chromosome 4 and FAIM2 from chromosome 12). GNPDA2 had a much stronger LD pattern than FAIM2 did, and they were not correlated (**Figure 3**). By employing the genipe module (Lemieux Perreault et al., 2016), an imputation pipeline on the genome-scale with PLINK, IMPUTE 2, and SHAPEIT, chromosomes 4 and 12 were imputed. After imputing, six SNPs were obtained from GNPDA2, and seven SNPs were obtained from FAIM2 (**Supplementary Table S1**).

2.2.2 Disease Model

Here, we generated a disease model with two loci. A disease model represents the relationship between two loci that correspond to the disease. With various combinations of odds ratios (OR), sample sizes, and population prevalence, we generated different disease models. Using the jointly recessive–dominant model (RD model) as an example, for each locus let the genotype OR be $(1 + \theta)$ and the population prevalence of the disease be p (**Supplementary Table S2**).

$\Pr(D|g_i)$ indicates the probability of a sample being a case given the genotype combination of g_i and named the penetrance of g_i, and $\Pr(\bar{D}|g_i)$ denotes the probability of a sample being a control given the genotype combination of g_i. Then, the odds of disease are:

$$ODD_{g_i} = \frac{\Pr(D|g_i)}{\Pr(\bar{D}|g_i)} = \frac{\Pr(D|g_i)}{1 - \Pr(D|g_i)}. \quad (8)$$

The penetrance of genotype g_i can be calculated using:

$$\Pr(g_i) = \frac{ODD_{g_i}}{1 + ODD_{g_i}}. \quad (9)$$

The corresponding penetrance table is shown in **Supplementary Table S3**.

With a specific genotype OR $(1 + \theta)$ and a population prevalence p, the baseline value γ represents the disease odds with the two loci that do not have the disease alleles, and it can be calculated by applying **Eq. 10** with the terms from **Supplementary Table S3**.

$$p = Pr(D) = \sum \Pr(D|g_i) \times \Pr(g_i). \quad (10)$$

We used six integrated disease models in gs2.0, which included a recessive–dominant model, a dominant–dominant model, an XOR model, a threshold model, a multiplicative model, and a recessive–recessive model. We generated different datasets by various parameter settings, and we compared the performances of KCCU, GBIGM, and AGGrEGATOr with our method.

Evaluation of Type-I Error: The type-I error indicates the ability of a method to reject the null hypothesis when it is

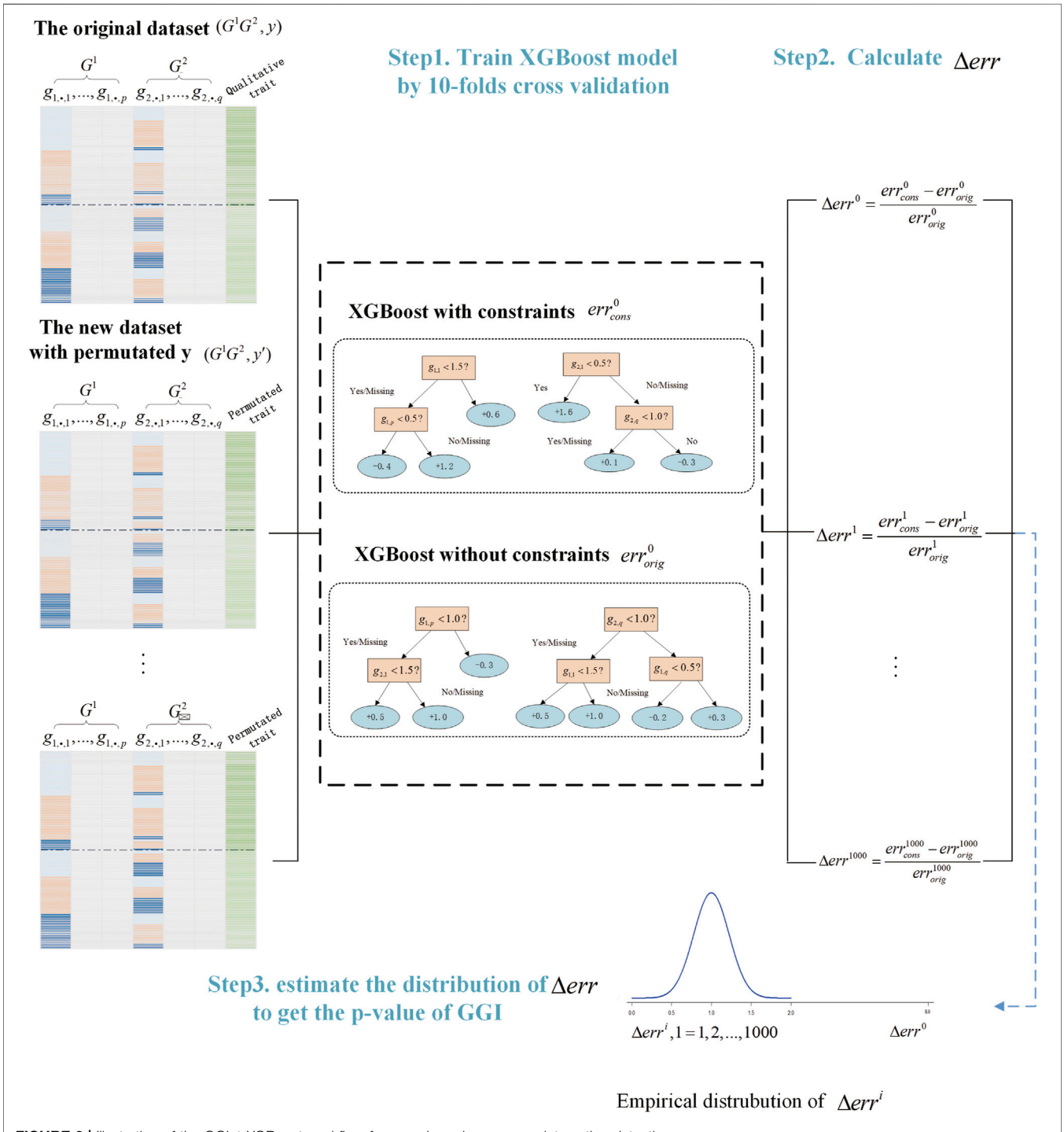


FIGURE 2 | Illustration of the GGInt-XGBoost workflow for gene-based gene–gene interaction detection.

true. In this study, the significance level α was set to 0.05. The simulation used in the model is shown in **Supplementary Table S4** and run 100 times with each sample size $n \in \{1k, 2k, 3k, 4k, 5k\}$ with the odds ratio set at 1.

Evaluation of Power of the Test: The power of a test indicates the probability that the method can reject the null hypothesis correctly when the alternative hypothesis is true. In this study, we generated 100 datasets for each parameter set under six disease models (**Supplementary Table S5**). The power under each parameter setting was expressed by the frequency with which the null hypothesis of the dataset was rejected correctly at the significance level of $\alpha = 0.05$. To evaluate the influence of sample size, we chose $n \in \{1k, 2k, 3k, 4k, 5k\}$ with a specific population prevalence $p = 0.01$ and $OR = 2$. To assess the impact of the OR, we considered varying $OR \in \{1.5, 2, 2.5, 3, 3.5, 4\}$ given a sample size of $n = 4000$ (cases and controls were both 2000, balanced dataset) and $p = 0.01$.

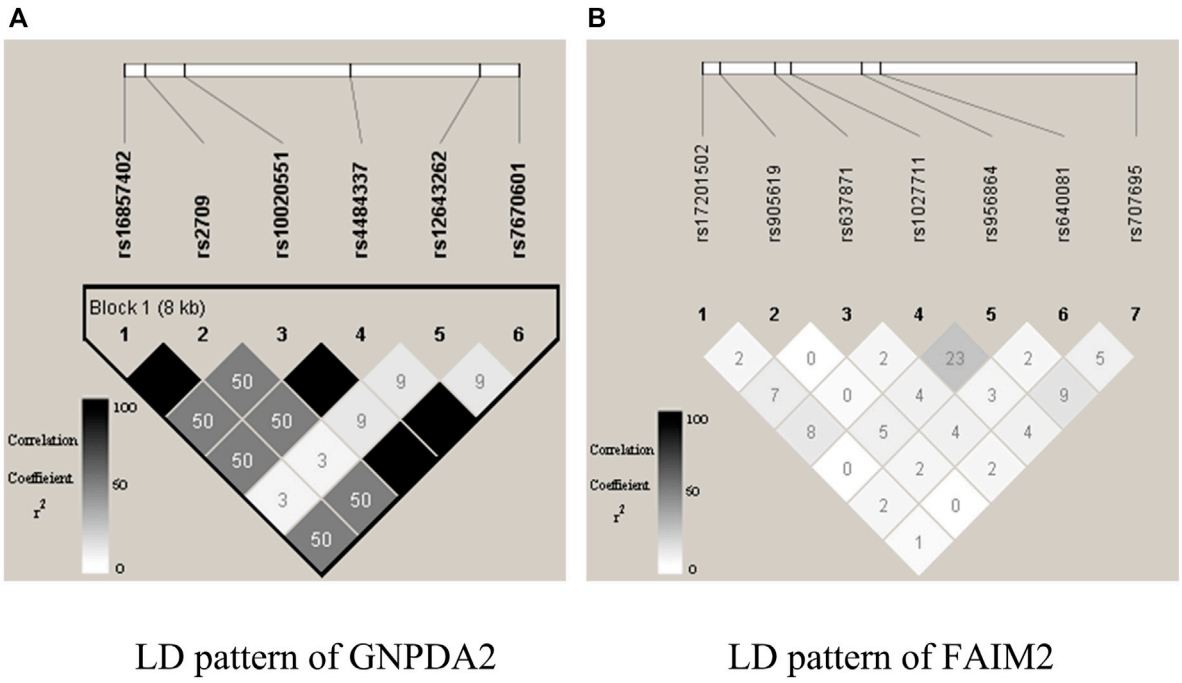


FIGURE 3 | Illustration of LD structures within genes GNPDA2 and FAIM2. The plots are generated by Haploview. r^2 measures the LD strength of each pair of SNPs in each square, $0 \leq r^2 \leq 1$, where 0 indicates no LD and 1 indicates complete LD. The GNPDA2 has a much stronger LD pattern than that within FAIM2, and they are not correlated.

For GGInt-XGBoost, KCCU, AGGrEGATOr, and GBIGM, if the number of datasets with a significance level less than α is m_1, then the power can be calculated by the following formula:

$$power = \frac{m_1}{100}. \tag{11}$$

GBIGM and AGGrEGATOr methods are nonparametric methods so no parameters need to be specific. We only set the ratio of the trimmed jackknife to 0.05 ($\omega = 0.05$) for KCCU. For GGInt-XGBoost, we set the number of trees to 1,000 (num_round = 1,000), the maximum depth of trees to 3 (max_depth = 3), the type of objective to "binary:logistic" (objective = "binary:logistic"), the learning rate to 0.01 (eta = 0.01), and the evaluate metric to error (eval_metric = "error"). We recommend that when dealing with real-data analysis, the depth of the trees is not to be set too deep in order to avoid overfitting. For a dataset with thousands of samples, a maximum depth of 2–4 is usually sufficient.

2.3 Experiments Using Rheumatoid Arthritis Data

To evaluate GGInt-XGBoost's ability to process real GGIs in a qualitative dataset, we analyzed the susceptibility of a series of pairs of genes in rheumatoid arthritis (RA), a chronic systemic disease with inflammatory synovitis with unknown etiology. It causes progressive bone destruction and affects bone remodeling. In this article, we chose the WTCCC (2007) dataset, which contained British population genotype data generated by the Affymetrix GeneChip 500 k. We preprocessed our dataset in the following ways:

i. To verify the GGIs in the RA, we selected the pathway hsa05323 from the KEGG pathway database. The genotyping coordinates of the WTCCC dataset can be found in UCSC hg18/NCBI Build36. There were 90 genes in this pathway. Among them, many genes belonged to MHCII and V-ATPase, which are two protein combinations. Because many GGIs occurred by themselves, we only selected representative genes from each protein combination, and then we excluded other genes. Finally, 48 genes remained, which resulted in $C_{48}^2 = 1128$ pairs of genes to be evaluated.

ii. The detailed gene information was obtained from the annotation file of NCBI Build36. For each gene, we added a 10 kb buffer region both downstream and upstream of the originally defined gene position. All SNPs within the region were selected for each gene.

iii. According to the quality control of GWAS, samples that included gender that did not match the chromosome X heterozygote rates were removed. SNPs were also excluded when they met any of the following conditions: the missing rate in the sample was $\geq 10\%$, the minor allele frequency (MAF) was ≤ 0.05, or the frequency of the control violated the Hardy–Weinberg equilibrium ($p < 0.0001$). Finally, 385 SNPs remained with 4,966 samples that consisted of 2,993 control subjects and 1973 case subjects.

3 RESULTS AND DISCUSSION

The experimental environment of the following results was a workstation with an Intel Xeon CPU E5-2,620 v2 at 2.10GHz, 96 GB of DDR3, Python3.6, and RStudio programming implementation.

TABLE 1 | Type-I error for KCCU, GBIGM, AGGrEGATOr, and GGInt-XGBoost when varying the sample size.

Method	Sample size				
	1,000	2,000	3,000	4,000	5,000
KCCU	0.02	0.02	0.01	0.05	0.07
GBIGM	0.13	0.06	0.07	0.07	0.07
AGGrEGATOr	0.05	0.06	0.07	0.04	0.02
GGInt-XGBoost	0.03	0.06	0.07	0.04	0.06

3.1 Simulation Study

3.1.1 Evaluation of Type-I Error

For type-I error, by setting the significance level at $\alpha = 0.05$, we varied the sample size from 1,000 to 5,000. All the methods tested, except for GBIGM when $n = 1,000$, had a type-I error comparable to the significance level (**Table 1**), which implied that these methods well controlled type-I error for various sample sizes.

3.1.2 Evaluation of the Power of the Test

Impact of Odds Ratio

We investigated the performance of the various methods in detecting GGIs under the six disease models, with a population prevalence of 0.01, the sample size of 4,000, and odds ratios that varied from 1.5 to 4 (**Figure 4**). For this experiment, a single pair of SNPs that belonged to different genes was chosen randomly for the disease models in the generation of the simulated dataset, and the genes that contained these two SNPs were considered to be interacting. A larger OR resulted in better performance for all methods, and, when $OR = 4$, some methods had a power that approached 1 (**Figure 4**). Our method was the best among all methods tested except for when $OR = 1.5$, which might be because the base

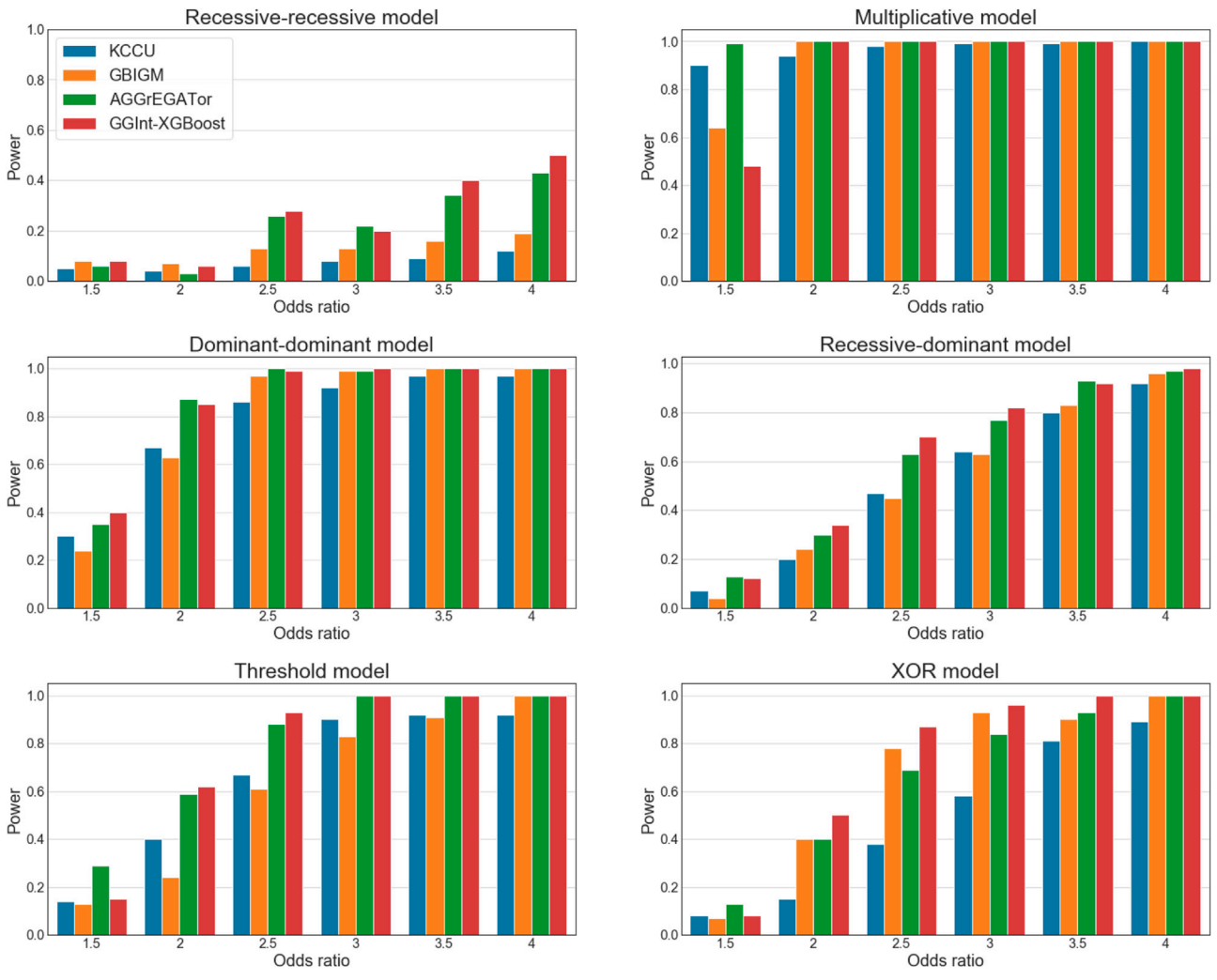


FIGURE 4 | Statistical power of simulation studies for KCCU (blue), GBIGM (yellow), AGGrEGATOr (green), and GGInt-XGBoost (red) under six disease models with $OR \in \{1.5, 2, 2.5, 3, 3.5, 4\}$

learner of XGBoost was a regression tree that might be prone to overfitting. It would be difficult to distinguish the signal from noise when the interaction strength was too low.

It is worth noting that in the recessive–recessive model (RR model) (**Supplementary Table S5A**), the detection power was consistently $\leq 20\%$ as the OR value changed gradually from 1.5 to 4. AGGrEGATOr and GGInt-XGBoost reached approximately 45%. According to the RR model penetrance table, the baseline γ was very small when the population prevalence was $p = 0.01$. Therefore, of the nine genotype combinations, eight of them tended to be zero. The only causal genotype (aabb) contained two minor alleles. Typically, the MAF of a SNP ranged from 0.2 to 0.4, and few genotypes (aabb) appeared in the simulation dataset. Consequently, it was difficult to detect the GGI under the disease phenotype. This was the main reason for the poor performance of these methods under the RR model.

Impact of Sample Size

We also investigated the influence of the sample size. Let the sample size be $n \in \{1k, 2k, 3k, 4k, 5k\}$, $p = 0.01$, and $OR = 2$ (**Supplementary Figure S1**). As the sample size increased, the detection power of all methods increased monotonically under all disease models, except for the RR model. In all methods, a larger sample size corresponded to better performance.

In conclusion, GGInt-XGBoost performed better in simulation studies than the other methods tested in almost every setting. The reason was probably that our method, by making use of constrained and unconstrained XGBoost models, made weak assumptions on the kind of interaction because any deviation from the additivity in the prediction of log odds ratios indicated an underlying GGI, which resulted in better statistical power. Furthermore, our method was more robust with respect to the LD pattern among the SNPs within each gene because the additivity constraint did not destroy the LD structure within each gene.

3.2 Experiments Using Rheumatoid Arthritis Data

Rheumatoid arthritis (RA) is an autoimmune disease with symptoms that typically include pannus formation in the synovial joints and destruction of cartilage and bones. The genes IL-17, IL-6, $TNF - \alpha$, RANK, and MMP3 are related to the development of RA (Majithia and Geraci, 2007). There were 48 genes in our dataset chosen from the RA pathway hsa05323, which resulted in 1,128 pairs of genes. We set significance level to $\alpha = 0.01$, and for our method, the number of permutations were set to $m = 1000$. GBIGM and KCCU resulted in 134 and 159 pairs of detected interacting genes, respectively. A total number of 65 of those pairs that were detected by GBIGM and 30 of the pairs that KCCU detected had a p-value of 0. AGGrEGATOr detected 17 pairs of interacting genes, and GGInt-XGBoost detected 58 pairs of interacting genes.

Because there were too many detected interacting gene pairs in KCCU and GBIGM with a p-value = 0, we could not analyze all of them in detail, so we focused on the 10 most significant gene pairs detected by AGGrEGATOr and, by our method, GGInt-XGBoost (**Table 2**). After a literature search, we found 7 of the 10 most significant gene pairs under GGInt-XGBoost and 3 of the 10 most significant gene pairs under AGGrEGATOr were supported by prior research. There was also a greater correlation between the results of GGInt-XGBoost and KCCU or GBIGM than the correlation between AGGrEGATOr and GBIGM or KCCU.

TABLE 2 | Calculated p-value for the 20 gene pairs using all four different methods. p-values in bold font indicate that they are significant. The ``Ref'' column indicates that the pair can be found as direct interaction in our literature search.

Gene1	Gene2	Ref	*p*-value			
			GGInt-XGBoost	AGGrEGATOr	KCCU	GBIGM
HLA class II	TGF β	Navarrete Santos et al. (1998)	**0.0**	0.588	**0.0**	**0.037**
HLA class II	LFA-1	Steere and Glickstein, (2004)	**0.0**	0.591	0.195	0.373
HLA class II	TEK		**0.0**	0.213	0.521	0.226
IL-8	ANG-1	Kabala et al. (2020)	**0.0**	1.0	0.818	0.32
MMP-3	April		**0.0**	0.164	0.161	0.063
HLA-DQA1	ANG-1		**0.0**	1.0	0.788	0.962
CTLA4	HLA class II	Karlson et al. (2008)	**0.0**	0.663	0.292	**0.0**
MMP-3	BLYs	Narasimhan et al. (2018)	**0.0**	0.473	**0.0001**	0.5
JUN	FOS	Huber et al. (2019)	**0.0**	0.441	0.692	**0.025**
GM-CSF	HLA class II	Field, (1995)	**0.0**	0.391	**0.047**	**0.0**
CD80	April		0.549	**0.0006**	0.941	0.334
CTSK	BLYS		**0.0**	**0.0008**	0.356	0.056
AP-1	IL-6		0.764	**0.0018**	0.098	0.287
CD86	CTSL		0.235	**0.0019**	0.519	0.252
CXCL6	FLT-1		0.098	**0.0023**	**0.004**	0.52
CTLA4	AP-1	Schneider and Rudd, (2014)	0.843	**0.0023**	**0.042**	0.102
FLT-1	LFA-1		0.117	**0.0031**	0.063	**0.028**
CCL3	TRAP	Jordan et al. (2018)	0.098	**0.0032**	0.682	**0**
IL-18	TGF β		0.137	**0.0036**	0.149	0.22
IL-1	SDF-1		0.647	**0.004**	0.116	0.636

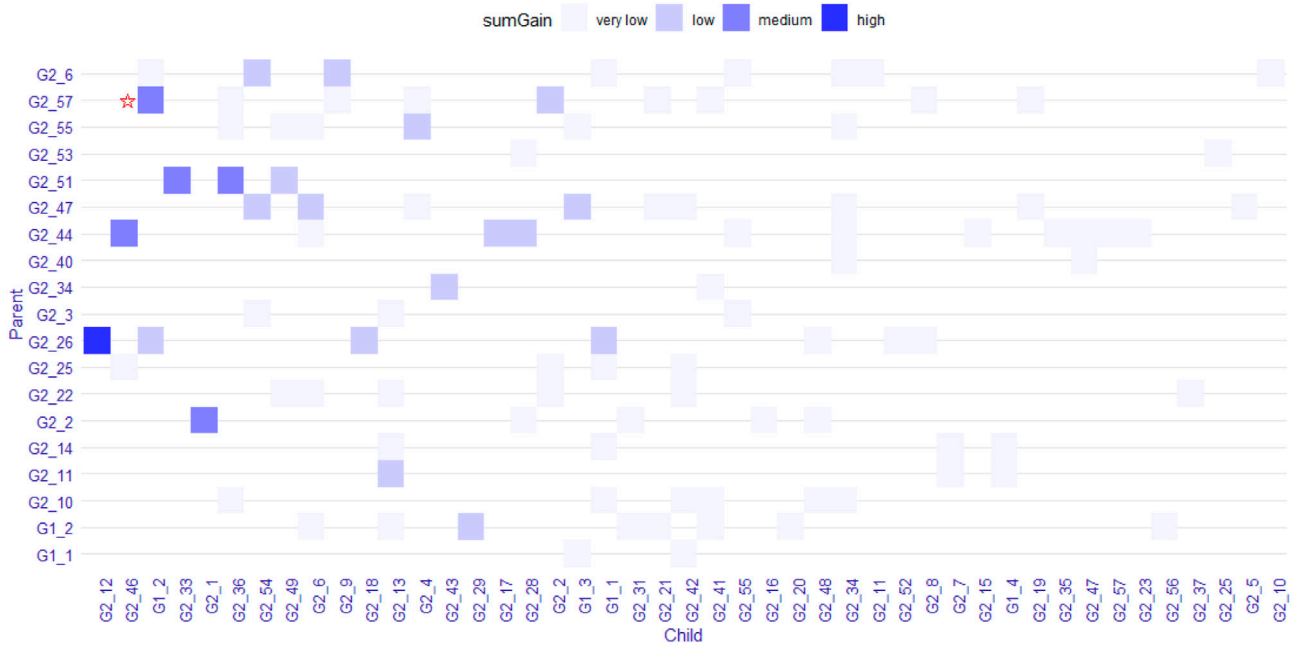


FIGURE 5 | Plot that shows pairs of SNPs that lie in two nodes of a regression tree connected by an edge. The color indicates the sumGain measure for the SNP pairs. The pair with the red star indicates the interacting SNP pair from IL-8 and Ang-I.

Furthermore, when using GGInt-XGBoost, after the detection of interacting gene pairs, one can also use the ensemble tree mechanism of XGBoost to investigate marker-based interactions further; this is because the regression tree, which is the base learner used in XGBoost, is a powerful tool for the discovery of interactions among features. For a regression tree model, one considers features that appear in the same traversal path from the root to leaf to be interacting. As an example, the gene pairs IL-8 and Ang-1 were found to interact using our method. Pawel et al.(Kabala et al., 2020) reported that Ang-1 induced the production of IL-8 in synovial tissue explants of RA patients. In the first tree in the unconstrained XGBoost model, it was clear that one SNP from the gene IL-8 on chromosome 4 interacted with rs121937926 in the gene Ang-1 (**Supplementary Figure S2**). The interaction form was flexible because our method imposed no functional form.

We explored the structure of the unconstrained XGBoost model further with R package EIX (Karbowiak and Biecek, 2021), which produced an interaction plot (**Figure 5**). For the convenience of display, all the SNPs in IL-8 were named "G1_X", and all SNPs in ANG-1 were named "G2_X", where "X" was the index. We chose the sumGain as a measure for the interaction strength. The sumGain was the sum of the gain value in all nodes in which the given SNP occurred. The intensity of the sumGain was divided into four equal parts and represented by different colored squares in the legend. The interacting SNP pairs in **Supplementary Figure S2** from IL-8 and Ang-1 exhibited median strength in **Figure 5** (with red star), which demonstrated that it was possible to use the results of GGInt-XGBoost for a more fine-grained analysis of GGIs at the marker level. Also, our method was robust with respect to LD because the LD structure within each gene was still expressed in the tree model and did not directly impact the performance of our method (**Figure 5**). **Table 3** gives the information of the top 10 interacted SNP pairs by sumGain and occurrence frequency in the ensemble boosting trees.

TABLE 3 | sumGain measure of the 10 most significant interacting SNP pairs from IL-8 and Ang-I. "Frequency" is the number of occurrences of the SNP pair in the trained model.

Index	Parent_SNP	Child_SNP	sumGain	Frequency
1	G2_26	G2_12	387.846,816	87
2	G2_44	G2_46	237.672,225	50
3	G2_57	G1_2	228.947,974	75
4	G2_51	G2_33	218.003046	33
5	G2_2	G2_1	214.650,794	68
6	G2_51	G2_36	213.414,267	28
7	G2_6	G2_54	184.671,309	81
8	G2_51	G2_49	178.749,247	27
9	G2_47	G2_6	154.492,033	31
10	G2_6	G2_9	140.040435	54

4 CONCLUSION

Gene–gene interactions (GGIs) are important in the study of complex diseases and traits. In this article, we developed a gene-based GGI detection algorithm called GGInt-XGBoost. We treated the GGI detection problem as a measure of how much the log odds ratio of qualitative traits deviated from the additive structure. GGInt-XGBoost benefits from the attractive built-in mechanism of XGBoost that allows for an elegant expression of the additive structure by adding feature interaction constraints.

Because of the weak assumptions of the interaction form and powerful and practical ability of XGBoost to fit nonlinear relationships, our method detected more types of interpretable GGIs accurately and effectively than other methods.

Combined with logistic regression, GGInt-XGBoost can be used for the GWAS of complex qualitative traits. To test its performance, we conducted a semi-empirical simulation study and a retrospective analysis of rheumatoid arthritis. For most of the settings tested, GGInt-XGBoost outperformed prior methods in statistical power for detecting GGIs. Also, because the base learner we used in XGBoost was the regression tree, GGIng-XGBoost can detect GGIs under quantitative traits. Furthermore, the base learner of XGBoost is a tree model that has a natural way of expressing marker-based interactions, which allows further investigations of interactions at the marker level after two genes are known to interact. For example, we looked for interactions between the genes IL-8 and Ang-1 and found that it was largely accounted for by the interaction between a single pair of SNPs from these two genes. Also, through the analysis of IL-8 and Ang-1, we found that GGInt-XGBoost was robust with respect to the LD structure within genes. The workflow designed for detection of GGIs did not damage the LD structure, and the assumption of the additive structure allowed marker-based interaction within genes. Last, GGInt-XGBoost might be improved further or generalized by incorporating ideas from causal inferences that would be applied more effectively to multi-gene settings and the study of gene pathways. In conclusion, GGInt-XGBoost is a helpful addition to the existing toolbox of statistical methods for studying gene–gene interaction in genome-wide association studies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://www.wtccc.org.uk/info/access_to_data_samples.html

AUTHOR CONTRIBUTIONS

YZ and LX contributed to conceptualization and project administration. YG contributed to conceptualization, methodology, investigation, funding acquisition, and writing-original draft. CW contributed to methodology, formal analysis, and writing-original draft. ZY contributed to software and formal analysis. YW (4th author) contributed to data curation and visualization. ZL contributed to resources and data curation. YW (6th author) contributed to formal analysis, writing-review, and editing. All authors contributed to the article and approved the submitted version.

FUNDING

The work was supported by the National Natural Science Foundation of China (No. 62002243, No. 31900306), and the Research Foundation of ShenZhen Polytechnic (6021310019K).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.801113/full#supplementary-material

REFERENCES

Babajide Mustapha, I., and Saeed, F. (2016). Bioactive Molecule Prediction Using Extreme Gradient Boosting. *Molecules* 21 (8). doi:10.3390/molecules21080983

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of Published Genome-wide Association Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Res.* 47 (D1), D1005–D1012. doi:10.1093/nar/gky1120

Chang, Y.-C., Wu, J. T., Wu, J.-T., Hong, M.-Y., Tung, Y.-A., Hsieh, P.-H., et al. (2020). GenEpi: Gene-Based Epistasis Discovery Using Machine Learning. *BMC Bioinformatics* 21 (1), 68. doi:10.1186/s12859-020-3368-2

Chen, T., and Guestrin, C. (2016). "XGBoost : A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA; New York, NY: Association for Computing Machinery), 785–794.

Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018). EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association Prediction. *Cell Death Dis* 9 (1), 3. doi:10.1038/s41419-017-0003-x

Cordell, H. J. (2009). Detecting Gene-Gene Interactions that Underlie Human Diseases. *Nat. Rev. Genet.* 10 (6), 392–404. doi:10.1038/nrg2579

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., et al. (2008). Exploration of Gene-Gene Interaction Effects Using Entropy-Based Methods. *Eur. J. Hum. Genet.* 16 (2), 229–235. doi:10.1038/sj.ejhg.5201921

Emily, M.: A Survey of Statistical Methods for Gene-Gene Interaction in Case-Control Genome-wide Association Studies. 2018.

Emily, M. (2016). AGGrEGATOr: A Gene-Based GEne-Gene interActTiOn Test for Case-Control Association Studies. *Stat. Appl. Genet. Mol. Biol.* 15 (2), 151–171. doi:10.1515/sagmb-2015-0074

Emily, M., Sounac, N., Kroell, F., and Houée-Bigot, M. (2020). Gene-Based Methods to Detect Gene-Gene Interaction in R: The GeneGeneInteR Package. *J. Stat. Softw.* 95 (12). doi:10.18637/jss.v095.i12

Emily, M. (2012). IndOR: a New Statistical Procedure to Test for SNP-SNP Epistasis in Genome-wide Association Studies. *Statist. Med.* 31 (21), 2359–2373. doi:10.1002/sim.5364

Fang, G., Wang, W., Paunic, V., Heydari, H., Costanzo, M., Liu, X., et al. (2019). Discovering Genetic Interactions Bridging Pathways in Genome-wide Association Studies. *Nat. Commun.* 10 (1), 4274. doi:10.1038/s41467-019-12131-7

Field, M. (1995). Colony-stimulating Factors. *Clin. Immunother.* 3 (4), 255–261. doi:10.1007/bf03259277

Friedman, J. H., and Stuetzle, W. (1981). Projection Pursuit Regression. *J. Am. Stat. Assoc.* 76 (376), 817–823. doi:10.1080/01621459.1981.10477729

Guo, F., Wang, D., and Wang, L. (2018). Progressive Approach for SNP Calling and Haplotype Assembly Using Single Molecular Sequencing Data. *Bioinformatics* 34 (12), 2012–2018. doi:10.1093/bioinformatics/bty059

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. London, United Kingdom: Chapman & Hall.

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020). TOOme: A Novel Computational Framework to Infer Cancer Tissue-Of-Origin by Integrating Both Gene Mutation and Expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential Etiologic and Functional Implications of Genome-

wide Association Loci for Human Diseases and Traits. *Proc. Natl. Acad. Sci.* 106 (23), 9362–9367. doi:10.1073/pnas.0903103106

Hu, Y., Sun, J. Y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (2021). rs1990622 Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19 (1), 11. doi:10.1186/s12916-020-01883-5

Huber, R., Kirsten, H., Näkki, A., Pohlers, D., Thude, H., Eidner, T., et al. (2019). Association of Human FOS Promoter Variants with the Occurrence of Knee-Osteoarthritis in a Case Control Association Study. *Int. J. Mol. Sci.* 20 (6). doi:10.3390/ijms20061382

Jiang, L., Wang, C., Tang, J., and Guo, F. (2019). LightCpG: a Multi-View CpG Sites Detection on Single-Cell Whole Genome Sequence Data. *Bmc Genomics* 20, 306. doi:10.1186/s12864-019-5654-9

Jiang, Q., Jin, S., Jiang, Y., Liao, M., Feng, R., Zhang, L., et al. (2017). Alzheimer's Disease Variants with the Genome-wide Significance Are Significantly Enriched in Immune Pathways and Active in Immune Cells. *Mol. Neurobiol.* 54 (1), 594–600. doi:10.1007/s12035-015-9670-8

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting Human microRNA-Disease Associations Based on Support Vector Machine. *Ijdmb* 8 (3), 282–293. doi:10.1504/ijdmb.2013.056078

Jordan, L. A., Erlandsson, M. C., Fenner, B. F., Davies, R., Harvey, A. K., Choy, E. H., et al. (2018). Inhibition of CCL3 Abrogated Precursor Cell Fusion and Bone Erosions in Human Osteoclast Cultures and Murine Collagen-Induced Arthritis. *Rheumatology (Oxford)* 57 (11), 2042–2052. doi:10.1093/rheumatology/key196

Kabala, P. A., Malvar-Fernández, B., Lopes, A. P., Carvalheiro, T., Hartgring, S. A. Y., Tang, M. W., et al. (2020). Promotion of Macrophage Activation by Tie2 in the Context of the Inflamed Synovia of Rheumatoid Arthritis and Psoriatic Arthritis Patients. *Rheumatology (Oxford)* 59 (2), 426–438. doi:10.1093/rheumatology/kez315

Karbowiak, E., and Biecek, P. (2021). "EIX: Explain Interactions in 'XGBoost'," in *R Package Version 1.1*. Available at: https://github.com/ModelOriented/EIX.

Karlson, E. W., Chibnik, L. B., Cui, J., Plenge, R. M., Glass, R. J., Maher, N. E., et al. (2008). Associations between Human Leukocyte Antigen, PTPN22, CTLA4 Genotypes and Rheumatoid Arthritis Phenotypes of Autoantibody Status, Age at Diagnosis and Erosions in a Large Cohort Study. *Ann. Rheum. Dis.* 67 (3), 358–363. doi:10.1136/ard.2007.071662

Larson, N. B., Jenkins, G. D., Larson, M. C., Vierkant, R. A., Sellers, T. A., Phelan, C. M., et al. (2013). Kernel Canonical Correlation Analysis for Assessing Gene-Gene Interactions and Application to Ovarian Cancer. *Eur. J. Hum. Genet.* 22 (1), 126–131. doi:10.1038/ejhg.2013.69

Lemieux Perreault, L. P., Legault, M. A., Asselin, G., and Dubé, M. P. (2016). Genipe: an Automated Genome-wide Imputation Pipeline with Automatic Reporting and Statistical Tools. *Bioinformatics* 32 (23), 3661–3663. doi:10.1093/bioinformatics/btw487

Li, H.-L., Pang, Y.-H., and Liu, B. (2021). BioSeq-BLM: a Platform for Analyzing DNA, RNA and Protein Sequences Based on Biological Language Models. *Nucleic Acids Res.*, gkab829. doi:10.1093/nar/gkab829

Li, J., and Chen, Y. (2008). Generating Samples for Association Studies Based on HapMap Data. *BMC bioinformatics* 9 (1), 44–13. doi:10.1186/1471-2105-9-44

Li, J., Huang, D., Guo, M., Liu, X., Wang, C., Teng, Z., et al. (2015). A Gene-Based Information Gain Method for Detecting Gene-Gene Interactions in Case-Control Studies. *Eur. J. Hum. Genet.* 23 (11), 1566–1572. doi:10.1038/ejhg.2015.16

Li, M.-X., Gui, H.-S., Kwan, J. S. H., and Sham, P. C. (2011). GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *Am. J. Hum. Genet.* 88 (3), 283–293. doi:10.1016/j.ajhg.2011.01.019

Li, P., Guo, M., Wang, C., Liu, X., and Zou, Q. (2015). An Overview of SNP Interactions in Genome-wide Association Studies. *Brief. Funct. Genomics* 14 (2), 143–155. doi:10.1093/bfgp/elu036

Lin, H., Mueller-Nurasyid, M., Smith, A. V., Arking, D. E., Barnard, J., Bartz, T. M., et al. (2016). Gene-gene Interaction Analyses for Atrial Fibrillation. *Sci. Rep.* 6 (1), 35371–35379. doi:10.1038/srep35371

Liu, G., Hu, Y., Han, Z., Jin, S., and Jiang, Q. (2019). Genetic Variant Rs17185536 Regulates SIM1 Gene Expression in Human Brain Hypothalamus. *Proc. Natl. Acad. Sci. USA* 116 (9), 3347–3348. doi:10.1073/pnas.1821550116

Liu, G., Hu, Y., Jin, S., and Jiang, Q. (2017). Genetic Variant Rs763361 Regulates Multiple Sclerosis CD226 Gene Expression. *Proc. Natl. Acad. Sci. USA* 114 (6), E906–E907. doi:10.1073/pnas.1618520114

Liu, G., Hu, Y., Jin, S., Zhang, F., Jiang, Q., and Hao, J. (2016). Cis-eQTLs Regulate Reduced LST1 Gene and NCR3 Gene Expression and Contribute to Increased Autoimmune Disease Risk. *Proc. Natl. Acad. Sci. USA* 113 (42), E6321–E6322. doi:10.1073/pnas.1614369113

Liu, G., and Jiang, Q. (2016). Alzheimer's Disease CD33 Rs3865444 Variant Does Not Contribute to Cognitive Performance. *Proc. Natl. Acad. Sci. USA* 113 (12), E1589–E1590. doi:10.1073/pnas.1600852113

Liu, G., Jin, S., Hu, Y., and Jiang, Q. (2018). Disease Status Affects the Association between Rs4813620 and the Expression of Alzheimer's Disease Susceptibility geneTRIB3. *Proc. Natl. Acad. Sci. USA* 115 (45), E10519–E10520. doi:10.1073/pnas.1812975115

Liu, G., Zhang, Y., Wang, L., Xu, J., Chen, X., Bao, Y., et al. (2018). Alzheimer's Disease Rs11767557 Variant Regulates EPHA1 Gene Expression Specifically in Human Whole Blood. *Jad* 61 (3), 1077–1088. doi:10.3233/jad-170468

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-Of-Origin. *Front. Cell Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. *Am. J. Hum. Genet.* 87 (1), 139–145. doi:10.1016/j.ajhg.2010.06.009

Loos, R. J. F. (2020). 15 Years of Genome-wide Association Studies and No Signs of Slowing Down. *Nat. Commun.* 11 (1), 5900. doi:10.1038/s41467-020-19653-5

Luo, J., Meng, Y., Zhai, J., Zhu, Y., Li, Y., and Wu, Y. (2020). Screening of SLE-Susceptible SNPs in One Chinese Family with Systemic Lupus Erythematosus. *Cbio* 15 (7), 778–787. doi:10.2174/1574893615666200120105153

Lv, Z., Wang, D., Ding, H., Zhong, B., and Xu, L. (2020). Escherichia Coli DNA N-4-Methycytosine Site Prediction Accuracy Improved by Light Gradient Boosting Machine Feature Selection Technology. *IEEE Access* 8, 14851–14859. doi:10.1109/access.2020.2966576

Lyu, P., Hou, J., Yu, H., and Shi, H. (2020). High-density Genetic Linkage Map Construction in Sunflower (Helianthus Annuus L.) Using SNP and SSR Markers. *Curr. Bioinformatics* 15 (8), 889–897. doi:10.2174/1574893615666200324134725

Ma, L., Clark, A. G., and Keinan, A. (2013). Gene-based Testing of Interactions in Association Studies of Quantitative Traits. *Plos Genet.* 9 (2), e1003321. doi:10.1371/journal.pgen.1003321

Majithia, V., and Geraci, S. A. (2007). Rheumatoid Arthritis: Diagnosis and Management. *Am. J. Med.* 120 (11), 936–939. doi:10.1016/j.amjmed.2007.04.005

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the Missing Heritability of Complex Diseases. *Nature* 461 (7265), 747–753. doi:10.1038/nature08494

Moore, J. H., Asselbergs, F. W., and Williams, S. M. (2010). Bioinformatics Challenges for Genome-wide Association Studies. *Bioinformatics* 26 (4), 445–455. doi:10.1093/bioinformatics/btp713

Mrozek, D., Daniłowicz, P., and Małysiak-Mrozek, B. (2016). HDInsight4PSi: Boosting Performance of 3D Protein Structure Similarity Searching with HDInsight Clusters in Microsoft Azure Cloud. *Inf. Sci.* 349-350, 77–101. doi:10.1016/j.ins.2016.02.029

Narasimhan, R., Coras, R., Rosenthal, S. B., Sweeney, S. R., Lodi, A., Tiziani, S., et al. (2018). Serum Metabolomic Profiling Predicts Synovial Gene Expression in Rheumatoid Arthritis. *Arthritis Res. Ther.* 20 (1), 164. doi:10.1186/s13075-018-1655-3

Navarrete Santos, A., Kehlen, A., Schütte, W., Langner, J., and Riemann, D. (1998). Regulation by Transforming Growth Factor-Beta1 of Class II mRNA and Protein Expression in Fibroblast-like Synoviocytes from Patients with Rheumatoid Arthritis. *Int. Immunol.* 10 (5), 601–607. doi:10.1093/intimm/10.5.601

Nobre, R., Ilic, A., Santander-Jimenez, S., and Sousa, L. (2021). Retargeting Tensor Accelerators for Epistasis Detection. *IEEE Trans. Parallel Distrib. Syst.* 32 (9), 2160–2174. doi:10.1109/tpds.2021.3060322

Peng, Q., Zhao, J., and Xue, F. (2010). A Gene-Based Method for Detecting Gene-Gene Co-association in a Case-Control Association Study. *Eur. J. Hum. Genet.* 18 (5), 582–587. doi:10.1038/ejhg.2009.223

Ritchie, M. D., Hahn, L. W., and Moore, J. H. (2003). Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the

Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity. *Genet. Epidemiol.* 24 (2), 150–157. doi:10.1002/gepi.10218

Ritchie, M. D., and Van Steen, K. (2018). The Search for Gene-Gene Interactions in Genome-wide Association Studies: Challenges in Abundance of Methods, Practical Considerations, and Biological Interpretation. *Ann. Transl. Med.* 6 (8), 157. doi:10.21037/atm.2018.04.05

Schneider, H., and Rudd, C. E. (2014). Diverse Mechanisms Regulate the Surface Expression of Immunotherapeutic Target Ctla-4. *Front. Immunol.* 5, 619. doi:10.3389/fimmu.2014.00619

Shao, J., and Liu, B. (2021). ProtFold-DFG: Protein Fold Recognition by Combining Directed Fusion Graph and PageRank Algorithm. *Brief Bioinform* 22 (3), bbaa192. doi:10.1093/bib/bbaa192

Shao, J., Yan, K., and Liu, B. (2021). FoldRec-C2C: Protein Fold Recognition by Combining Cluster-To-Cluster Model and Protein Similarity Network. *Brief Bioinform* 22 (3), bbaa144. doi:10.1093/bib/bbaa144

Steere, A. C., and Glickstein, L. (2004). Elucidation of Lyme Arthritis. *Nat. Rev. Immunol.* 4 (2), 143–152. doi:10.1038/nri1267

Stone, C. J. (1985). Additive Regression and Other Nonparametric Models. *Ann. Stat.* 13, 689–705. doi:10.1214/aos/1176349548

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., et al. (2010). BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *Am. J. Hum. Genet.* 87 (3), 325–340. doi:10.1016/j.ajhg.2010.07.021

Wang, H., Jijun, T., Ding, Y., and Guo, F. (2021). Exploring Associations of Non-coding RNAs in Human Diseases via Three-Matrix Factorization with Hypergraph-Regular Terms on center Kernel Alignment. *Brief. Bioinform.* 22, bbaa409. doi:10.1093/bib/bbaa409

Wang, Z., He, W., Tang, J., and Guo, F. (2020). Identification of Highest-Affinity Binding Sites of Yeast Transcription Factor Families. *J. Chem. Inf. Model.* 60 (3), 1876–1883. doi:10.1021/acs.jcim.9b01012

Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: A Sequence-Based Predictor for Identifying N6-Methyladenosine Sites Using Ensemble Learning. *Mol. Ther. - Nucleic Acids* 12, 635–644. doi:10.1016/j.omtn.2018.07.004

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017). A Novel Hierarchical Selective Ensemble Classifier with Bioinformatics Application. *Artif. Intelligence Med.* 83, 82–90. doi:10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved Prediction of Protein-Protein Interactions Using Novel Negative Samples, Features, and an Ensemble Classifier. *Artif. Intelligence Med.* 83, 67–74. doi:10.1016/j.artmed.2017.03.001

Yang, J., Huang, T., Huang, T., Petralia, F., Long, Q., Zhang, B., et al. (2015). Synchronized Age-Related Gene Expression Changes across Multiple Tissues in Human and the Link to Complex Diseases. *Sci. Rep.* 5, 15145. doi:10.1038/srep15145

Young, A. I. (2019). Solving the Missing Heritability Problem. *Plos Genet.* 15 (6), e1008222. doi:10.1371/journal.pgen.1008222

Yu, B., Qiu, W., Chen, C., Ma, A., Jiang, J., Zhou, H., et al. (2020). SubMito-XGBoost: Predicting Protein Submitochondrial Localization by Fusing Multiple Feature Information and eXtreme Gradient Boosting. *Bioinformatics* 36 (4), 1074–1081. doi:10.1093/bioinformatics/btz734

Yu, X., Zhou, J., Zhao, M., Yi, C., Duan, Q., Zhou, W., et al. (2020). Exploiting XG Boost for Predicting Enhancer-Promoter Interactions. *Curr. Bioinformatics* 15 (9), 1036–1045.

Yuan, Z., Gao, Q., He, Y., Zhang, X., Li, F., Zhao, J., et al. (2012). Detection for Gene-Gene Co-association via Kernel Canonical Correlation Analysis. *BMC Genet.* 13, 83. doi:10.1186/1471-2156-13-83

Zhang, S., Jiang, W., Ma, R. C., and Yu, W. (2019). Region-based Interaction Detection in Genome-wide Case-Control Studies. *BMC Med. Genomics* 12 (Suppl. 7), 133. doi:10.1186/s12920-019-0583-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Guo, Wu, Yuan, Wang, Liang, Wang, Zhang and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

METHODS
published: 17 December 2021
doi: 10.3389/fcell.2021.753027

GBDTLRL2D Predicts LncRNA–Disease Associations Using MetaGraph2Vec and K-Means Based on Heterogeneous Network

Tao Duan, Zhufang Kuang, Jiaqi Wang and Zhihao Ma*

School of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, China

In recent years, the long noncoding RNA (lncRNA) has been shown to be involved in many disease processes. The prediction of the lncRNA–disease association is helpful to clarify the mechanism of disease occurrence and bring some new methods of disease prevention and treatment. The current methods for predicting the potential lncRNA–disease association seldom consider the heterogeneous networks with complex node paths, and these methods have the problem of unbalanced positive and negative samples. To solve this problem, a method based on the Gradient Boosting Decision Tree (GBDT) and logistic regression (LR) to predict the lncRNA–disease association (GBDTLRL2D) is proposed in this paper. MetaGraph2Vec is used for feature learning, and negative sample sets are selected by using K-means clustering. The innovation of the GBDTLRL2D is that the clustering algorithm is used to select a representative negative sample set, and the use of MetaGraph2Vec can better retain the semantic and structural features in heterogeneous networks. The average area under the receiver operating characteristic curve (AUC) values of GBDTLRL2D obtained on the three datasets are 0.98, 0.98, and 0.96 in 10-fold cross-validation.

Keywords: long noncoding RNA, heterogeneous network, MetaGraph2Vec, K-means, Gradient Boosting Decision Tree, logistic regression

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Yongjun Tang,
Central South University, China
Bingbo Wang,
Xidian University, China

****Correspondence:***
Zhufang Kuang
zfkuangcn@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *04 August 2021*
Accepted: *22 November 2021*
Published: *17 December 2021*

Citation:
Duan T, Kuang Z, Wang J and Ma Z (2021) GBDTLRL2D Predicts LncRNA–Disease Associations Using MetaGraph2Vec and K-Means Based on Heterogeneous Network. Front. Cell Dev. Biol. 9:753027. doi: 10.3389/fcell.2021.753027

1 INTRODUCTION

In the human genome, more than 98% of the genes are noncoding protein sequences. The remaining 2% can only be transcribed into noncoding RNAs (ncRNAs). The ncRNAs can be divided into microRNA (miRNA), long ncRNA (lncRNA), etc. NcRNAs between 200 and 100,000 in length are lncRNAs.

At first, lncRNAs are considered useless RNAs without any biological function (Mercer and Mattick, 2013). This is because they are expressed at a lower level than protein-coding RNAs. However, with the development of experimental methods and computing power, the change of lncRNA has been found to be associated with many diseases, such as colorectal cancer (Xiong et al., 2021), lung adenocarcinoma (Hou et al., 2021), and gastrointestinal cancer (Abdi et al., 2021). With the deepening of research on lncRNAs, there are many pieces of evidence that lncRNAs play a key role in many important biological processes, including transcription, translation, splicing, differentiation, epigenetic regulation, immune response, and cell cycle control. For example, lncRNA loc105377478 promotes NPs-Nd2O3 in 16HBE cells and thus induces inflammation in human bronchial epithelial cells (Yu et al., 2021). LncRNA HOTAIR is considered a potential

biomarker (Maass et al., 2014). The expression level of HOTAIR in breast cancer tissues is 100 to approximately 2,000 times higher than normal tissues and is associated with the proliferation and survival of colorectal cancer (Ge et al., 2013). Some research has shown that lncRNA BCAR4 is expressed in 27% of primary breast tumors. LncRNA Braveheart has also been demonstrated to control heart development by interacting with the epigenetic modifier PRC2. And increasing the expression of lncRNA Linc-MD1 can promote muscle differentiation (Cesana et al., 2011). LncRNA NEAT1 can regulate the development of Parkinson's. Therefore, the research of the potential lncRNA–disease association can better comprehend the potential mechanism of human diseases and help diagnose and treat diseases. This research has important practical implications.

Biological experiments to identify potential associations are time-consuming, labor-intensive, and very expensive. Therefore, in order to effectively reduce the time consumed by biological experiments and economic costs, there has been much research based on bioinformatics and computational power. For example, the method KATZLGO is proposed by Zhang et al. (2017) to predict the interaction of lncRNA–lncRNA. The PLPIHS is proposed by Xiao et al. (2017) to predict lncRNA–protein interactions using HeteSim score. A computational framework for predicting lncRNA–protein interactions is proposed by Liu (2020). An approach to explore miRNA sponge networks in breast cancer is proposed by Tian and Wang (2021). A method to predict the subcellular localization of lncRNAs is proposed by Yang et al. (2020). The potential roles of oral squamous cell carcinoma (OSCC)-related mRNA and lncRNA are revealed by Li et al. (2021) through protein interaction network and co-expression network analysis. The model GBDTL2E is proposed by Wang et al. (2020) to predict the association between lncRNA and environmental factors. With the deepening of research, research on the prediction of lncRNA–disease association is mainly divided into the following categories:

1) Based on machine learning methods, the main idea of these methods is to prioritize candidate lncRNAs by training known and unknown lncRNA–disease correlation. The semi-supervised learning framework LRLSLDA is proposed by Chen and Yan (2013). A graph regularization non-negative matrix factorization (LDGRNMF) is proposed by Wang M.-N. et al. (2021). Based on the weight algorithm and the improved projection algorithm, LDAP-WMPS is proposed by Wang B. et al. (2021). A model proposed by Zhou et al. (2021) uses high-order proximity reserved embedding to embed nodes into the network. The model VGAELDA, which integrates variational reasoning and graph autoencoder, is proposed by Shi et al. (2021). A multi-label fusion collaborative matrix decomposition (MLFCMF) method is proposed by Gao et al. (2021) to predict lncRNA–disease associations. The model PSPA-LA-PCRA is proposed by Wang and Zhang (2021), which uses the data of pathological stages. The random distribution logical regression framework (RDLRF) is proposed by Sun et al. (2021), and the RDLRF combines simboost feature extraction with logistic regression (LR). The method FVTLDA is proposed by Xiao et al. (2020), which combines multiple linear regression and artificial neural network. The BLM-NPA is proposed by Cui et al. (2019) to predict based on the nearest neighbor. The alternate least squares method of matrix factor factorization (ALSBMF) is proposed by Zhu et al. (2020). A computational method based on graphical autoencoder matrix completion (GAMCLDA) is proposed by Wu et al. (2020). A deep matrix factorization method (DMFLDA) is proposed by Zeng et al. (2020). A graph-based method (PANDA) is proposed by Silva and Spinosa (2021). The PANDA takes the association prediction of lncRNAs and diseases as a link prediction problem.

2) Based on network methods, the main idea of these methods is using a similarity network to predict lncRNA–disease association. Based on the combination of incremental principal component analysis (IPCA) and random forest (RF), a lncRNA–disease association prediction method IPCARF is proposed by Zhu et al. (2021). The prediction method for lncRNA–disease associations (PCSLDA) based on Point Cut Set is proposed by Kuang et al. (2019). The model GAERF is proposed by Wu et al. (2021); GAREF uses graph autocoding (GAE) and RF to identify disease-related lncRNAs. A random walk-based multi-similarity fusion and bidirectional label propagation method RWSF-BLP is proposed by Xie et al. (2021). Based on the assumption that there is a potential association between an lncRNA and a disease, if they are associated with the same set of miRNAs, similar diseases tend to be closely related to IneRNAs with similar functions; the method LDLMD is proposed by (Wang et al., 2019). A method of internal confidence-based local radial basis biological network (ICLRBBN) is proposed by Wang Y. et al. (2021). A two-stage prediction model (DRW-BNSP) is proposed by Zhang et al. (2021). HAUBRW algorithm is proposed by Xie et al. (2020b), which combines thermal diffusion algorithm and probabilistic diffusion algorithm to redistribute resources. An lncRNA–disease association prediction model based on RF and feature selection, RFLDA, is proposed by Yao et al. (2020). A predictive lncRNA–disease prediction model based on heterogeneous networks is proposed by Song et al. (2020). The LDAMAN is proposed by Zhang et al. (2020), which uses a structural deep network embedding model. The method based on linear neighborhood similarity and unbalanced double random walk (LDA-LNSUBRW) is proposed by Xie et al. (2020a). The MHRWR is proposed by Zhao et al. (2020) to integrate the similarity network of lncRNAs, diseases, and genes, with the known lncRNA–disease association network, lncRNA–gene network, and disease–gene network. The method LDAH2V is proposed by Deng et al. (2021), which uses HIN2Vec to calculate the meta path and eigenvector of each lncRNA–disease pair in heterogeneous information networks.

It can be seen that the association prediction of lncRNAs and diseases has become a research hotspot. Currently, the existing methods simply regard all objects in the network as the same type. However, in heterogeneous networks, there are many types of nodes, and the relationship between nodes is very complex, which is not considered by traditional methods. At the same time, unknown correlation is far greater than known correlation, which brings great challenges to model training. To solve

TABLE 1 | LncRNA–disease association relationship dataset.

Dataset	Number of lncRNA	Number of diseases	Number of associations
DataSet1 (DS1)	112	150	276
DataSet2 (DS2)	131	169	319
DataSet3 (DS3)	285	226	621

Note. lncRNA, long noncoding RNA.

these problems, a method based on the Gradient Boosting Decision Tree (GBDT) and LR to predict the lncRNA–disease association (GBDTLRL2D) is proposed in this paper. The GBDTLRL2D uses MetaGraph2Vec for feature learning and the K-means clustering method to select negative sample sets. The contributions of our method are included:

- The GBDTLRL2D comprehensively considers the topological structure characteristics and meta-path characteristics of nodes in heterogeneous networks. MetaGraph2Vec is used to learn more information by capturing more semantic relationships between remote nodes.
- The GBDTLRL2D uses the K-means to get the clustering of the unknown correlation. The same number of negative samples as the positive samples is selected from the clusters.
- The GBDTLRL2D combines GBDT and LR. The GBDT + LR is a special classification algorithm. Its ability to find features and combine features is very powerful. The classification accuracy is high.

2 MATERIALS AND METHODS

The known lncRNA and disease-associated data used in this paper are downloaded from the lncRNADisease (Chen et al., 2012), which includes three versions, namely, the version of June 2012, the version of January 2014, and the version of June 2015. **Table 1** shows the data after deduplication.

In this section, a method based on the GBDT and LR to predict the lncRNA–disease association (GBDTLRL2D) is proposed. The GBDTLRL2D uses MetaGraph2Vec for feature learning and the K-means clustering method to select negative sample sets. The main steps of GBDTLRL2D are as follows: 1) according to the downloaded data, the set of lncRNAs and diseases as well as the association matrix A of lncRNA–diseases is obtained after deduplication. 2) The disease semantic similarity matrix SSD and lncRNA functional similarity matrix FSL are calculated, and then the Gaussian interaction profile kernel similarity matrix of disease (GSD) and lncRNA (GSL) are calculated. 3) LncRNA similarity matrix SL is constructed according to GSL and FSL, and disease similarity matrix SD is constructed according to GSD and SSD. 4) The association matrix A of lncRNA–disease, lncRNA similarity matrix SL, and the disease similarity matrix SD are integrated to construct the global heterogeneous network G. On the G, the feature of each node is learned by MetaGraph2Vec to obtain the feature representation of each node. 5) K-means is used to select negative samples to obtain train sets. 6) The GBDT and LR classifier are used to predict the lncRNA–disease association. **Figure 1** is a flowchart of the GBDTLRL2D. Each step of GBDTLRL2D is detailed in the next section.

2.1 Calculate Disease Semantic Similarity

The computing method of disease semantic similarity SSD in this experiment is based on the Disease Ontology. The method presents the disease tissue as a directed acyclic graph (DAG). As shown in **Figure 2**, the relationships between diseases are described in a DAG. Each node is a disease, and the arrow points from a disease to its ancestor disease. It can be seen from **Figure 3** that if one lncRNA is associated with a disease, then this lncRNA may be associated with sub-diseases of the disease. From this perspective, the correct identification of these new associations may help to understand the mechanisms underlying RNA levels and improve the speed of accurate diagnosis and treatment of diseases. Therefore, the SSD between diseases is calculated according to the DAG.

For disease d_i, the semantic value is obtained. The contribution of each ancestral disease in the DAG of disease d_u to the semantic value of d_i is firstly calculated as shown in **formula (1)**:

$$C_i(u) = \begin{cases} 1, & \text{if } u = i \\ \max\{\Delta \times C_i(u') \mid u' \in \text{ children of } u\}, & \text{if } u \neq i \end{cases} \tag{1}$$

where Δ is the weight of the edge connecting disease d_u and its sub-diseases, namely, semantic contribution factor. According to the above formula, as the distance between disease d_i and other diseases increases, semantic contributions decrease. Therefore, Δ should be selected between 0 and 1. In this paper, Δ= 0.5. Then the semantic value of d_i is calculated as the sum of the contributions of d_i's ancestor disease and d_i itself, as shown in **formula (2)**:

$$C(i) = \sum_{u \in Z(i)} C_i(u) \tag{2}$$

where Z(i) represents the node set in the DAG of disease d_i.

For the disease d_i and d_j, when the DAG of disease d_i and d_j has more overlapping nodes, their semantic similarity is higher. Therefore, the semantic similarity matrix SSD of diseases can be obtained as shown in **formula (3)**:

$$SSD(i, j) = \frac{\sum_{u \in Z(i) \cap Z(j)} \left(C_i(u) + C_j(u)\right)}{C(i) + C(j)} \tag{3}$$

SSD (i, j) is denoted as the semantic similarity value between disease d_i and d_j.

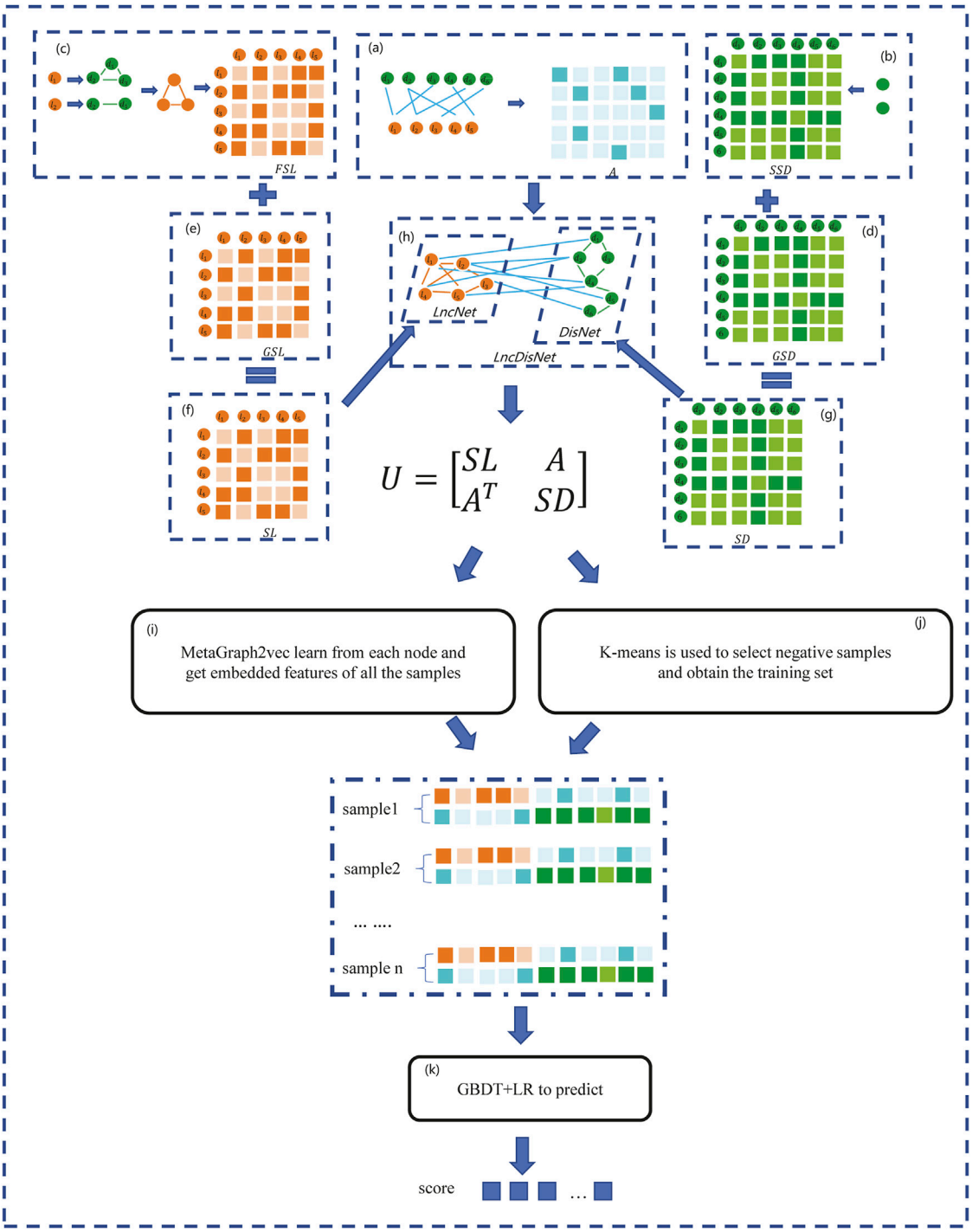


FIGURE 1 | Flowchart of the GBDTLRL2D. **(A)** Obtained the association matrix A. **(B)** Calculated the disease semantic similarity matrix SSD. **(C)** Calculated the long noncoding RNA (lncRNA) functional similarity matrix FSL. **(D)** Calculated the disease Gaussian interaction profile kernel similarity GSD. **(E)** Calculated the lncRNA Gaussian interaction profile kernel similarity GSL. **(F)** Obtained the lncRNA similarity SL. **(G)** Obtained the disease similarity SD. **(H)** Integrated three subnets A, SL, and SD to construct a global heterogeneous network. **(I)** Obtained the embedded features of nodes. **(J)** Selected the negative sample and obtained the training set. **(K)** Trained the Gradient Boosting Decision Tree combined with logistic regression classifier (GBDT + LR).

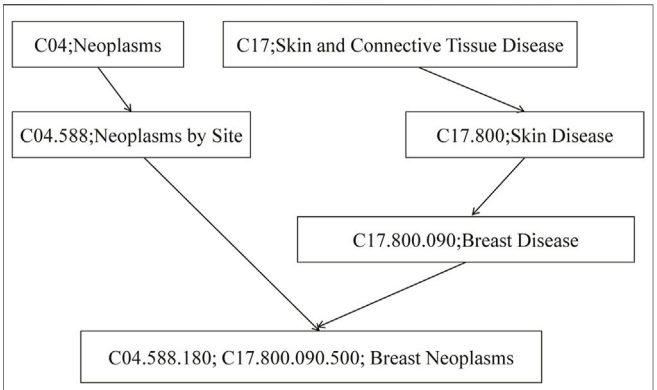


FIGURE 2 | The disease DAG of breast neoplasms. DAG, directed acyclic graph.

2.2 Calculate Long Noncoding RNA Functional Similarity

According to the LNCSIM (Chen et al., 2015), the lncRNA functional similarity is described as follows. Diseases associated with the same lncRNA are grouped into a set. The DL1 is the disease set related to lncRNA l_m, including x diseases. The DL2 is the disease set related to lncRNA l_n, including y diseases. When the semantic similarity between diseases in DS1 and DS2 is higher, the functional similarity between lncRNA l_m and l_n may be higher, as shown in **formula (4)**:

$$\mathrm{FSL}(l_m, l_n) = \frac{\sum_{d \in \mathrm{DL2}} maxS(d, \mathrm{DL1}) + \sum_{d \in \mathrm{DL1}} maxS(d, \mathrm{DL2})}{x + y} \quad (4)$$

$$\mathrm{maxS}(d, \mathrm{DL1}) = \max_{d \in DL1}(\mathrm{SS}(d, d_1)) \quad (5)$$

where maxS (d, DL1 (l_m)) is the maximum semantic similarity of all diseases in the set DL1 related to lncRNA l_m.

2.3 Calculate Gaussian Interaction Profile Kernel Similarity

In this section, the adjacency matrix A is constructed according to the known lncRNA–disease association. The A (l_i, d_j) indicates whether lncRNA l_i and disease d_j are related. The A as shown in **formula (6)**:

$$A(l_i, d_j) = \begin{cases} 1 & l_i \text{ is associated with } d_j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

In lncRNA functional similarity matrix FSL and the disease semantic similarity matrix SSD, many elements are 0. In order to make the similarity network not sparse, Gaussian kernel similarity is calculated, as shown in **formula (7)**:

$$\mathrm{GSL}(l_m, l_n) = \exp(-\delta_l \|\mathrm{A}(m, :) - \mathrm{A}(n, :)\|^2) \quad (7)$$

$$\mathrm{GSD}(d_i, d_j) = \exp(-\delta_d \|\mathrm{A}(:, i) - \mathrm{A}(:, j)\|^2) \quad (8)$$

where GSL (l_m, l_n) is the Gaussian interaction profile kernel similarity score of lncRNA l_m and l_n, and GSD (d_i, d_j) is the Gaussian interaction profile kernel similarity of diseases d_i and d_j.

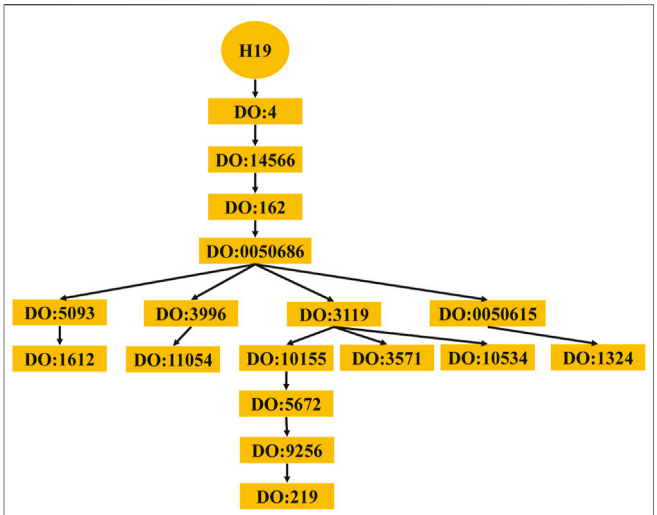


FIGURE 3 | Diseases associated with lncRNA H19. lncRNA, long noncoding RNA.

The A (m,) is mth row of A, and A (. i) is ith col of A. Parameters δ_l and δ_d are obtained as shown in **Eq. 9** and **Eq. 10**:

$$\delta_l = \delta_l' / \left(\frac{1}{p} \sum_{m=1}^{p} \|\mathrm{A}(m, :)\|^2 \right) \quad (9)$$

$$\delta_d = \delta_d' / \left(\frac{1}{q} \sum_{i=1}^{q} \|\mathrm{A}(:, i)\|^2 \right) \quad (10)$$

where p is the number of lncRNAs and q is the number of diseases.

2.4 Obtain Similarity Network

In this section, lncRNA similarity network and disease similarity network are constructed. The lncRNA similarity network is represented as SL. SL is fused by FSL and GSL. For lncRNA l_m and l_n, if FSL (l_m, l_n) = 0, then SL (l_m, l_n) = GSL (l_m, l_n); otherwise, SL (l_m, l_n) = FSL (l_m, l_n), as shown in **formula (11)**:

$$\mathrm{SL}(l_m, l_n) = \begin{cases} \mathrm{GSL}(l_m, l_n) & \text{if } \mathrm{FSL}(l_m, l_n) = 0 \\ \mathrm{FSL}(l_m, l_n) & \text{otherwise} \end{cases} \quad (11)$$

$$\mathrm{SD}(d_i, d_j) = \begin{cases} \mathrm{GSD}(d_i, d_j) & \text{if } \mathrm{SSD}(d_i, d_j) = 0 \\ \mathrm{SSD}(d_i, d_j) & \text{otherwise} \end{cases} \quad (12)$$

Similarly, the disease similarity network is expressed as SD. The SD is fused by SSD and GSD, as shown in **formula (12)**:

2.5 Obtain Node Features Through MetaGraph2Vec

In this section, the heterogeneous network is constructed by integrating the adjacency matrix A of lncRNA–disease association, lncRNA similarity network SL, and the disease similarity network SD. Because there are many types of nodes and complex relationships among nodes in the heterogeneous

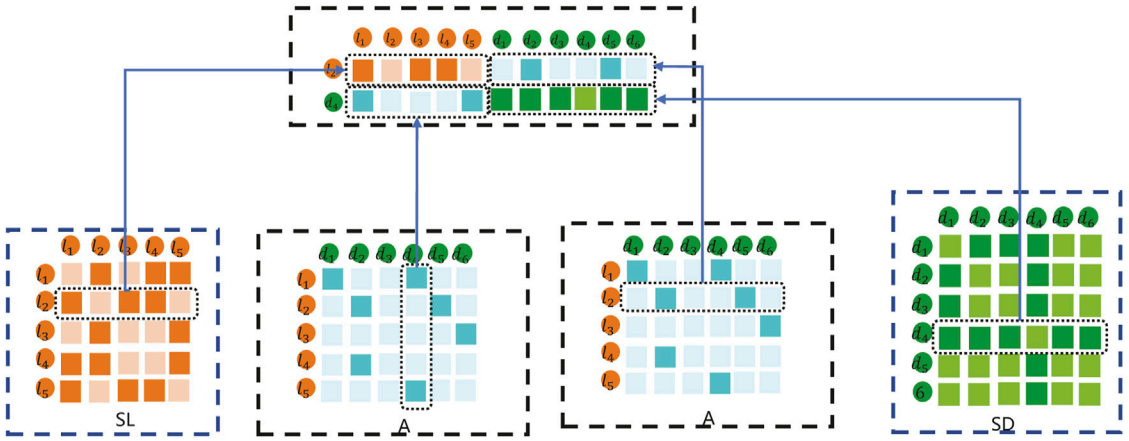


FIGURE 4 | An example of K-means embedding matrix.

network, embedding networks is difficult to implement. In GBDTLRL2D algorithm, MetaGraph2Vec (Zhang et al., 2018) is used for feature learning, which preserves both structural and semantic features in heterogeneous networks. MetaGraph2Vec uses metagraph to guide the generation of a random walk and learn about the potential embedding of nodes in a heterogeneous network of multiple types. Metagraph can represent features of sparse heterogeneous networks based on meta-paths. The specific substeps are shown below.

2.5.1 Building Heterogeneous Networks

The heterogeneous network G = (*V*, *E*) is constructed, where *V* represents the set of nodes and *E* represents the set of edges. The G is denoted as U. The dimension of the matrix U is (*nl* + *nd*)*(*nl* + *nd*), where *nl* is the number of lncRNAs and *nd* is the number of diseases, as shown in **formula (13)**:

$$U = \begin{bmatrix} SL & A \\ A^T & SD \end{bmatrix} \tag{13}$$

where A^T is the transpose of A.

2.5.2 Random Walk Guided by metagraph

Given a metaGraph $g = (N, M, n_s, n_t)$ with $n_s = n_t$, the metaGraph is defined as a DAG on G, where n_s represents the source node, n_t represents the target node, *N* is the node set, and *M* is the edge set. The $g^{\infty} = (N, M, n_s^{\infty}, n_t^{\infty})$ is recursive metaGraph of *g*. The g^{∞} is constructed by any number of *g*'s connected tail to head. A node of type n_s is selected to start a random walk guided by the metatype.

In step *i*, node v_{i-1} is selected as the start of a random walk guided by the metaGraph. The random walk obtains the edge types of node v_{i-1} in a heterogeneous network that meets the constraints in the metagraph with all neighboring nodes. One edge type is randomly selected. Then, an edge of the selected edge type is randomly selected to get the next node v_i. The random walk terminates when there is no edge type that satisfies the constraint.

The transition probability of step *i* guided by metaGraph *g* is denoted as T $(v_i|v_{i-1}; g^{\infty})$, v_{i-1} is the current node, and v_i is the next hop node. If the node v_{i-1} and its neighbors in the heterogeneous network G do not satisfy the edge type of the constraint of the g^{∞}, T $(v_i|v_{i-1}; g^{\infty}) = 0$. Otherwise, T $(v_i|v_{i-1}; g^{\infty})$ is shown in **formula (14)**:

$$T(v_i \mid v_{i-1}; g^{\infty}) = \frac{1}{NUM_{g^{\infty}}(v_{i-1})} \times \frac{1}{\left|\{\mu \mid (v_{i-1}, \mu) \in E, \phi(v_i) = \phi(\mu)\}\right|} \tag{14}$$

where $|\{\mu \mid (v_{i-1}, \mu) \in E, \phi(v_i) = \phi(\mu)\}|$ is the number of neighbor nodes of the same type as v_{i-1}. $NUM_{g^{\infty}}(v_{i-1})$ is the number of edge types that satisfy the constraint in the g^{∞} starting from v_{i-1}, as shown in **formula (15)**:

$$NUM_{g^{\infty}}(v_{i-1}) = \left|\left\{j \middle| (\phi(v_{i-1}), \phi(\mu)) \cdot \in M \cap (N[l^{\infty}(\phi(v_{i-1}))] \times N[j]), \middle| (v_{i-1}, \mu) \in E\right\}\right| \tag{15}$$

After several walks, a node sequence $S_g = v_1, v_2, \ldots, v_L$ of length L is finally obtained.

2.5.3 Obtain Node Features Through MetaGraph2Vec

By learning the mapping function Ψ, the nodes of heterogeneous networks are embedded into a *d*-dimensional space to obtain the embedding feature. The network G has a large number of nodes with different semantics. The nodes with similar semantics in heterogeneous networks are guaranteed to have similar low-dimensional representations $\Psi(v)$.

The node sequence $S_g = v_1, v_2, \ldots, v_L$ of length L is obtained by a random walk guided by metaGraph *g*. The embedding function $\Psi(\cdot)$ is learned by maximizing the occurrence probability of nodes before and after v_i in the window, and the window size is *b*. $\Psi(\cdot)$ is shown in **formula (16)**:

$$\min_{\Psi} -\log T(\{v_{i-b}, \ldots, v_{i+b}\} / v_i \mid \Psi(v_i)) \tag{16}$$

where $T(\{v_{i-b}, \ldots, v_{i+b}\} / v_i \mid \Psi(v_i)) = \prod_{j=i-b, j\neq i}^{i+b} T(v_j \mid \Psi(v_i))$.

Following MetaPath2Vec, the $T(v_j|\Psi(v_i))$ is related to the type of v_j, as shown in **formula (17)**:

$$\mathrm{T}\left(v_j \mid \Psi(v_i)\right) = \mathrm{T}\left(v_j \mid \Psi(v_i), \psi\left(v_j\right)\right)\mathrm{T}\left(\psi\left(v_j\right) \mid \Psi(v_i)\right) \quad (17)$$

where the probability $\mathrm{T}(v_j \mid \Psi(v_i), \psi(v_j))$ is shown in **formula (18)**:

$$\mathrm{T}\left(v_j \mid \Psi(v_i), \psi\left(v_j\right)\right) = \frac{\exp\left(\Phi\left(v_j\right) \cdot \Psi(v_i)\right)}{\sum_{\mu \in V, \psi(\mu)=\psi\left(v_j\right)} \exp\left(\Phi'(\mu) \cdot \Psi(v_i)\right)} \quad (18)$$

After that, stochastic gradient descent is used to learn the parameters. At each iteration, a node context pair (v_i, v_j) is sampled according to the distribution of P (v_i, v_j), and the P (v_i, v_j) is the occurrence frequency of each node context pair (v_i, v_j) within b window size. The gradients are updated to minimize the following objective:

$$\mathcal{O}_{ij} = -\log T\left(v_j \mid \Psi(v_i)\right) \quad (19)$$

To speed up training, negative sampling is used to approximate the objective function:

$$\mathcal{O}_{ij} = \log \rho\left(\Phi\left(v_j\right) \cdot \Psi(v_i)\right) + \sum_{u=1}^{U} \log \rho\left(-\Phi\left(v_{N_{j,u}}\right) \cdot \Psi(v_i)\right) \quad (20)$$

where $\rho(\cdot)$ is the sigmoid function, $v_{N_{j,u}}$ is the uth negative node sampled for node v_j, and U is the number of negative samples, $v_{N_{j,u}}$ sampled from nodes with type $\Psi(v_j)$. Formally, parameters Ψ and Φ are updated as follows:

$$\Psi = \Psi - \lambda \frac{\partial \mathcal{O}_{ij}}{\partial \Psi}; \Phi = \Phi - \lambda \frac{\partial \mathcal{O}_{ij}}{\partial \Phi} \quad (21)$$

where λ is the learning rate.

The embedding function Ψ embeds the nodes of a heterogeneous network into a low-dimensional space, embedding each node and obtaining a low-dimensional representation $\Phi(v)$. Finally, the d-dimensional matrix X is obtained.

2.6 Obtain Negative Samples Using K-Means Clustering

Since the number of negative samples is far greater than that of positive samples in the set, it is necessary to balance the train set. The proposed method uses a novel and advanced data balancing method K-means clustering. The K-means is a segmentation technique based on centroid. The centroid of a cluster is used to represent the cluster. The centroid of a cluster is defined as the mean value of points within the cluster. The K-means is relatively simple and easy to implement. The specific implementation steps are as follows:

1. The initial k cluster centers are randomly selected from the unknown sample.
2. According to the distance from the point to the center of each cluster, the point is assigned to the closest cluster center category.
3. All points are assigned, and k cluster centers are recalculated.
4. If the recalculated k cluster centers are not the same as the previous cluster centers, we go to 2; otherwise, go to 5.
5. The clustering center is not changed, and the clustering results are output.

In this method, the data feature input into the K-means clustering method is composed of the fusion of SL, SD, and A. For example, the embedding matrix for sample lncRNA l_2 and disease d_4 pairs is shown in **Figure 4**.

As shown in **Figure 4**, the embedding matrix of lncRNA l_2 and disease d_4 pairs includes the following parts: 1) the first part is the second row of lncRNA similarity matrix SL; 2) the second part is composed of the vector corresponding to the adjacency matrix A of d_4; 3) the third part is composed of the vector corresponding to l_2 of the adjacency matrix A; and 4) the fourth part is the second row of disease similarity matrix SD. Combined with the representations in the first part, second part, third part, and fourth part, the final lncRNA l_2 and disease d_4 samples are constructed to carry out the K-means embedding matrix.

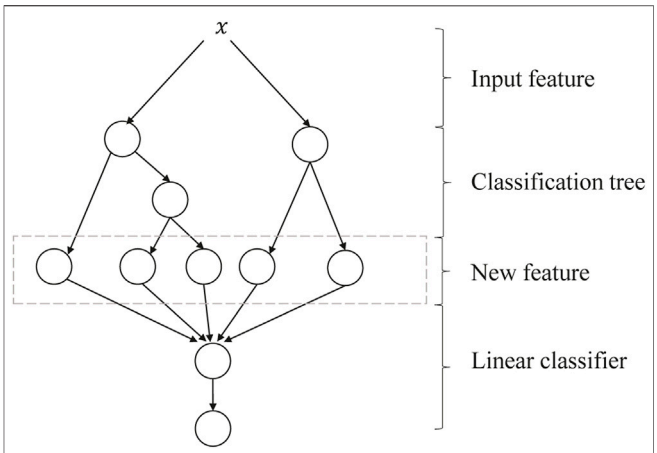


FIGURE 5 | Algorithm structure of GBDT + LR. GBDT, Gradient Boosting Decision Tree; LR, logistic regression.

2.7 Train the Gradient Boosting Decision Tree Combined With Logistic Regression Classifier

After the sample and features are obtained, the GBDT + LR classifier is trained. Parameters of the model are initialized. The training data are regressed through the GBDT model and generate a decision tree. The leaf nodes of the decision tree are combined to find the new feature. The feature is used as input to the LR classifier model. Thus, the training process of GBDT + LR classifier is completed.

GBDT + LR is a process of feature crossing, and the path of GBDT can be directly used as the input feature of LR, avoiding the process of manual combination of cross features. Its algorithm structure is shown in **Figure 5**. The two trees in the figure are regression tree models trained by GBDT. The left tree has three leaf nodes, and the right tree has two leaf nodes. The final feature is a five-dimensional vector. For input x, it is assumed that it falls

TABLE 2 | The partial experimental parameters of GBDTLRL2D.

Notation	Value	Definition
nl_1	112	The number of lncRNAs in dataset1
nd_1	150	The number of diseases in dataset1
n_1	262	Total number of diseases and lncRNAs in dataset1
nl_2	131	The number of lncRNAs in dataset2
nd_2	169	The number of diseases in dataset2
n_2	300	Total number of diseases and lncRNAs in dataset2
nl_3	285	The number of lncRNAs in dataset3
nd_3	226	The number of diseases in dataset3
n_3	511	Total number of diseases and lncRNAs in dataset3
γ_l'	1	Gaussian interaction properties of lncRNA kernel similar bandwidth
γ_d'	1	Gaussian interaction properties of lncRNA kernel similar bandwidth
k	10	K-means clustering divides the unknown samples into k clusters
K	5	The number of negative samples taken in MetaGraph2Vec

Note. lncRNA, long noncoding RNA.

TABLE 3 | Comparison of prediction performance using other machine learning methods.

Dataset	Method	ACC	Recall	$F1_{score}$	MCC	AUC
DS1	GBDT + LR	0.928	0.920	0.927	0.858	0.975
DS2		0.934	0.928	0.934	0.870	0.982
DS3		0.887	0.871	0.885	0.777	0.961
DS1	RF + LR	0.787	0.767	0.780	0.581	0.880
DS2		0.800	0.802	0.801	0.603	0.898
DS3		0.796	0.767	0.790	0.601	0.889
DS1	GBDT	0.570	0.658	0.608	0.125	0.619
DS2		0.600	0.724	0.645	0.210	0.654
DS3		0.636	0.631	0.636	0.282	0.647
DS1	LR	0.570	0.659	0.609	0.125	0.649
DS2		0.601	0.724	0.645	0.211	0.705
DS3		0.636	0.631	0.636	0.282	0.667

Note. lncRNA, long noncoding RNA; ACC, Accuracy; MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve; GBDT, Gradient Boosting Decision Tree; LR, logistic regression; RF, random forest.

on the first node of the left tree and encodes (1, 0, 0); if it falls on the second node of the right tree, it encodes (0, 1), so the overall code is (1, 0, 0, 0, 1). Such codes are input into LR for classification. The steps of GBDT + LR for the algorithm are as follows:

Step 1) The original training data are trained with GDBT to generate a decision tree, and grid search is used to find the best parameter combination.

A: The initialization parameter of GDBT is shown in **formula (22)**:

$$\Theta_0(x) = \frac{1}{2} * \log\left(\frac{\sum_{i=1}^{NUM} y_i}{\sum_{i=1}^{NUM} 1 - y_i}\right) \tag{22}$$

There are NUM samples to be trained, and y_i is the label for sample i. The loss function J (y, $\Theta_t(x)$) is defined as shown in **formula (23)**:

$$J\left(y, \Theta_t(x)\right) = \log\left(1 + \exp\left(-y\Theta_t(x)\right)\right) \tag{23}$$

where y is the label and $\Theta_t(x)$ is the weak model in the tth round.

B: Cycle t, in turn, where $t = 1, 2, \ldots T$.

1) The negative gradient of the loss function of sample ith in wheel tth is calculated, as follows:

$$r_{t,i} = -\frac{\partial J\left(y_i, \Theta_{t-1}(x_i)\right)}{\partial \Theta_{t-1}(x_i)} = \frac{y_i}{\left(1 + \exp\left(y_i\right)\Theta(x_i)\right)}, i = 1, 2, \ldots, NUM \tag{24}$$

where $i = 1, 2, 3, \ldots, NUM$.

2) Construct the tth decision tree, and then get the corresponding leaf node area as R_{tn}, where $n = 1, 2, \ldots, N$. N is the number of leaf nodes of the tree.

3) For the samples in each leaf node, we calculated the c_{tn}, which minimizes the loss function, as shown in **formula (25)**:

$$c_{tn} = \arg\min_c \sum_{x \in R_{tn}} \log\left(1 + \exp\left(-y_i\Theta(x_i) + c\right)\right) \tag{25}$$

4) Update the tth weak model as shown in **formula (26)**:

$$\Theta_t(x) = \Theta_{t-1}(x) + \alpha * \sum_{n=1}^{N} c_{tn} I(x \in R_{tn}) \tag{26}$$

where $I(x \in R_{tn})$ means that if x falls on a leaf node corresponding to R_{tn}, then the corresponding term is 1, and α means the learning rate.

5) Determine whether t is greater than T. If t is less than T, $t = t + 1$ and jump to 1) for the next iteration. Otherwise, it means that all T weak learners have been constructed and jump to **C** to end the training.

C: The final strong learner model is shown in **formula (27)**:

$$\Theta(x) = \Theta_0(x) + \alpha * \sum_{t=1}^{T} \sum_{n=1}^{N} c_{tn} I(x \in R_{tn}) \tag{27}$$

where α is the learning rate.

Step 2) After the training of GDBT, for each tree in the model, the calculated probability value of the leaf node is denoted as 1, and new training data are constructed. In this paper, One-Hot Encoding is used to process the results of GDBT and construct a new training dataset.

One-hot Encoding is also known as one-bit Efficient coding. Single-hot coding encodes N states by using an N-bit status

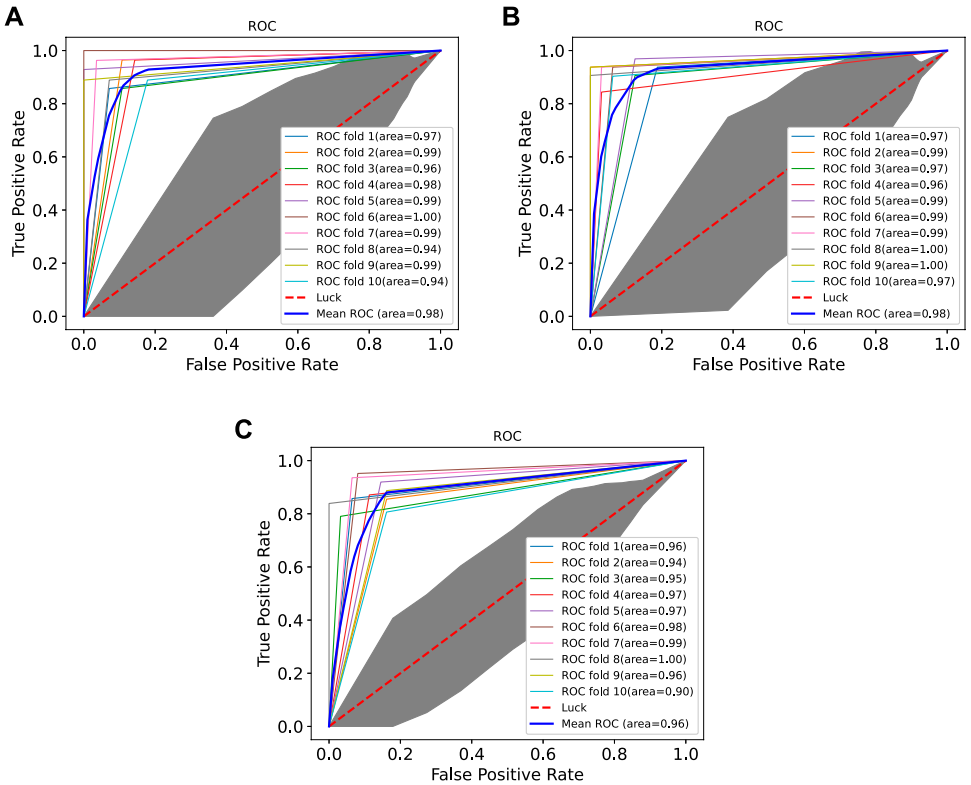


FIGURE 6 | ROC curve of GBDTLRL2D on three datasets. ROC, receiver operating characteristic.

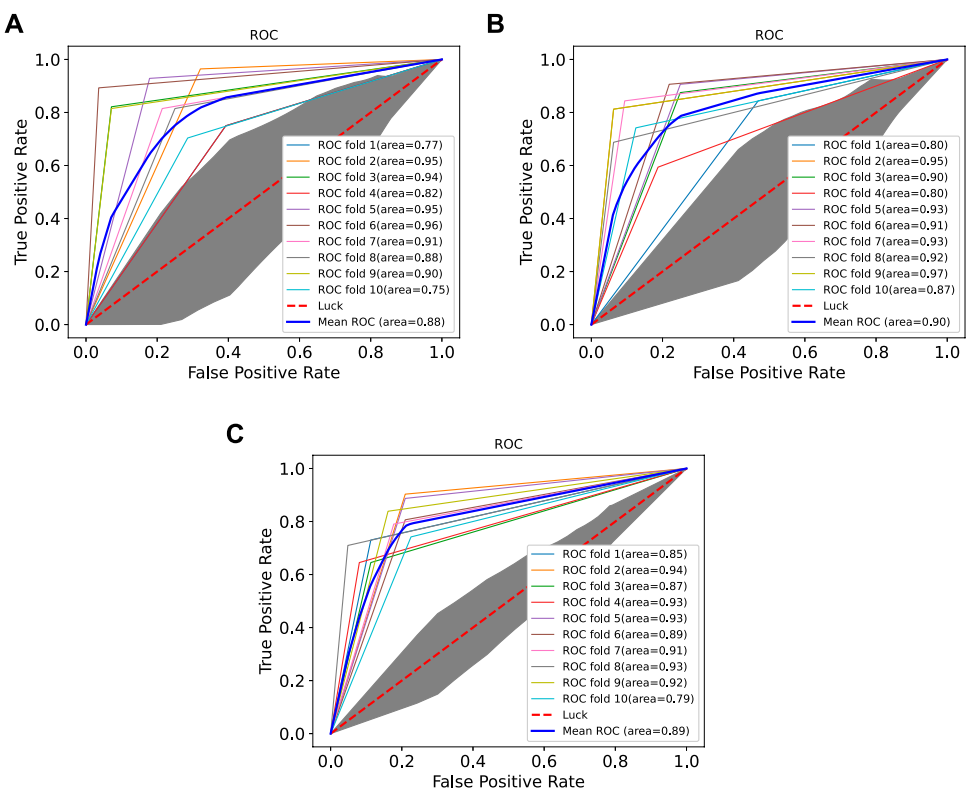


FIGURE 7 | ROC curves on three datasets using RF + LR. ROC, receiver operating characteristic; RF, random forest; LR, logistic regression.

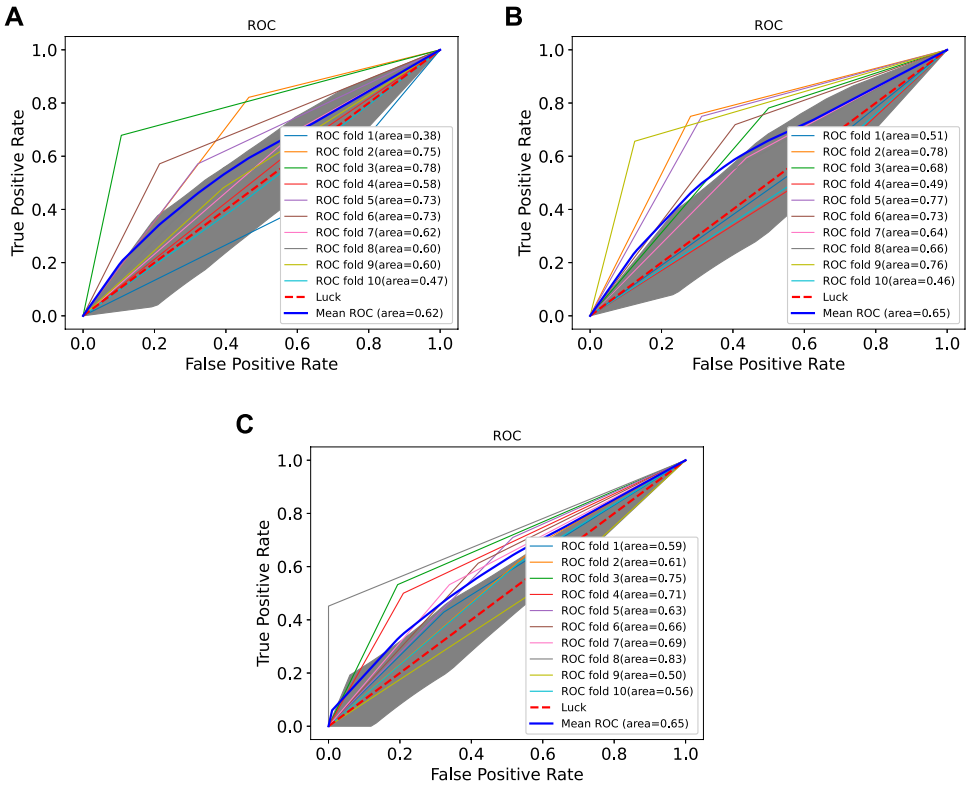


FIGURE 8 | ROC curves on three datasets using GBDT. ROC, receiver operating characteristic.

register. Each of the N states has its own independent register bit, and only one is valid at any time such as the following.

General status Encoder: 000, 001, 010, 011, 100, 101

One-Hot Encoder: 000 001, 000 010, 000 100, 001 000, 010 000, 100 000

Step 3) The new features obtained and the label data of the original training data are input into the LR classifier for the training. The hypothesis function of LR is shown in **formula (28)**. Given x and θ, the possibility that x belongs to a positive sample is shown in **formula (29)**. θ is obtained by training to minimize the loss function in **formula (30)**.

$$h_\theta(x) = g\left(\theta^T x\right), g(z) = \frac{1}{1+e^{-z}} \tag{28}$$

$$Pr(y = 1 \mid x; \theta) = h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \tag{29}$$

$$L(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right] \tag{30}$$

3 RESULT AND DISCUSSION

3.1 Dataset

The data are downloaded from the lncRNA–disease-associated data from the lncRNADisease database, including the data of three versions, namely, the version of June 2012, the version of January 2014, and the version of June 2015, labeled as DS1, DS2, and DS3, respectively. The training samples are obtained by all positive samples and randomly selecting negative samples by K-means clustering.

3.2 Performance Measures

In this paper, the 10-fold cross-validation is selected to measure the performance of the proposed method. The parameters of GBDTLRL2D are shown in **Table 2**. The main steps of 10-fold cross-validation are as follows: the training set is randomly divided into 10 subsets of the same size, nine of which are used as training data, and the remaining one is used as validation data in each training. After ten times of the above process training, each of the ten subsets, in turn, is used as validation data to obtain ten performance results. The final performance evaluation is obtained by averaging the ten performance results. Various evaluation indexes are used in this experiment, including Recall (REC), F1-score, Accuracy (ACC), Matthews correlation coefficient (MCC), and area under the receiver operating characteristic (ROC) curve (AUC). Their definition is as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{31}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{32}$$

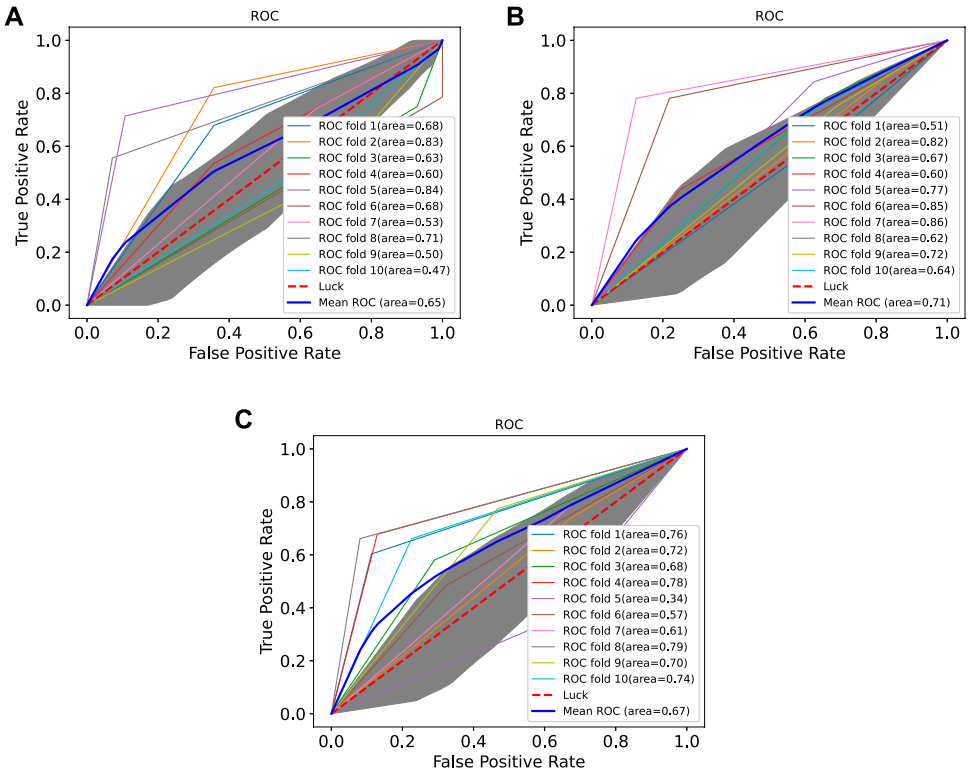


FIGURE 9 | ROC curves on three datasets using LR. ROC, receiver operating characteristic; LR, logistic regression.

$$\mathrm{F1}_{\mathrm{score}} = \frac{2*TP}{2TP + FP + FN} \tag{33}$$

$$\mathrm{MCC} = \frac{TP*TN - FP*FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{34}$$

where TP represents the number of correct prediction of positive samples as positive samples, TN represents the number of correct prediction of negative samples as negative samples, FN represents the number of incorrect prediction of positive samples as negative samples, and FP represents the number of prediction of negative samples as positive samples. We plot the ROC based on the true positive rate (TPR) and false positive rate (FPR), and we calculate the AUC as an important index to measure the model.

3.3 Performance Comparison With Existing Machine Learning Methods

In order to prove the advantages of GBDT combined with LR classifier, we carried out several experiments to compare with GBDTLRL2D, including using RF + LR as the classifier, using GBDT only as the classifier, and using LR only as a classifier. It can be seen that GBDTLRL2D obtains the best performance among these methods. The 10-fold cross-validation is selected to measure the performance of the proposed method. **Table 3** shows the predictive performance of GBDT + LR compared with other methods. The ROC curves of 10-fold cross-validation of GBDTLRL2D, RF + LR, GBDT, and LR are shown in **Figures**

TABLE 4 | Performance comparison of representation learning without MetaGraph2vec.

Dataset	Whether to use MetaGraph2Vec	ACC	Recall	$F1_{score}$	MCC	AUC
DS1	Yes	0.928	0.920	0.927	0.858	0.975
DS2		0.934	0.928	0.934	0.870	0.982
DS3		0.887	0.871	0.885	0.777	0.961
DS1	No	0.773	0.785	0.779	0.555	0.871
DS2		0.786	0.758	0.778	0.581	0.877
DS3		0.829	0.826	0.829	0.667	0.923

Note. lncRNA, long noncoding RNA; ACC, Accuracy; MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve.

TABLE 5 | Performance comparison without K-means.

Dataset	Whether to use K-means	ACC	Recall	$F1_{score}$	MCC	AUC
DS1	Yes	0.928	0.920	0.927	0.858	0.975
DS2		0.934	0.928	0.934	0.870	0.982
DS3		0.887	0.871	0.885	0.777	0.961
DS1	No	0.802	0.713	0.778	0.617	0.888
DS2		0.769	0.705	0.745	0.553	0.871
DS3		0.779	0.726	0.763	0.562	0.876

Note. ACC, Accuracy; MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve.

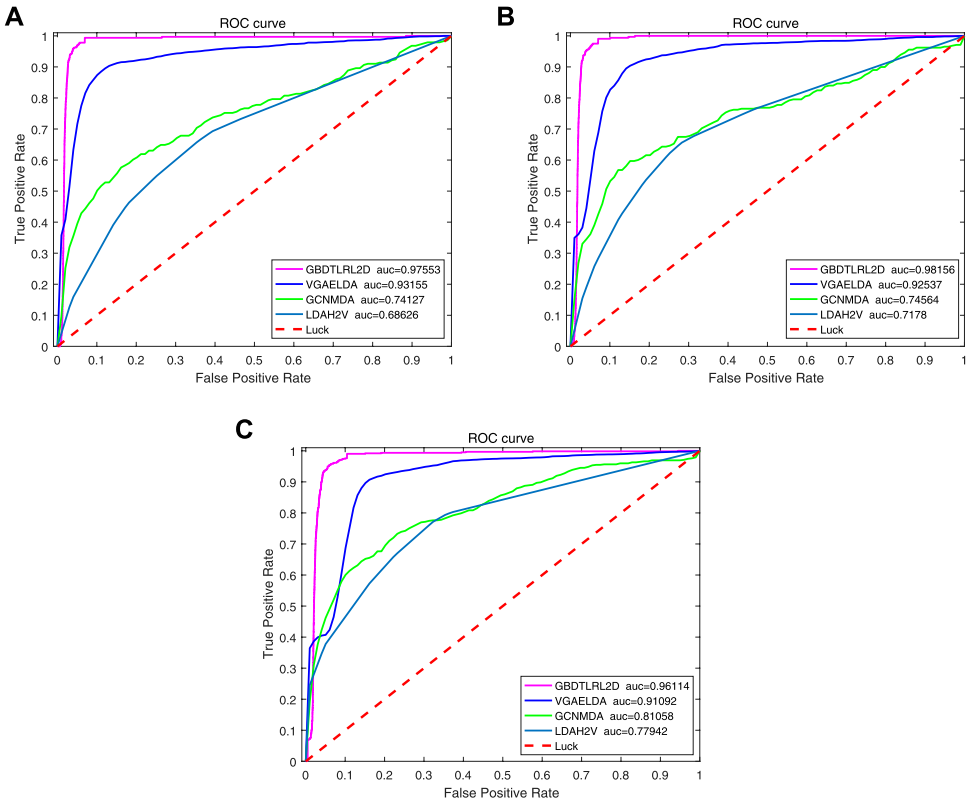


FIGURE 10 | ROC curve comparison with existing methods. ROC, receiver operating characteristic.

TABLE 6 | The top 10 predicted diseases related to "PVT1."

Rank	Disease	Score
1	Lymphoma	0.999 169 371
2	Cancer	0.998 948 531
3	Breast cancer	0.998 948 531
4	Prostate cancer	0.998 948 531
5	Ovarian cancer	0.998 948 531
6	Type 2 diabetes	0.995 265 292
7	Type 1 diabetes	0.995 265 292
8	Diabetic nephropathy	0.987 846 199
9	Hodgkin's lymphoma	0.984 907 726
10	Burkitt's lymphomas	0.983 042 458

6–9, respectively. The AUCs of GBDTLRL2D, RF + LR, GDBBT, and LR in DS1 are 0.976, 0.880, 0.619, and 0.649, respectively. The AUCs of GBDTLRL2D, RF + LR, GDBBT, and LR in DS2 are 0.983, 0.898, 0.654, and 0.705, respectively. The AUCs of GBDTLRL2D, RF + LR, GDBBT, and LR in DS3 are 0.961, 0.889, 0.647, and 0.667, respectively. It can be seen that GBDTLRL2D obtains the best performance among these methods.

3.4 Performance Comparison With Different Topological Features

In order to demonstrate the performance of the experimental features, different feature groups (not using MetaGraph2Vec for representation learning, but using MetaGraph2Vec for representation learning) and different negative samples (not using K-means for clustering, but using K-means for clustering) are used for performance comparison in this section. **Table 4** and **Table 5** show the performance comparison with different topological features. In **Table 4**, on the same dataset, the result shows that the features obtained through MetaGraph2Vec embedding learning are trained to achieve better performance. Similarly, in **Table 5**, the performance of negative samples obtained through K-means cluster screening is better than that of negative samples randomly selected for training.

3.5 Performance Comparison With Existing Methods

To further illustrate the advantages of the proposed model, several existing methods based on embedding are compared with GBDTLRL2D, such as LDAH2V, VGAELDA (Shi et al., 2021), and GCNMDA (Long et al., 2020). The 10-fold cross-validation is selected to measure the performance.

LDAH2V: The LDAH2V uses the HIN2Vec to calculate the meta-path and feature vector for each lncRNA–disease pair in the heterogeneous information network (HIN), which consists of lncRNA similarity network, disease similarity network, miRNA similarity network, and the associations between them. Then, a Gradient Boosting Tree (GBT) classifier to

predict lncRNA–disease associations is built with the feature vectors.

VGAELDA: The VGAELDA integrates graph embedding learning and the alternate training via variational inference. Variational graph autoencoders (VGAEs) infer representations from features of lncRNAs and diseases, while graph autoencoders propagate labels via known lncRNA–disease associations. These two kinds of autoencoders are trained alternately by adopting variational expectation–maximization algorithm.

GCNMDA: The graph convolution network is used for network embedding in GCNMDA. The GCNMDA exploited the Conditional Random Field (CRF), which can ensure that similar nodes have similar representations. At the same time, the attention mechanism is designed in CRF layer.

Figure 10 shows the comparison results. Among these methods, the proposed model GBDTLRL2D achieves the best performance. There are several reasons: 1) the features learned by MetaGraph2Vec can better preserve node information and semantic information in a heterogeneous information network. 2) K-means clustering is used to select more representative negative samples. 3) The GBDTLRL2D uses the combined machine learning method of GBDT + LR with good performance to make predictions.

Despite that our method is obviously superior to previous methods in all aspects, there are some limitations to GBDTLRL2D. The number of lncRNA–disease associations confirmed by biological experimental methods is limited. In addition, it is important to select classifiers. Currently, GBDT + LR is the best classifier for our model. In the future, we will have to try to combine other classifiers to achieve more accurate predictions.

3.6 Case Study

In this section, to further show the performance of the proposed model GBDTLRL2D in predicting the lncRNA–disease association, a case study is conducted on lncRNA "PVT1." A proven association between "PVT1" and many diseases has been found in biology. In this paper, the proposed model GBDTLRL2D is used to predict the association between "PVT1" and disease. After processing by our algorithm, the list of diseases associated with lncRNA "PVT1" and their predicted scores is obtained. Ranking the diseases according to the predicted score from large to small, we can find that all the diseases in the top 10 associated with lncRNA "PVT1" are confirmed to be associated with "PVT1" in the lncRNADisease database. The top 10 diseases associated with lncRNA "PVT1" and their predicted scores are shown in **Table 6**.

CONCLUSIONS

LncRNAs have been found by biologists to be closely related to diseases. Predicting the lncRNA–disease associations is conducive to research on the pathogenesis of a disease. But traditional biological methods have a large amount of data and are expensive, labor-intensive, and time-consuming. In recent years, there has been much research on computational models of biological experiments. In this paper, a method for predicting lncRNA–disease association is proposed. The proposed method uses MetaGraph2Vec to learn the features of nodes in a heterogeneous network and then uses K-means to select representative negative samples to solve the problem of imbalance between positive and negative samples, and the GBDT combined with LR is used as a classifier to predict lncRNA–disease associations. At last, the average AUCs of GBDTLRL2D obtained on the three datasets are 0.98, 0.98, and 0.96 in 10-fold cross-validation. Compared with the SIMCLDA, IIRWR, NCPLDA, and other experiments, the GBDTLRL2D greatly improves accuracy and performance.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: http://www.cuilab.cn/lncrnadisease.

AUTHOR CONTRIBUTIONS

TD, ZK, JW, and ZM conceived this work and designed the experiments. TD, JW, and ZK carried out the experiments. TD and ZM collected the data and analyzed the results. TD and ZK wrote, revised, and approved the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62072477, 61309027, 61702562, and 61702561; the Hunan Provincial Natural Science Foundation of China under Grant No. 2018JJ3888; the Scientific Research Fund of Hunan Provincial Education Department under Grant No. 18B197; the National Key R&D Program of China under Grant No. 2018YFB1700200; the Open Research Project of Key Laboratory of Intelligent Information Perception and Processing Technology (Hunan Province) under Grant No. 2017KF01; and the Hunan Key Laboratory of Intelligent Logistics Technology 2019TP1015.

ACKNOWLEDGMENTS

We would like to thank the Experimental Center of School of Computer and Information Engineering, Central South University of Forestry and Technology, for providing computing resources.

REFERENCES

Abdi, E., Latifi-Navid, S., Latifi-Navid, H., and Safaralizadeh, R. (2021). Lncrna Polymorphisms and Upper Gastrointestinal Cancer Risk. *Pathol. - Res. Pract.* 218, 153324. doi:10.1016/j.prp.2020.153324

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., et al. (2011). A Long Noncoding Rna Controls Muscle Differentiation by Functioning as a Competing Endogenous Rna. *Cell* 147, 358–369. doi:10.1016/j.cell.2011.09.028

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). Lncrnadisease: a Database for Long-Non-Coding Rna-Associated Diseases. *Nucleic Acids Res.* 41, D983–D986. doi:10.1093/nar/gks1099

Chen, X., Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing Lncrna Functional Similarity Network Based on Lncrna-Disease Associations and Disease Semantic Similarity. *Sci. Rep.* 5, 11338–11412. doi:10.1038/srep11338

Chen, X., and Yan, G.-Y. (2013). Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics* 29, 2617–2624. doi:10.1093/bioinformatics/btt426

Cui, Z., Liu, J. X., Gao, Y. L., Zhu, R., and Yuan, S. S. (2019). Lncrna-disease Associations Prediction Using Bipartite Local Model with Nearest Profile-Based Association Inferring. *IEEE J. Biomed. Health Inform.* 24, 1519–1527. doi:10.1109/JBHI.2019.2937827

Deng, L., Li, W., and Zhang, J. (2021). Ldah2v: Exploring Meta-Paths across Multiple Networks for Lncrna-Disease Association Prediction. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18, 1572–1581. doi:10.1109/TCBB.2019.2946257

Gao, M.-M., Cui, Z., Gao, Y.-L., Wang, J., and Liu, J.-X. (2021). Multi-label Fusion Collaborative Matrix Factorization for Predicting Lncrna-Disease Associations. *IEEE J. Biomed. Health Inform.* 25, 881–890. doi:10.1109/JBHI.2020.2988720

Ge, X., Chen, Y., Liao, X., Liu, D., Li, F., Ruan, H., et al. (2013). Overexpression of Long Noncoding Rna Pcat-1 Is a Novel Biomarker of Poor Prognosis in Patients with Colorectal Cancer. *Med. Oncol.* 30, 588. doi:10.1007/s12032-013-0588-6

Hou, M., Wu, N., and Yao, L. (2021). LncRNA CBR3-AS1 Potentiates Wnt/β-Catenin Signaling to Regulate Lung Adenocarcinoma Cells Proliferation, Migration and Invasion. *Cancer Cel Int* 21, 36–12. doi:10.1186/s12935-020-01685-y

Kuang, L., Zhao, H., Wang, L., Xuan, Z., and Pei, T. (2019). A Novel Approach Based on point Cut Set to Predict Associations of Diseases and Lncrnas. *Cbio* 14, 333–343. doi:10.2174/1574893613666181026122045

Li, J., Zhang, C., Shi, Y., Li, Q., Li, N., and Mi, Y. (2021). Identification of Key Lncrnas and Mrnas Associated with Oral Squamous Cell Carcinoma Progression. *Cbio* 16, 207–215. doi:10.2174/1573411016999200729125745

Liu, Z.-P. (2020). Predicting Lncrna-Protein Interactions by Machine Learning Methods: a Review. *Curr. Bioinformatics* 15, 831–840.

Long, Y., Wu, M., Kwoh, C. K., Luo, J., and Li, X. (2020). Predicting Human Microbe-Drug Associations via Graph Convolutional Network with Conditional Random Field. *Bioinformatics* 36, 4918–4927. doi:10.1093/bioinformatics/btaa598

Maass, P. G., Luft, F. C., and Bähring, S. (2014). Long Non-coding Rna in Health and Disease. *J. Mol. Med.* 92, 337–346. doi:10.1007/s00109-014-1131-8

Mercer, T. R., and Mattick, J. S. (2013). Structure and Function of Long Noncoding Rnas in Epigenetic Regulation. *Nat. Struct. Mol. Biol.* 20, 300–307. doi:10.1038/nsmb.2480

Shi, Z., Zhang, H., Jin, C., Quan, X., and Yin, Y. (2021). A Representation Learning Model Based on Variational Inference and Graph Autoencoder for Predicting Lncrna-Disease Associations. *BMC bioinformatics* 22, 1–20. doi:10.1186/s12859-021-04073-z

Silva, A. B. O. V., and Spinosa, E. J. (2021). Graph Convolutional Auto-Encoders for Predicting Novel Lncrna-Disease Associations. *Ieee/acm Trans. Comput. Biol. Bioinf.* 1, 1. doi:10.1109/TCBB.2021.3070910

Song, X.-Y., Liu, T., Qiu, Z.-Y., You, Z.-H., Sun, Y., Jin, L.-T., et al. (2020). Prediction of Lncrna-Disease Associations from Heterogeneous Information Network Based on Deepwalk Embedding Model. In International Conference on Intelligent Computing. Springer, 291–300. doi:10.1007/978-3-030-60796-8_25

Sun, Y., Zhao, H., Zhou, G., Guan, T., Wang, Y., and Gao, J. (2021). Random Distributed Logistic Regression Framework for Predicting Potential lncRNA–disease Association. *J. Mol. Cel Biol.* 13, 386–388. doi:10.1093/jmcb/mjab005

Tian, L., and Wang, S.-L. (2021). Exploring Mirna Sponge Networks of Breast Cancer by Combining Mirna-Disease-Lncrna and Mirna-Target Networks. *Cbio* 16, 385–394. doi:10.2174/1574893615999200711171530

Wang, B., Zhang, C., Du, X.-x., and Zhang, J.-f. (2021a). Lncrna-Disease Association Prediction Based on Weight Matrix and Projection Score. *BMC Bioinformatics*. doi:10.21203/rs.3.rs-428221/v1

Wang, B., and Zhang, J. (2021). Principal Component Regression Analysis for Lncrna-Disease Association Prediction Based on Pathological Stage Data. *IEEE Access* 9, 20629–20640. doi:10.1109/access.2021.3053839

Wang, J., Kuang, Z., Ma, Z., and Han, G. (2020). Gbdtl2e: Predicting lncrna-ef associations using diffusion and hetesim features based on a heterogeneous network. *Front. Genet.* 11, 272. doi:10.3389/fgene.2020.00272

Wang, L., Xuan, Z., Zhou, S., Kuang, L., and Pei, T. (2019). A Novel Model for Predicting Lncrna-Disease Associations Based on the Lncrna-Mirna-Disease Interactive Network. *Cbio* 14, 269–278. doi:10.2174/1574893613666180703105258

Wang, M.-N., You, Z.-H., Wang, L., Li, L.-P., and Zheng, K. (2021b). Ldgrnmf: Lncrna-Disease Associations Prediction Based on Graph Regularized Non-negative Matrix Factorization. *Neurocomputing* 424, 236–245. doi:10.1016/j.neucom.2020.02.062

Wang, Y., Li, H., Kuang, L., Tan, Y., Li, X., Zhang, Z., et al. (2021c). Iclrbbn: a Tool for Accurate Prediction of Potential Lncrna Disease Associations. *Mol. Ther. - Nucleic Acids* 23, 501–511. doi:10.1016/j.omtn.2020.12.002

Wu, Q. W., Xia, J. F., Ni, J. C., and Zheng, C. H. (2021). GAERF: Predicting lncRNA-Disease Associations by Graph Auto-Encoder and Random forest. *Brief Bioinform* 22. doi:10.1093/bib/bbaa391.Bbaa391

Wu, X., Lan, W., Chen, Q., Dong, Y., Liu, J., and Peng, W. (2020). Inferring Lncrna-Disease Associations Based on Graph Autoencoder Matrix Completion. *Comput. Biol. Chem.* 87, 107282. doi:10.1016/j.compbiolchem.2020.107282

Xiao, Y., Xiao, Z., Feng, X., Chen, Z., Kuang, L., and Wang, L. (2020). A Novel Computational Model for Predicting Potential Lncrna-Disease Associations Based on Both Direct and Indirect Features of Lncrna-Disease Pairs. *BMC bioinformatics* 21, 555–622. doi:10.1186/s12859-020-03906-7

Xiao, Y., Zhang, J., and Deng, L. (2017). Prediction of Lncrna-Protein Interactions Using Hetesim Scores Based on Heterogeneous Networks. *Sci. Rep.* 7, 3664–3712. doi:10.1038/s41598-017-03986-1

Xie, G., Huang, B., Sun, Y., Wu, C., and Han, Y. (2021). Rwsf-blp: a Novel Lncrna-Disease Association Prediction Model Using Random Walk-Based Multi-Similarity Fusion and Bidirectional Label Propagation. *Mol. Genet. Genomics*, 1–11. doi:10.1007/s00438-021-01764-3

Xie, G., Jiang, J., and Sun, Y. (2020a). Lda-lnsubrw: Lncrna-Disease Association Prediction Based on Linear Neighborhood Similarity and Unbalanced Bi-random Walk, *Ieee/acm Trans. Comput. Biol. Bioinf.*, 1, 1. doi:10.1109/TCBB.2020.3020595

Xie, G., Wu, C., Gu, G., and Huang, B. (2020b). Haubrw: Hybrid Algorithm and Unbalanced Bi-random Walk for Predicting Lncrna-Disease Associations. *Genomics* 112, 4777–4787. doi:10.1016/j.ygeno.2020.08.024

Xiong, M., Wu, M., Peng, D., Huang, W., Chen, Z., Ke, H., et al. (2021). Lncrna Dancr Represses Doxorubicin-Induced Apoptosis through Stabilizing Malat1 Expression in Colorectal Cancer Cells. *Cel Death Dis.* 12, 1–17. doi:10.1038/s41419-020-03318-8

Yang, X.-F., Zhou, Y.-K., Zhang, L., Gao, Y., and Du, P.-F. (2020). Predicting Lncrna Subcellular Localization Using Unbalanced Pseudo-k Nucleotide Compositions. *Cbio* 15, 554–562. doi:10.2174/1574893614666190902151038

Yao, D., Zhan, X., Zhan, X., Kwoh, C. K., Li, P., and Wang, J. (2020). A Random forest Based Computational Model for Predicting Novel Lncrna-Disease Associations. *BMC bioinformatics* 21, 126–218. doi:10.1186/s12859-020-3458-1

Yu, F., Zhang, X., Gao, L., Xue, H., Liu, L., Wang, S., et al. (2021). LncRNA Loc105377478 Promotes NPs-Nd2O3-Induced Inflammation in Human Bronchial Epithelial Cells through the ADIPOR1/NF-Kb axis. *Ecotoxicology Environ. Saf.* 208, 111609. doi:10.1016/j.ecoenv.2020.111609

Zeng, M., Lu, C., Fei, Z., Wu, F., Li, Y., Wang, J., et al. (2020). DMFLDA: A Deep Learning Framework for Predicting lncRNA-Disease Associations, *Ieee/acm Trans. Comput. Biol. Bioinf.*, 1. doi:10.1109/TCBB.2020.2983958

Zhang, D., Yin, J., Zhu, X., and Zhang, C. (2018). "MetaGraph2vec: Complex Semantic Path Augmented Heterogeneous Network Embedding," in

Pacific-Asia conference on knowledge discovery and data mining (Springer), 196–208. doi:10.1007/978-3-319-93037-4_16

Zhang, P., Zhao, B.-W., Wong, L., You, Z.-H., Guo, Z.-H., and Yi, H.-C. (2020). "A Novel Computational Method for Predicting Lncrna-Disease Associations from Heterogeneous Information Network with Sdne Embedding Model," in International Conference on Intelligent Computing (Springer), 505–513. doi:10.1007/978-3-030-60802-6_44

Zhang, Y., Chen, M., Xie, X., Shen, X., and Wang, Y. (2021). Two-stage Inference for Lncrna-Disease Associations Based on Diverse Heterogeneous Information Sources. *IEEE Access* 9, 16103–16113. doi:10.1109/ACCESS.2021.3053030

Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2017). Katzlgo: Large-Scale Prediction of Lncrna Functions by Using the Katz Measure Based on Multiple Networks. *Ieee/acm Trans. Comput. Biol. Bioinform* 16, 407–416. doi:10.1109/TCBB.2017.2704587

Zhao, X., Yang, Y., and Yin, M. (2020). Mhrwr: Prediction of Lncrna-Disease Associations Based on Multiple Heterogeneous Networks, *Ieee/acm Trans. Comput. Biol. Bioinf.*, 1. doi:10.1109/TCBB.2020.2974732

Zhou, J.-R., You, Z.-H., Cheng, L., and Ji, B.-Y. (2021). Prediction of Lncrna-Disease Associations via an Embedding Learning hope in Heterogeneous Information Networks. *Mol. Ther. - Nucleic Acids* 23, 277–285. doi:10.1016/j.omtn.2020.10.040

Zhu, R., Wang, Y., Liu, J. X., and Dai, L. Y. (2021). Ipcarf: Improving Lncrna-Disease Association Prediction Using Incremental Principal Component Analysis Feature Selection and a Random forest Classifier. *BMC bioinformatics* 22, 175–217. doi:10.1186/s12859-021-04104-9

Zhu, W., Huang, K., Xiao, X., Liao, B., Yao, Y., and Wu, F.-X. (2020). Alsbmf: Predicting Lncrna-Disease Associations by Alternating Least Squares Based on Matrix Factorization. *IEEE Access* 8, 26190–26198. doi:10.1109/access.2020.2970069

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Copyright © 2021 Duan, Kuang, Wang and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 23 December 2021
doi: 10.3389/fcell.2021.808208

Comprehensive Analysis Revealed that *CDKN2A* is a Biomarker for Immune Infiltrates in Multiple Cancers

Zheng Chen[1,2], *Yingjie Guo*[1,3], *Da Zhao*[1,2], *Quan Zou*[1], *Fusheng Yu*[4]*, *Lijun Zhang*[2]* and *Lei Xu*[3]*

[1]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [2]School of Applied Chemistry and Biological Technology, Shenzhen Polytechnic, Shenzhen, China, [3]School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China, [4]Beidahuang Industry Group General Hospital, Harbin, China

The *CDKN2A* (cyclin dependent kinase inhibitor 2A/multiple tumor suppressor 1) gene, also known as the *P16* gene, encodes multiple tumor suppressor 1 (*MTS1*), which belongs to the INK4 family. In tumor tissue, *CDKN2A* has a high expression level compared with normal tissue and reflects prognosis in tumor patients. Our research targeted the analysis of *CDKN2A* expression in 33 tumors and clinical parameters, patient prognosis and tumor immunity roles. The *CDKN2A* expression level was significantly correlated with the tumor mutation burden (TMB) in 10 tumors, and the expression of *CDKN2A* was also correlated with MSI (microsatellite instability) in 10 tumors. *CDKN2A* expression was associated with infiltrating lymphocyte (TIL) levels in 22 pancancers, thus suggesting that *CDKN2A* expression is associated with tumor immunity. Enrichment analysis indicated that *CDKN2A* expression was involved in natural killer cell-mediated cytotoxicity pathways, antigen processing and presentation, olfactory transduction pathways, and regulation of the autophagy pathway in multiple cancers. *CDKN2A* was significantly associated with several immune cell infiltrates in pantumors. *CDKN2A* may serve as a promising prognostic biomarker and is associated with immune infiltrates across cancers.

Keywords: *CDKN2A*, TMB, MSI, immune infiltrates, pan-cancer

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan
Jialiang Yang,
Geneis (Beijing) Co. Ltd., China

***Correspondence:**
Lei Xu
csleixu@szpt.edu.cn
Lijun Zhang
c7zlj@szpt.edu.cn
Fusheng Yu
fushengyu1@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *03 November 2021*
Accepted: *06 December 2021*
Published: *23 December 2021*

Citation:
Chen Z, Guo Y, Zhao D, Zou Q, Yu F, Zhang L and Xu L (2021) Comprehensive Analysis Revealed that CDKN2A is a Biomarker for Immune Infiltrates in Multiple Cancers. Front. Cell Dev. Biol. 9:808208. doi: 10.3389/fcell.2021.808208

INTRODUCTION

In recent years, malignant tumor incidence has increased year by year, and finding effective treatment methods for malignant tumors has been a research focus worldwide (Cheng et al., 2018; Zhao et al., 2019). At present, the effect of tumor treatment is unsatisfactory, and tumorigenesis is associated with gene variations. Mutations in cell cycle-related genes usually lead to tumor formation, such as *TP53*, *CDKN2A*, *RB1* and *BRCA1*.

The *CDKN2A* gene, also known as the *P16* gene, encodes multiple tumor suppressor 1 (*MTS1*), which belongs to the INK4 family. The *CDKN2A* gene was identified by yeast two-hybrid protein correlation screening for proteins that interact with *CDK4* (cyclin-dependent kinase 4) (Serrano et al., 1993). The *CDKN2A* gene, located on chromosome 9p21 and 8.5 kb in length, contains 3 exons and encodes a protein composed of 148 amino acids (Serra and Chetty, 2018). The P16 protein can bind to CDK4 and CDK6 (cyclin-dependent kinase 6) and inhibit the formation of kinase activity complexes by cyclin D (CD) and CDK4 (Yang et al., 1996). The kinase activity complex can prevent RB protein phosphorylation. Cells are stopped in the G phase by blocking the phosphorylation of RB protein and regulating the cell

cycle. Mutation of the *CDKN2A* gene will remove the inhibition of the CyclinD-CDK4 complex. Then, RB protein phosphorylation results in abnormal cell cycle progression, and cells gain unlimited proliferation ability (Romagosa et al., 2011).

CDKN2A tumor suppressors with mutations or gene loss are related to various tumors. Methylation of the *P16* gene may be an important mechanism in the development of ovarian cancer. Compared with patients without *P16* promoter methylation, ovarian cancer patients with *P16* promoter methylation have a significantly higher risk of disease progression (Todd et al., 2000; Cheng et al., 2016). However, promoter methylation of P16 genes cannot be used as a marker in early ovarian cancer diagnosis (Jiang et al., 2017). Melanoma is a cancer formed by the gradual accumulation of pathological mutations in normal melanocytes. Loss of the *CDKN2A* gene may participate in early invasion and metastasis of melanoma and suppress the initiation of invasion through inhibition of BRN2 in melanoma (Zeng et al., 2018). The most common gene mutations in human pancreatic cancer are *Kras* activation and *P16* inactivation. According to research on human pancreatic cancer, the proto-oncogene *Kras* can induce the expression of *P16,* and blocking the induction of *P16* by *Kras* can lead to tumor transformation and cancer cell metastasis (Chang et al., 2014). The *CDKN2A* methylation frequency is significantly higher in pancreatic cancer patients, which is related to patient survival (Tang et al., 2015). Such genetic events in the *CDKN2A* gene may play an important role in pancreatic ductal carcinoma.

A homozygous deletion mutation of *CDKN2A* (P16) was found in 31 patients with 127 EGFR mutations in lung cancer treated with EGFR-TKIs. For patients with *CDKN2A* (*P16*) deletion mutations, the median progression-free survival was 5.3 and 10.5 months, respectively. Therefore, EGFR mutation with *CDKN2A* (*P16*) deletion mutation is associated with the development of lung cancer (Jiang et al., 2016). In a study of patients with lymphoma, 25% of *CDKN2A* (*P16*) deletion mutations and 22% of *TP53* deletion mutations were detected in tumor tissues. Patients with deletion mutations in both *CDKN2A* and *TP53* had an average survival of 1.8 years, significantly lower than 4.3 and 5.1 years for patients with deletion mutations in only one of the genes. Both *CDKN2A* and *TP53* deletion mutations are closely associated with shortened overall survival (Delfau-Larue et al., 2015). Thus, *CDKN2A* deletion mutation is a poor prognostic indicator in mantle cell lymphoma. Studies on malignant glioma have shown that deletion mutations of *CDKN2A* can be seen in 40.3% of cases, of which homozygous deletion accounts for 74%, and homozygous deletion is more common in patients with primary malignant glioma (Cen et al., 2012; Hu et al., 2021a).

CDKN2A is a multiple tumor suppressor 1 (MTS1) which encoded a protein named P16. P16 can bind CDK4 and CDK6. P16 protein inhibits cyclin D (CD) and CDK4 to form a complex with kinase activity then regulates the cell cycle. Phase G and cell cycle are abnormal if the *CDKN2A* gene is mutated or deleted, and the cells obtain unlimited proliferation ability. Therefore, the study of *CDKN2A* gene is very important to further understand the impact of gene mutation on the development of cancer. Here, we sought to determine the *CDKN2A* gene deletion status in 33 kinds of tumors, and through analysis of the *CDKN2A* gene and patient survival, correlations between clinical tumor stage, tumor mutation load, microsatellite correlation, the relationships between tumor microenvironment, immune cell infiltration, gene expression and enrichment of GSEA, the relationship between the *CDKN2A* gene and tumor progression was assessed.

DATA AND METHODS

Data Acquisition and Processing

Thirty-three tumor transcription datasets, somatic mutation, and survival data were obtained from the UCSC Xena (http://xena.ucsc.edu/) database. There were 11,057 samples (10,327 tumor samples and 730 normal samples) for transcription data. The BiomaRt (version 2.44.4) package of R (version 4.0.2) was used for gene ID conversion. The dplyr (version 1.0.5) package of R (version 4.0.2) was used for *CDKN2A* gene extraction from the transcription data. To analyze the differential expression of the *CDKN2A* gene, the ggpubr (version 0.4.0) (https://CRAN.R-project.org/package=ggpubr) package was used in 33 tumors.

Relevant Analysis of *CDKN2A* Expression and Prognosis of Tumor Patients

Overall survival status information for tumor samples (10,327) was downloaded from theUCSC Xena (http://xena.ucsc.edu/) database for 33 tumors. DSS (disease-specific survival), DFI (disease-free interval) and PFI (progression-free interval) status and time information were downloaded from TCGA Pan-Cancer (PANCAN) in the UCSC Xena (http://xena.ucsc.edu/) database. The survival status and time information for the 33 tumors were extracted and used for prognostic analysis. According to the *CDKN2A* expression level, two groups with high and low expression levels were distinguished, and the prognostic value of *CDKN2A* for the two groups in each tumor was analyzed using the OS (overall survival), DSS, DFI and PFI data. The *CDKN2A* risk ratio forest plot was plotted. The limma (version 3.44.3) (Ritchie et al., 2015), survival (version 3.2-10), survminer (version 0.4.9) and forestplot (version 1.10.1) R packages were used to perform this analysis.

Analysis of *CDKN2A* Expression and TMB, Tumor Stage and MSI

The tumor microenvironment contains two components, immune and stromal cells, which possess diagnostic and prognostic value in cancers (Hanahan and Weinberg, 2011; Fridman et al., 2012; Zhang et al., 2019; Hu et al., 2020; Islam et al., 2020; Hu et al., 2021b). Stage information was obtained from the UCSC Xena (http://xena.ucsc.edu/) database for all tumor samples, and we divided the stages into 4. With all of the tumors, approximately 8099 samples were used for tumor stage analysis. The R packages limma (version 3.44.3) and ggpubr (version 0.4.0) were used for the analysis of tumor stages associated with *CDKN2A* expression. A total of 10,114 samples of 33 tumors for mutation data were used for tumor mutation

burden (TMB) analysis. The Spearman correlation test and fmsb package (https://CRAN.R-project.org/package=fmsb) of R were applied for analysis of the correlation between the *CDKN2A* gene and TMB. A total of 10,415 tumor samples were used to calculate the MSI score combined with the *CDKN2A* gene expression data for MSI analysis. The Spearman correlation test method and fmsb package were also applied to create a radar plot between the tumor and *CDKN2A* genes.

Analysis of the Tumor Microenvironment and *CDKN2A* Expression Levels

The tumor microenvironment is mainly composed of stromal cells and immune cells, which play an important role in cancer prognosis. ESTIMATE is a tool for predicting tumor purity in tumor tissues (Yoshihara et al., 2013). Estimate and limma packages were used to estimate tumor purity by calculating the proportion of immune and stromal cells based on 11,057 samples from 33 tumors. The correlations of *CDKN2A* expression and the tumor microenvironment score were tested by the Spearman correlation test method. The R packages ggplot2 (version 3.3.3), ggpubr (version 0.4.0) and ggExtra (version 0.9) were applied to determine the correlation distribution.

Correlation Analysis of Immune Cell Infiltration and *CDKN2A* Expression

The 22 immune cell types found in tumors include naive B cells, naive CD4 T cells, memory B cells and resting memory CD4 T cells and so on (Newman et al., 2015). To estimate the infiltration of these 22 immune cell types, we used CIBERSORT (https://cibersort.stanford.edu/) to calculate their percentages based on RNAseq data (Newman et al., 2019). The limma (version 3.44.3) package was used for expression data processing, and CIBERSORT was used to compute these immune cell infiltration scores. Correlations between *CDKN2A* expression and the 22 immune cell infiltration scores were tested by the Spearman correlation test method. The R packages ggpubr (version 0.4.0), ggplot2 (version 3.3.3) and ggExtra (version 0.9) were applied to determine the correlation distribution.

PPI Networks and Correlation of *CDKN2A* With Marker Genes

The *CDKN2A* PPI network was constructed in the STRING database (https://www.string-db.org/) (Szklarczyk et al., 2019), which contains 31 coexpressed genes. The 31 coexpressed genes, 1 tumor-associated gene and 47 immune checkpoint genes from the literature were applied to calculate the correlation with *CDKN2A* in 33 tumors with the Spearman correlation test and limma (version 3.44.3) package. The correlation heatmap of CDKN2A and 79 genes was constructed by the reshape2 (version 1.4.4) and RColorBrewer (version 1.1-2) packages.

Analysis of Gene Enrichment

GSEA (genome set enrichment analysis) is a method to determine whether a set of genes shows concordant differences between two biological states. Thirty-three tumor expression datasets were separated according to the *CDKN2A* expression level. The phenotype label was *CDKN2A* expression level. Based on the KEGG database (https://www.kegg.jp/) (Kanehisa and Goto, 2000) and GO database, genes were enriched into clusters. In addition, normalized enrichment scores were applied to classify the enriched pathways (Subramanian et al., 2005). FDR <0.05 was used for the enriched standard. The R packages limma (version 3.44.3), org.Hs.eg.db (version 3.11.4), enrichplot (version 1.8.1) and clusterProfiler (version 3.16.1) were applied for GSEA.

RESULTS

The mRNA Expression Levels of *CDKN2A* in Different Types of Cancers

To verify the differential expression of *CDKN2A* in tumor tissues, the *CDKN2A* transcriptome data of multiple cancers and normal tissues were analyzed using the ggpubr (version 4.0.4) package of R. This analysis revealed that the CDKN2A expression level was higher in various tumor tissues. **Figure 1** shows the differential expression of *CDKN2A* in all 33 tumors and normal tissues. This analysis revealed that *CDKN2A* expression was extremely significant in 15 tumors compared with their normal tissue, including KICH (Kidney Chromophobe), BRCA (Breast invasive carcinoma), HNSC (Head and Neck squamous cell carcinoma), CHOL (Cholangiocarcinoma), COAD (Colon adenocarcinoma), KIRC (Kidney renal clear cell carcinoma), PRAD (Prostate adenocarcinoma), KIRP (Kidney renal papillary cell carcinoma), THCA (Thyroid carcinoma), LIHC (Liver hepatocellular carcinoma), STAD (Stomach adenocarcinoma), LUAD (Lung adenocarcinoma), UCEC (Uterine Corpus Endometrial Carcinoma) and READ (Rectum adenocarcinoma). In addition, SARC (Sarcoma), GBM (Glioblastoma multiforme), BLCA (Bladder Urothelial Carcinoma), LUSC (Lung squamous cell carcinoma) and CESC (Cervical squamous cell carcinoma and) tumors were significantly different from normal tissues. No change in expression levels was observed in the remaining 14 tumors. Abnormal expression of CDKN2A may be related to the occurrence of many cancers.

Prognostic Value of *CDKN2A* in Pan-Cancer

The impact of *CDKN2A* on the prognostic value of tumor samples was analyzed and compared with that of normal samples. *CDKN2A* with high expression was associated with OS (overall survival) in 9 tumors, and *CDKN2A* also correlated with poor DSS (disease-specific survival) in 7 tumors (**Figures 2A,B**). KIRP, LIHC, PRAD and UCEC tumors had poorer DFI (disease-free interval) due to the high expression level of *CDKN2A* (**Figure 2C**). Additionally, high *CDKN2A* expression was positively related to poor PFI (progression-free interval) in 8 tumors (**Figure 2D**).

To further clarify the role of *CDKN2A* in a single cancer, clinical data for single cancers were used for analysis. Similarly, high mRNA levels of *CDKN2A* correlated with overall survival (OS) in ACC (Adrenocortical carcinoma), KIRC, LIHC, MESO

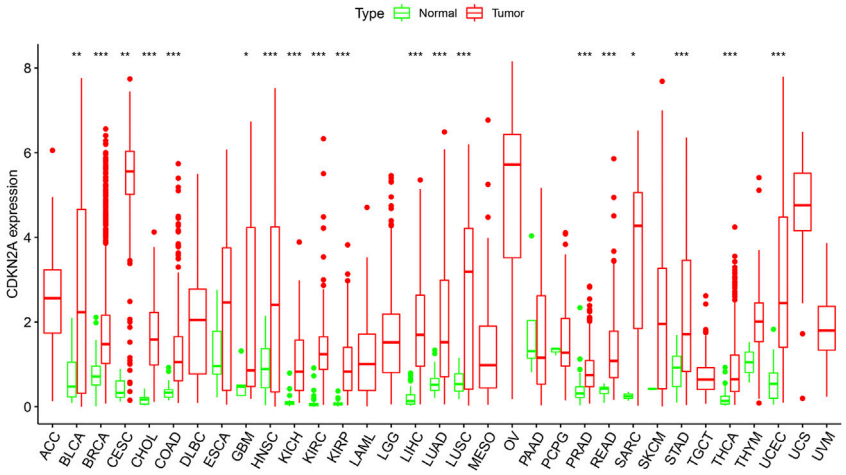


FIGURE 1 | *CDKN2A* expression levels in different tumors. Expression of *CDKN2A* in tumors were performed by ggpubr (version 4.0.4) package of R (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$).

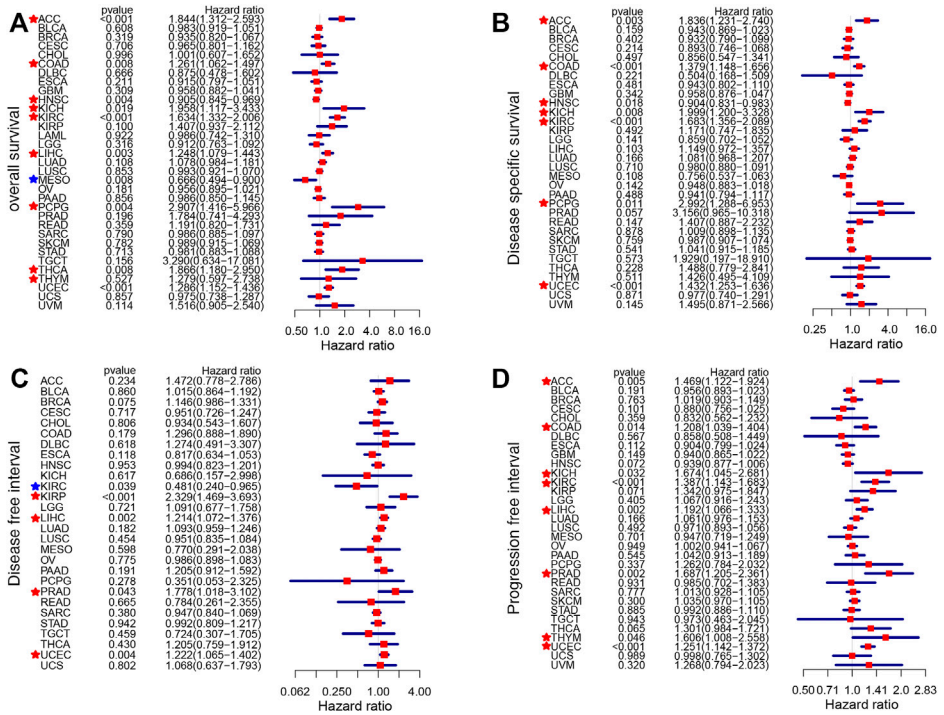


FIGURE 2 | Survival analysis of *CDKN2A* expression in 33 types of cancers. **(A)** OS (overall survival) **(B)** DSS (disease-specific survival) **(C)** DFI (disease-free interval) **(D)** PFI (progression-free interval).

(Mesothelioma), PCPG (Pheochromocytoma and Paraganglioma) and UCEC tumors, with *p*-values of 0.010, 0.029, 0.001, 0.001, 0.005 and 0.001, respectively. For disease specific survival (DSS), high mRNA levels of *CDKN2A* correlated with ACC, COAD, DLBC (Lymphoid Neoplasm Diffuse Large B), KIRC, MESO, PCPG and UCEC tumors, with *p*-values of 0.007, 0.017, 0.049, 0.006, 0.001, 0.015, and 0.001, respectively. The *p*-values of ACC, KIRC, LIHC, MESO, PARD, SKCM (Skin Cutaneous Melanoma) and UCEC tumors were 0.003, 0.028, 0.004, 0.007, 0.001, 0.037, and 0.001, respectively, due to the high expression level of *CDKN2A* in the progression-free interval (PFI). ESCA, KIRP and LIHC tumors were associated with

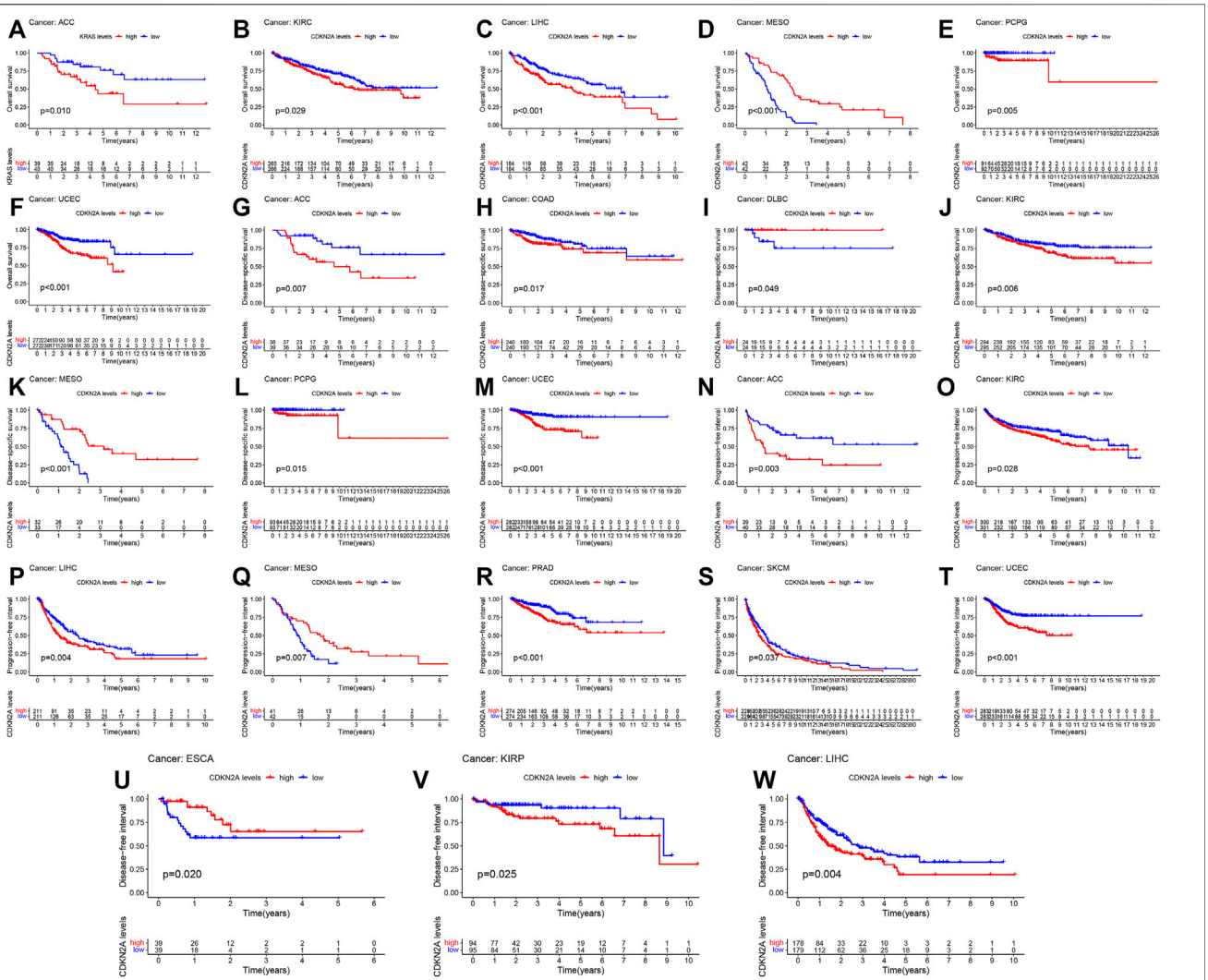


FIGURE 3 | Relationship of *CDKN2A* expression and prognosis value in 33 cancers. **(A–F)** OS **(G–M)** DSS **(N–T)** PFI **(U–W)** DFI.

high expression of *CDKN2A* for the disease-free interval (DFI), with *p*-values of 0.020, 0.025 and 0.004, respectively (**Figure 3**).

CDKN2A Expression Levels Associated With the Clinicopathological Features of Patients With Tumors

TMB (tumor mutation burden) and MSI (microsatellite instability) status are related to *CDKN2A* expression for the clinicopathological characteristics of tumor stage pathology grade. To investigate this relationship, 33 tumor sample data points from the database were downloaded for analysis. **Figures 4A–G** shows that there was some effect on the high expression level of *CDKN2A* in seven tumors at several stages, specifically ACC, COAD, KICH, KIRC, KIRP, LIHC, and THCA. Moreover, *CDKN2A* with high expression levels in 10 tumors, BLCA, ACC, BRCA, THCA, PRAD, UCEC, LUAD, KICH, SKCM, KIRC, LIHC, and HNSC, was significantly correlated with TMB (**Figure 4H**), and high expression levels of *CDKN2A* in BRCA, UCEC and PRAD tumors were also significantly correlated with microsatellite instability (MSI) (**Figure 4I**).

CDKN2A Expression Levels Associated With Tumor Microenvironment

For tumor microenvironment analysis, the estimate package of R was used to calculate the immune score and stromal score based on the *CDKN2A* expression level. Thirty-three tumor sample data points from the database were downloaded for analysis. As shown in **Figure 5**, the stromal score of LIHC and immune score of STAD were negatively correlated with *CDKN2A* expression levels. However, the immune scores of BRCA, CESC, KIRC, LGG (Brain Lower Grade Glioma), OV (Ovarian serous cystadenocarcinoma), READ, TGCT (Testicular Germ Cell Tumors), and THCA and the stromal scores of COAD, TGCT and THCA were negatively correlated with the *CDKN2A*

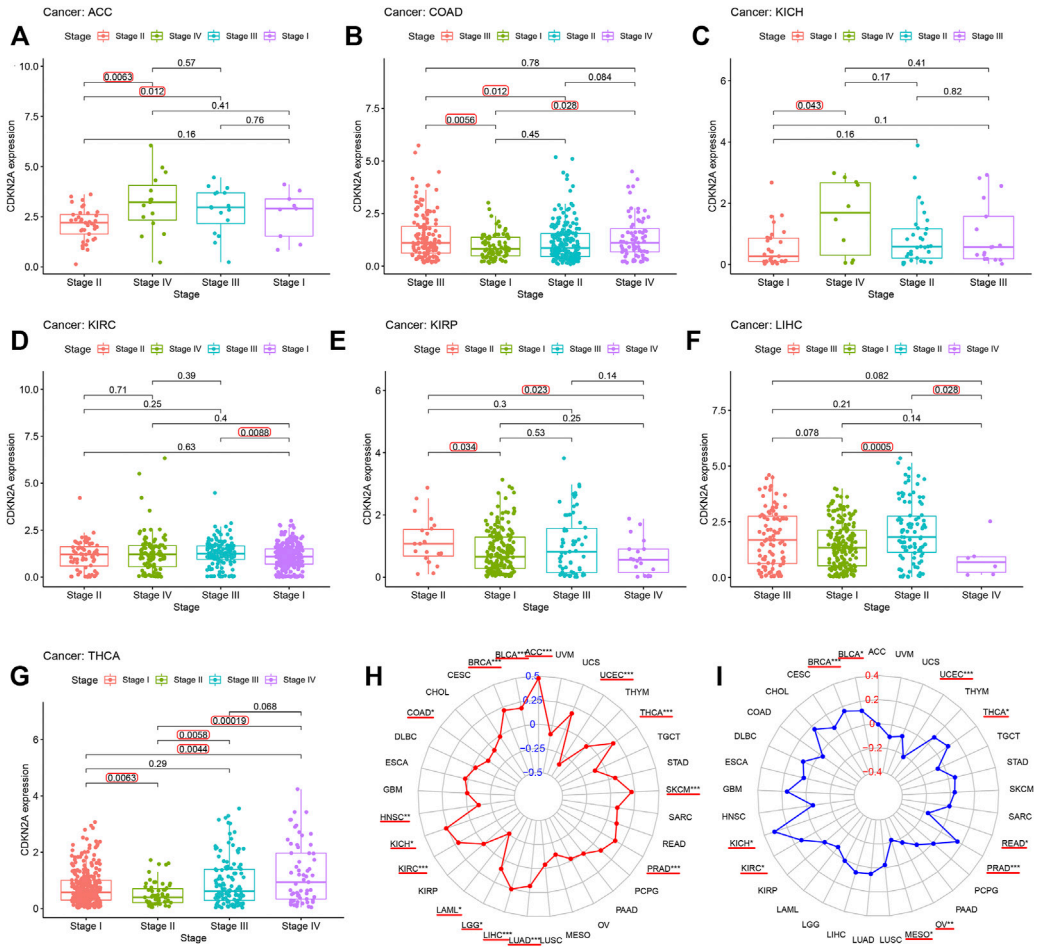


FIGURE 4 | Relationship between *CDKN2A* expression level and clinicopathological characteristics. **(A–G)** Relationship of *CDKN2A* expression level and stage grade. **(H)** Tumor mutation burden (TMB). **(I)** Microsatellite instability (MSI). *, $p < 0.05$. **$p < 0.01$, ***$p < 0.001$.

expression level. The results showed that all immune and stromal scores of these tumors were significantly correlated with the *CDKN2A* expression level.

The Association of *CDKN2A* With Tumor Immune Cell Infiltration Levels

To investigate the relationship between immune cells and *CDKN2A* expression in tumors, 22 immune infiltrating cell expression datasets were prepared from the CIBERSORT website (https://cibersort.stanford.edu/runcibersort.php). High *CDKN2A* expression levels were positively associated with memory activated CD4 T cells, CD8 T cells, activated NK cells, regulatory T cells (Tregs), activated dendritic cells, and follicular helper T cells and were negatively related to the resting immune cell levels of mast cells, M2 macrophages, and memory resting CD4 T cells in BRCA. In HNSC, high *CDKN2A* expression was positively correlated with CD8 T cells and regulatory T cells (Tregs). High *CDKN2A* expression in KIRC was positively correlated with regulatory T cells (Tregs) and CD8 T cells and negatively correlated with M2 macrophages, resting memory CD4 T cells, and resting mast cells. *CDKN2A* expression in THCA tumors was positively associated with activated dendritic cells, regulatory T cells (Tregs), resting dendritic cells, activated memory CD4 memory T cells, and follicular helper T cells and negatively correlated with M0 macrophages and M2 macrophages. Furthermore, the results indicated that high expression of *CDKN2A* was negatively correlated with resting-state immune cells and positively correlated with the activation of immune cells in multiple tumors (**Figure 6**).

Functional Analysis of *CDKN2A*

The *CDKN2A* gene encodes multiple tumor suppressor 1, which regulates various cancers through different biological processes. The PPI network was used to investigate the molecular function of the CDKN2A protein using the STRING database (**Figure 7A**). The interacting genes from the protein-protein interaction network and common immune checkpoint genes were used to analyze the relationship with *CDKN2A*. The results indicated that several immune checkpoint genes and the genes from the PPI network were correlated with *CDKN2A* expression levels in many tumors (**Figure 7B**). GSEA was used for the further analysis of

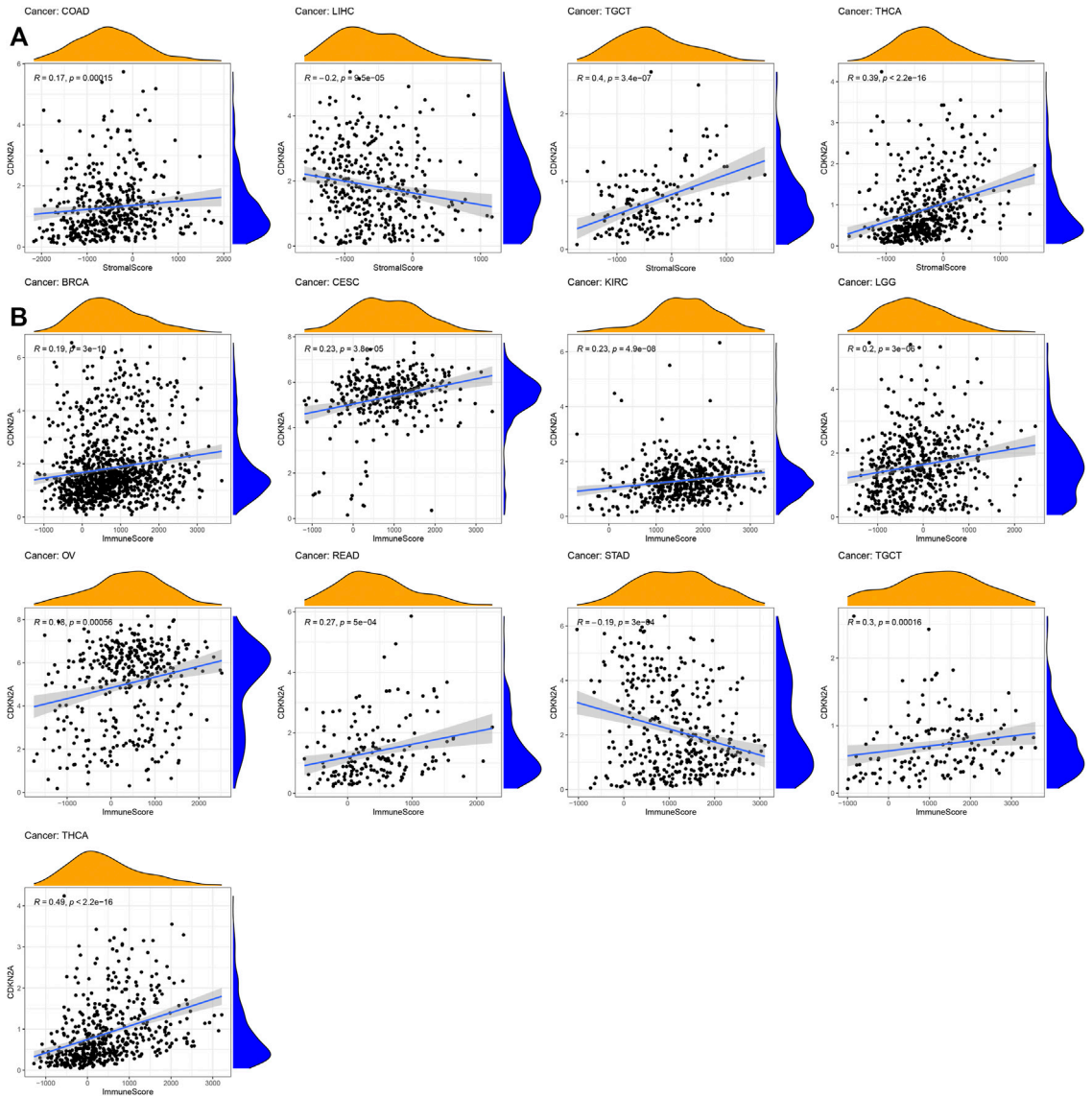


FIGURE 5 | Relationships between *CDKN2A* expression and tumor microenvironment. **(A)** Stromal score, **(B)** Immune score.

molecular mechanisms. High expression levels of *CDKN2A* are involves in the immune response, regulation of signaling pathways and regulation of immune effector processes in ACC. In BLCA, *CDKN2A* promoted blood vessel endothelial cell migration and mRNA binding. In addition, a high expression level of *CDKN2A* participated in epidermal cell differentiation, epidermal development and the humoral immune response in BRCA, KIRC and OV tumors. The T cell receptor complex is involved in KICH and SKCM tumors due to high *CDKN2A* expression. High expression levels of *CDKN2A* participated in keratinocyte differentiation and the intermediate filament cytoskeleton in LAML. High expression levels of *CDKN2A* participated in the immune response regulating signaling pathways and lymphocyte-mediated immunity intermediate filament cytoskeleton in LGG. In PCPGs, *CDKN2A* promoted B cell activation and the production of molecular mediators of the immune response. Circulating immunoglobulin complex and immune response regulating signaling pathways are involved in TGCT due to high *CDKN2A* expression (**Figure 8A**).

KEGG analysis suggested that a high *CDKN2A* expression level promoted many pathways in various tumors. In BRCA, CESE and OV, high *CDKN2A* expression levels are involved in antigen processing and presentation. High *CDKN2A* expression levels promoted natural killer cell-mediated cytotoxicity pathways in BRCA, OV and UVM (Uveal Melanoma). Olfactory transduction pathways are also involved in BRCA, OV, SARC, STAD, THCA, ESCA and LGG tumors. High *CDKN2A* expression levels promoted the regulation of the autophagy pathway in BRCA and CESC. In OV and UVM,

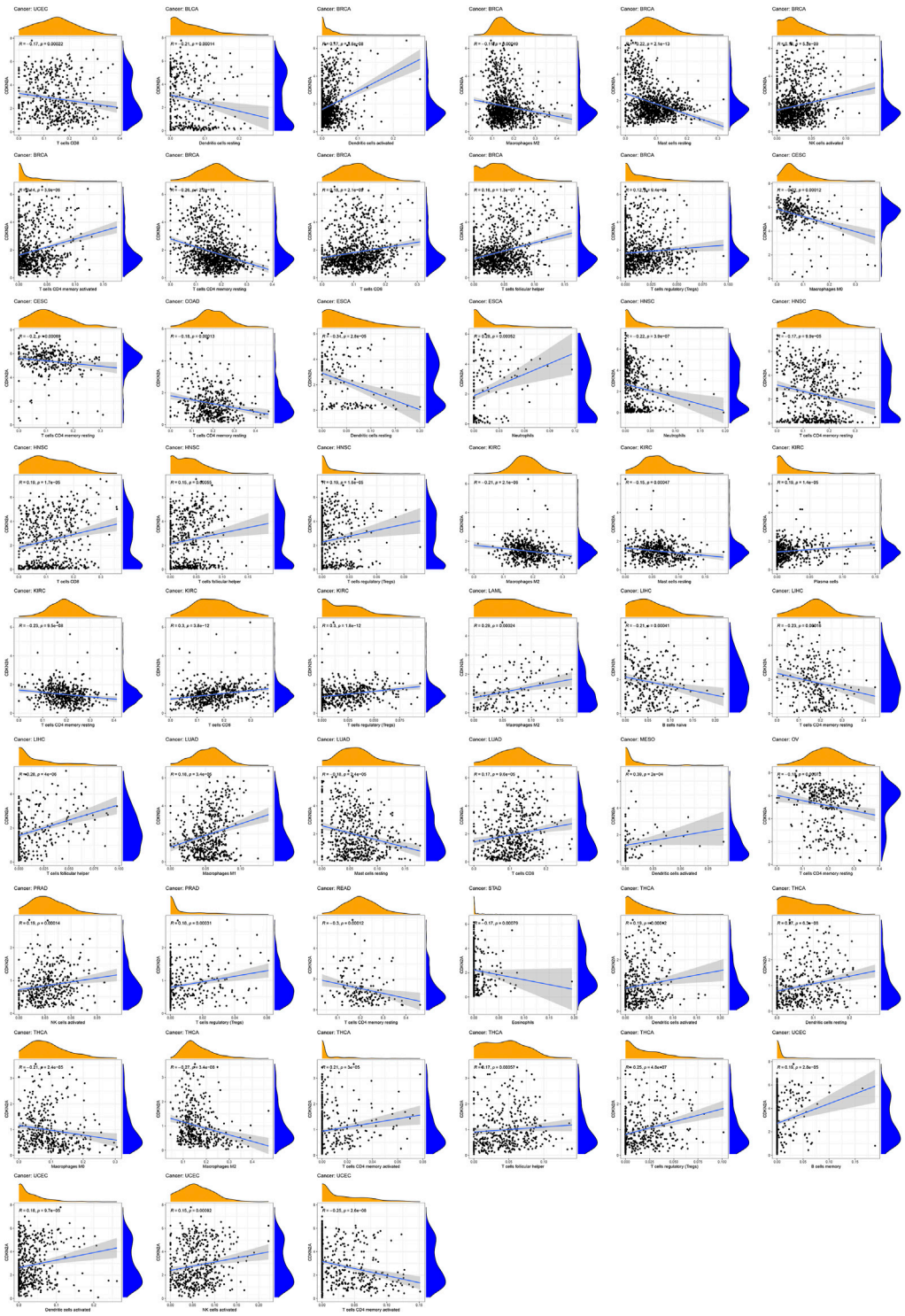


FIGURE 6 | Relationships between *CDKN2A* expression and different types of immune cells infiltration level in tumors.

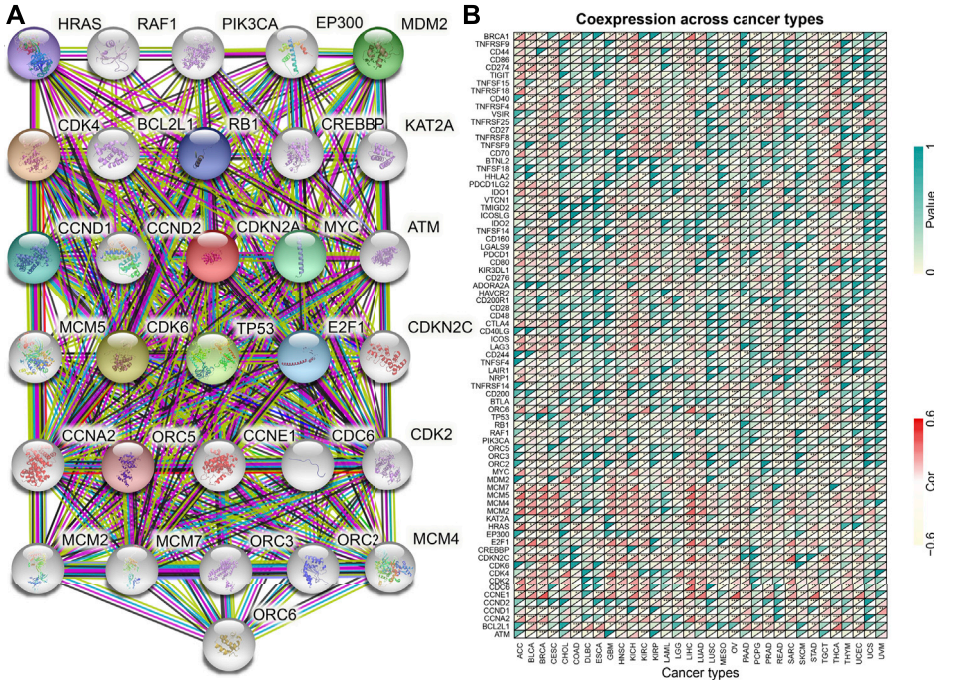


FIGURE 7 | *CDKN2A* protein-protein network and expression relationships between *CDKN2A* and related genes. **(A)** PPI network for *CDKN2A*-interaction genes. **(B)** Correlation between *CDKN2A* expression and related genes (immune checkpoint genes and interacted related genes) expression.

high *CDKN2A* expression levels were involved in TOLL-like receptor signaling and the RIG-I-like receptor signaling pathway. *CDKN2A* has the potential to serve as a biomarker in various cancers (**Figure 8B**).

DISCUSSION

There are two proteins, p14ARF and p16INK4A, with tumor suppressive functions encoded by the *CDKN2A* locus (Sharpless and DePinho, 1999). The p16INK4A protein is a CDK (cyclin-dependent kinase) inhibitor, and it functions as a predominant suppressor of melanoma (Krimpenfort et al., 2001; Sviderskaya et al., 2002; Goldstein et al., 2006; Bennett, 2015). The expression level of the *CDKN2A* gene was higher in tumor tissue than in normal tissue. Furthermore, *CDKN2A* expression was significantly higher in 15 tumors, which was consistent with tumor stage analysis in COAD, KIRC, KIRP, KIHC, LIHC and THCA. Several studies have indicated the prognostic role of *CDKN2A* in many tumors (Christopher et al., 2002; Ai et al., 2003; Zeng et al., 2018; Christodoulou et al., 2020; Ji et al., 2020; Xande et al., 2020). Therefore, *CDKN2A* plays an important prognostic role in these tumors.

Recently, tumor mutation burden (TMB) has been highlighted as a positive predictive factor and immunotherapy biomarker for immune checkpoint inhibitors and participates in immune checkpoint blockade therapy (Chan et al., 2018; Lee et al., 2019). MSI-high status is usually used to predict tumor-agnostic markers and salvage immunotherapy in critically ill patients with end-stage cancer (Pietrantonio et al., 2020). The results indicated that TMB and high MSI were associated with high expression of *CDKN2A* in a variety of tumors, such as BLCA, BRCA, UCEC, THCA, PRAD, KIRC and KICH. The relationship of *CDKN2A* with MSI and TMB in this research is novel in these tumors, and *CDKN2A* is related to TMB and MSI and deserves more in-depth study.

Previous studies have shown that immune cells can recognize CDKN2A frameshift products (Hastings and Rausch, 2014). In mice, knockdown of *CDKN2A* can reduce IL-4-induced IL-10 production, and IL-4 can induce IL-10-producing CD8(+) T cells (Zhao et al., 2013). *CDKN2A* gene occurs more frequently in cell lines than in pancreatic cancer tissues. Such genetic events on the CDKN2A gene may play an important role possibly at a later step in the progression of pancreatic ductal carcinoma (Sugimoto et al., 1998). Zeng et al. find that complete CDKN2A loss coincides with the onset of invasiveness in melanocytic tumors at distinct progression stages (Zeng et al., 2018). p16 inactivation was the major mechanism of RB pathway perturbation in non-small-cell lung carcinoma, with homozygous deletion being the most frequent method, followed by methylation and the rarer point mutations (Tam et al., 2013). The detection of homozygous CDKN2A deletion by FISH would have been helpful in confirming a diagnosis of mesothelioma over reactive mesothelial cells in 12 of 13 samples with positive or suspicious cytology (Illei et al., 2003). For tumor immune cell infiltration, the results suggested that *CDKN2A* expression was positively related to CD8 T cells, NK activated cells and CD4 T activated cells but negatively related to memory resting CD4 T cells and resting CD4

FIGURE 8 | GSEA enrichment analysis of *CDKN2A* expression in 33 tumors. **(A)** GO enrichment analysis in various tumors. **(B)** KEGG enrichment analysis in various tumors.

T cells. High expression of *CDKN2A* can increase activated immune cell numbers and reduce resting immune cell numbers. Our research showed that *CDKN2A* was associated with the infiltration of multiple immune cells in BRCA, HNSC, KIRC and THCA. In these tumors, *CDKN2A* may be a potential immunotherapeutic biomarker to affect tumor cell viability.

The PPI network was analyzed in the STRING database, and 30 proteins interacted with CDKN2A, including CCND1, CDK1, CCND2, CDK2, CDK6, CCNE1, CDK4, TP53, MYC, CDC6, and HRAS. Most of the interacting proteins belong to cell cycle-associated proteins and proto-oncogene-associated proteins. The apoptotic pathway contains 10 core genes (*BAX*, *TP53*, *TP53INP1*, *CDKN2A*, *TP53BP1*, *CDKN1A*, *MDM2*, *CDKN1B*, *CCDN1* and *BCL2*), and these genes take part in the control of critical processes involved in hepatocellular carcinoma (HCC) (Yu et al., 2015). These genes may interact to control tumor

progression. Disturbing the balance may result in a higher probability of tumor development (Cotter, 2009).

CDKN2A as a biomarker can be applied to assess prognosis in cancers. In addition, *CDKN2A* is expressed in various tumors and positively correlates with prognosis and immune cell infiltration, with high expression levels in most tumors. In general, the expression of *CDKN2A* played a detrimental role across cancers. In an analysis of immune cell infiltration, high *CDKN2A* expression was positively and significantly associated with many activated immune cells, which indicated that *CDKN2A* may be involved in tumor immunity. Together, these integrated analyses suggested that *CDKN2A* had prognostic value in cancers. Therefore, *CDKN2A* has potential as a therapeutic target for tumor treatment. However, our study promotes future immunotherapy research.

CONCLUSION

Overall, we identified a novel cancer-related gene *CDKN2A*, which encodes *multiple tumor suppressor 1* (*MTS1*) and belongs to the INK4 family. *CDKN2A* expression levels, clinical parameters, patient prognosis and tumor immunity were investigated in 33 tumors. In addition, we also explored the association between *CDKN2A* expression level and tumor mutation burden (TMB) and microsatellite instability (MSI). However, *CDKN2A* expression is associated with infiltrating lymphocyte (TIL) levels, suggested that *CDKN2A* expression is related with tumor immunity. Based on these data, *CDKN2A* may be a promising prognostic biomarker with a potential molecular mechanism that affects survival outcomes in cancer patients.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

LX, LZ, FY, and QZ designed the research; ZC and DZ performed the research; ZC and YG analyzed the data; ZC wrote the manuscript. All authors read and approved the manuscript.

FUNDING

The work was supported by the National Natural Science Foundation of China (No.62002244, No.62002242, No.62001311) and Post-doctoral Foundation Project of Shenzhen Polytechnic (No. 6020330002K), the Research Foundation of Shenzhen Polytechnic (6021310019K). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Ai, L., Stephenson, K. K., Ling, W., Zuo, C., Mukunyadzi, P., Suen, J. Y., et al. (2003). The P16 (CDKN2a/INK4a) Tumor-Suppressor Gene in Head and Neck Squamous Cell Carcinoma: A Promoter Methylation and Protein Expression Study in 100 Cases. *Mod. Pathol.* 16, 944–950. doi:10.1097/01.mp.0000085760.74313.dd

Bennett, D. (2015). Genetics of Melanoma Progression: The Rise and Fall of Cell Senescence. *Pigment Cel Melanoma Res.* 29, 122–140. doi:10.1111/pcmr.12422

Cen, L., Carlson, B. L., Schroeder, M. A., Ostrem, J. L., Kitange, G. J., Mladek, A. C., et al. (2012). p16-CDK4-RB axis Controls Sensitivity to a Cyclin-dependent Kinase Inhibitor PD0332991 in Glioblastoma Xenograft Cells. *Neuro-oncology* 14, 870–881. doi:10.1093/neuonc/nos114

Chan, T., Yarchoan, M., Jaffee, E., Swanton, C., Quezada, S. A., Stenzinger, A., et al. (2018). Development of Tumor Mutation Burden as an Immunotherapy Biomarker: Utility for the Oncology Clinic. *Ann. Oncol.* 30, 44. doi:10.1093/annonc/mdy495

Chang, Z., Ju, H., Ling, J., Zhuang, Z., Li, Z., Wang, H., et al. (2014). Cooperativity of Oncogenic K-Ras and Downregulated p16/INK4A in Human Pancreatic Tumorigenesis. *PloS one* 9, e101452. doi:10.1371/journal.pone.0101452

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics* 34 (11), 1953–1956. doi:10.1093/bioinformatics/bty002

Cheng, L., Shi, H., Wang, Z., Hu, Y., Yang, H., Zhou, C., et al. (2016). IntNetLncSim: an Integrative Network Analysis Method to Infer Human lncRNA Functional Similarity. *Oncotarget* 7 (30), 47864–47874. doi:10.18632/oncotarget.10012

Christodoulou, E., Nell, R. J., Verdijk, R. M., Gruis, N. A., van der Velden, P. A., and van Doorn, R. (2020). Loss of Wild-type CDKN2A Is an Early Event in the Development of Melanoma in FAMMM Syndrome. *J. Invest. Dermatol.* 140, 2298–e3. doi:10.1016/j.jid.2020.03.938

Christopher, S. A., Diegelman, P., Porter, C. W., and Kruger, W. D. (2002). Methylthioadenosine Phosphorylase, a Gene Frequently Codeleted with p16(cdkN2a/ARF), Acts as a Tumor Suppressor in a Breast Cancer Cell Line. *Cancer Res.* 62, 6639–6644. doi:10.1002/cncr.10960

Cotter, T. G. (2009). Apoptosis and Cancer: the Genesis of a Research Field. *Nat. Rev. Cancer* 9, 501–507. doi:10.1038/nrc2663

Delfau-Larue, M., Klapper, W., Berger, F., Jardin, F., Briere, J., Salles, G., et al. (2015). High-dose Cytarabine Does Not Overcome the Adverse Prognostic Value of CDKN2A and TP53 Deletions in Mantle Cell Lymphoma. *Blood* 126, 604–611. doi:10.1182/blood-2015-02-628792

Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The Immune Contexture in Human Tumours: Impact on Clinical Outcome. *Nat. Rev. Cancer* 12, 298–306. doi:10.1038/nrc3245

Goldstein, A. M., Chan, M., Harland, M., Gillanders, E. M., Hayward, N. K., Avril, M.-F., et al. (2006). High-risk Melanoma Susceptibility Genes and Pancreatic Cancer, Neural System Tumors, and Uveal Melanoma across GenoMEL. *Cancer Res.* 66, 9818–9828. doi:10.1158/0008-5472.can-06-0494

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell* 144, 646–674. doi:10.1016/j.cell.2011.02.013

Hastings, K. T., and Rausch, M. P. (2014). Innate and Adaptive Immune Responses to Cancer. *Fundamentals Cancer Prev.*, 81–121. doi:10.1007/978-3-642-38983-2_4

Hu, Y., Qiu, S., and Cheng, L. (2021). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 7036592. doi:10.1155/2021/7036592

Hu, Y., Sun, J-y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (2021). rs1990622 Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19 (1), 11. doi:10.1186/s12916-020-01883-5

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 Regulates TSPAN13 Expression and Contributes to Alzheimer's Disease with Sex Differences. *Brain* 143 (11), e95. doi:10.1093/brain/awaa302

Illei, P. B., Ladanyi, M., Rusch, V. W., and Zakowski, M. F. (2003). The Use of CDKN2A Deletion as a Diagnostic Marker for Malignant Mesothelioma in Body Cavity Effusions. *Cancer* 99 (1), 51–56. doi:10.1002/cncr.10923

Islam, M. A., Kundu, S., and Hassan, R. (2020). Gene Therapy Approaches in an Autoimmune Demyelinating Disease: Multiple Sclerosis. *Curr. Gene Ther.* 19 (6), 376–385. doi:10.2174/1566523220666200306092556

Ji, Z., Huo, C., and Yang, P. (2020). Genistein Inhibited the Proliferation of Kidney Cancer Cells via CDKN2a Hypomethylation: Role of Abnormal Apoptosis. *Int. Urol. Nephrol.* 52, 1049–1055. doi:10.1007/s11255-019-02372-2

Jiang, J., Gu, Y., Liu, J., Wu, R., Fu, L., Zhao, J., et al. (2016). Coexistence of p16/CDKN2A Homozygous Deletions and Activating EGFR Mutations in Lung Adenocarcinoma Patients Signifies a Poor Response to EGFR-TKIs. *Lung Cancer* 102, 101–107. doi:10.1016/j.lungcan.2016.10.015

Jiang, Y., Yan, F., Liang, L., Wan, Y., Liu, J., and Cheng, W. (2017). Meta-analysis Demonstrates No Association between P16 Ink4a Promoter Methylation and Epithelial Ovarian Cancer. *Arch. Gynecol. Obstet.* 295, 697–704. doi:10.1007/s00404-016-4264-x

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Krimpenfort, P., Quon, K. C., Mooi, W. J., Loonstra, A., and Berns, A. (2001). Loss of p16INK4a Confers Susceptibility to Metastatic Melanoma in Mice. *Nature* 413, 83–86. doi:10.1038/35092584

Lee, D.-W., Han, S. W., Bae, J. M., Jang, H., Han, H., Kim, H., et al. (2019). Tumor Mutation Burden and Prognosis in Colorectal Cancer Patients Treated with Adjuvant Fluoropyrimidine and Oxaliplatin. *Clin. Cancer Res.* 25, 6141. doi:10.1158/1078-0432.ccr-19-1105

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry. *Nat. Biotechnol.* 37, 773–782. doi:10.1038/s41587-019-0114-2

Pietrantonio, F., Loupakis, F., Randon, G., Raimondi, A., Salati, M., Trapani, D., et al. (2020). Efficacy and Safety of Immune Checkpoint Inhibitors in Patients with Microsatellite Instability-High End-Stage Cancers and Poor Performance Status Related to High Disease Burden. *Oncologist* 25, 803–809. doi:10.1634/theoncologist.2020-0014

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). LIMMA powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Romagosa, C., Simonetti, S., López-Vicente, L., Mazo, A., Lleonart, M. E., Castellvi, J., et al. (2011). p16Ink4a Overexpression in Cancer: a Tumor Suppressor Gene Associated with Senescence and High-Grade Tumors. *Oncogene* 30 (18), 2087–2097. doi:10.1038/onc.2010.614

Serra, S., and Chetty, R. (2018). p16. *J. Clin. Pathol.* 71 (10), 853–858. doi:10.1136/jclinpath-2018-205216

Serrano, M., Hannon, G. J., and Beach, D. (1993). A New Regulatory Motif in Cell-Cycle Control Causing Specific Inhibition of Cyclin D/CDK4. *Nature* 366 (6456), 704–707. doi:10.1038/366704a0

Sharpless, N. E., and DePinho, R. A. (1999). The INK4A/ARF Locus and its Two Gene Products. *Curr. Opin. Genet. Dev.* 9 (1), 22–30. doi:10.1016/s0959-437x(99)80004-5

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Sugimoto, Y., Morita, R., Hikiji, K., Imura, G., Ogata, Y., Yasuda, D., et al. (1998). Alteration of the CDKN2A Gene in Pancreatic Cancers: Is it a Late Event in the Progression of Pancreatic Cancer. *Int. J. Oncol.* 13, 669–676. doi:10.3892/ijo.13.4.669

Sviderskaya, E. V., Simon, P. H., Tracy, J. E. -W., Lynda, C., Seth, J. O., David, J. E., et al. (2002). p16Ink4a in Melanocyte Senescence and Differentiation. *J. Natl. Cancer Inst.* 94, 446–454. doi:10.1093/jnci/94.6.446

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–d613. doi:10.1093/nar/gky1131

Tam, K. W., Zhang, W., Soh, J., Stastny, V., Chen, M., Sun, H., et al. (2013). CDKN2A/p16 Inactivation Mechanisms and Their Relationship to Smoke Exposure and Molecular Features in Non-small-cell Lung Cancer. *J. Thorac. Oncol.* 8 (11), 1378–1388. doi:10.1097/jto.0b013e3182a46c0c

Tang, B., Li, Y., Qi, G., Yuan, S., Wang, Z., Yu, S., et al. (2015). Clinicopathological Significance of CDKN2A Promoter Hypermethylation Frequency with Pancreatic Cancer. *Sci. Rep.* 5, 13563. doi:10.1038/srep13563

Todd, M. C., Sclafani, R. A., and Langan, T. A. (2000). Ovarian Cancer Cells that Coexpress Endogenous Rb and P16 Are Insensitive to Overexpression of Functional P16 Protein. *Oncogene* 19, 258–264. doi:10.1038/sj.onc.1203289

Xande, J. G., Dias, A. P., Tamura, R. E., Cruz, M. C., Brito, B., Ferreira, R. A., et al. (2020). Bicistronic Transfer of CDKN2A and P53 Culminates in Collaborative Killing of Human Lung Cancer Cells *In Vitro* and *In Vivo*. *Gene Ther.* 27, 51–61. doi:10.1038/s41434-019-0096-1

Yang, R., Serrano, M., Slater, J., Leung, E., and Koeffler, H. P. (1996). Analysis of p16INK4aand its Interaction with CDK4. *Biochem. biophysical Res. Commun.* 218, 254–259. doi:10.1006/bbrc.1996.0045

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Yu, G. P., Xiao, Q. Y., Shi, Z. Q., Tang, L. S., Ma, X. P., Zhang, L. Y., et al. (2015). Genetic Polymorphisms in Apoptosis-Related Genes and the Prognosis of Hepatocellular Carcinoma. *Am. J. Cancer Res.* 5, 3249–3259.

Zeng, H., Jorapur, A., Shain, A. H., Lang, U. E., Torres, R., Zhang, Y., et al. (2018). Bi-allelic Loss of CDKN2A Initiates Melanoma Invasion via BRN2 Activation. *Cancer cell* 34, 56–68. e9. doi:10.1016/j.ccell.2018.05.014

Zhang, X., Shi, S., Shen, J., Zhao, M., and He, Q. (2019). Functional Immunoregulation by Heme Oxygenase 1 in Juvenile Autoimmune Diseases. *Curr. Gene Ther.* 19 (2), 110–116. doi:10.2174/1566523219666190710092935

Zhao, J., Sun, T., Wu, S., and Liu, Y. (2019). High Mobility Group Box 1: An Immune-Regulatory Protein. *Curr. Gene Ther.* 19 (2), 100–109. doi:10.2174/1566523219666190621111604

Zhao, Y., Zhao, H., Sun, Y., Hao, J., Qi, X., Zhou, X., et al. (2013). IL-4 Induces a Suppressive IL-10-producing CD8+ T Cell Population via a Cdkn2a-dependent Mechanism. *J. Leukoc. Biol.* 94, 1103–1112. doi:10.1189/jlb.0213064

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chen, Guo, Zhao, Zou, Yu, Zhang and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSARY

CDKN2A/MTS1 cyclin dependent kinase inhibitor 2A/multiple tumor suppressor 1

TMB tumor mutation burden

MSI microsatellite instability

TIL infiltrating lymphocyte

TP53 CDKN2A, RB1 and BRCA1.

CDK4 cyclin-dependent kinase 4

CDK6 cyclin-dependent kinase6

BRN2 POU class 3 homeobox 2

BAX BCL2 associated X, apoptosis regulator

TP53INP1 tumor protein p53 inducible nuclear protein 1

TP53BP1 tumor protein p53 binding protein 1

CDKN1A cyclin dependent kinase inhibitor 1A

MDM2 MDM2 proto-oncogene

CDKN1B cyclin dependent kinase inhibitor 1B

CCDN1 cyclin D1

BCL2 BCL2 apoptosis regulator

GSEA genome set enrichment analysis

UCSC Xena Xena Functional Genomics Explorer

OS overall survival

DSS disease-specific survival

DFI disease-free interval

PFI progression-free interval

PPI network protein-protein interact network

KEGG Kyoto Encyclopedia of Genes and Genomes

GO Gene Ontology

FDR false discovery rate

ACC Adrenocortical carcinoma

BLCA Bladder Urothelial Carcinoma

BRCA Breast invasive carcinoma

CESC Cervical squamous cell carcinoma and endocervical adenocarcinoma

CHOL Cholangiocarcinoma

COAD Colon adenocarcinoma

CDKN2A is a biomarker for multiple cancers

DLBC Lymphoid Neoplasm Diffuse Large B cell Lymphoma

ESCA Esophageal carcinoma

GBM Glioblastoma multiforme

HNSC Head and Neck squamous cell carcinoma

KICH Kidney Chromophobe

KIRC Kidney renal clear cell carcinoma

KIRP Kidney renal papillary cell carcinoma

LAML Acute Myeloid Leukemia

LGG Brain Lower Grade Glioma

LIHC Liver hepatocellular carcinoma

LUAD Lung adenocarcinoma

LUSC Lung squamous cell carcinoma

MESO Mesothelioma

OV Ovarian serous cystadenocarcinoma

PAAD Pancreatic adenocarcinoma

PCPG Pheochromocytoma and Paraganglioma

PRAD Prostate adenocarcinoma

READ Rectum adenocarcinoma

SARC Sarcoma

SKCM Skin Cutaneous Melanoma

STAD Stomach adenocarcinoma

TGCT Testicular Germ Cell Tumors

THCA Thyroid carcinoma

THYM Thymoma

UCEC Uterine Corpus Endometrial Carcinoma

UCS Uterine Carcinosarcoma

UVM Uveal Melanoma

IL-4 interleukin 4

IL-10 interleukin 10

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 12 January 2022
doi: 10.3389/fcell.2021.811585

Identification of Diagnostic Markers for Breast Cancer Based on Differential Gene Expression and Pathway Network

Shumei Zhang[1], *Haoran Jiang*[1], *Bo Gao*[2], *Wen Yang*[3]* *and Guohua Wang*[1]*

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]Department of Radiology, The Second Affiliated Hospital, Harbin Medical University, Harbin, China, [3]International Medical Center, Shenzhen University General Hospital, Shenzhen, China

Background: Breast cancer is the second largest cancer in the world, the incidence of breast cancer continues to rise worldwide, and women's health is seriously threatened. Therefore, it is very important to explore the characteristic changes of breast cancer from the gene level, including the screening of differentially expressed genes and the identification of diagnostic markers.

Methods: The gene expression profiles of breast cancer were obtained from the TCGA database. The edgeR R software package was used to screen the differentially expressed genes between breast cancer patients and normal samples. The function and pathway enrichment analysis of these genes revealed significant enrichment of functions and pathways. Next, download these pathways from KEGG website, extract the gene interaction relations, construct the KEGG pathway gene interaction network. The potential diagnostic markers of breast cancer were obtained by combining the differentially expressed genes with the key genes in the network. Finally, these markers were used to construct the diagnostic prediction model of breast cancer, and the predictive ability of the model and the diagnostic ability of the markers were verified by internal and external data.

Results: 1060 differentially expressed genes were identified between breast cancer patients and normal controls. Enrichment analysis revealed 28 significantly enriched pathways ($p < 0.05$). They were downloaded from KEGG website, and the gene interaction relations were extracted to construct the gene interaction network of KEGG pathway, which contained 1277 nodes and 7345 edges. The key nodes with a degree greater than 30 were extracted from the network, containing 154 genes. These 154 key genes shared 23 genes with differentially expressed genes, which serve as potential diagnostic markers for breast cancer. The 23 genes were used as features to construct the SVM classification model, and the model had good predictive ability in both the training dataset and the validation dataset (AUC = 0.960 and 0.907, respectively).

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Leyi Wei,
Shandong University, China
Yi Xiong,
Shanghai Jiao Tong University, China

***Correspondence:**
Guohua Wang
ghwang@nefu.edu.cn
Wen Yang
13159850336@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *09 November 2021*
Accepted: *13 December 2021*
Published: *12 January 2022*

Citation:
Zhang S, Jiang H, Gao B, Yang W and Wang G (2022) Identification of Diagnostic Markers for Breast Cancer Based on Differential Gene Expression and Pathway Network. Front. Cell Dev. Biol. 9:811585. doi: 10.3389/fcell.2021.811585

Conclusion: This study showed that the difference of gene expression level is important for the diagnosis of breast cancer, and identified 23 breast cancer diagnostic markers, which provides valuable information for clinical diagnosis and basic treatment experiments.

Keywords: breast cancer, KEGG pathway network, SVM, diagnostic markers, gene expression

BACKGROUND

Breast cancer is the phenomenon of runaway proliferation of mammary epithelial cells under the action of various carcinogenic factors. The initial stage of the disease is usually marked by breast mass, nipple discharge, and axillary lymph node enlargement. In advanced stages, cancer cells can spread far away and cause cachexia, which can be manifested by loss of appetite, anorexia, weight loss, fatigue, anemia and fever, and in severe cases can be life-threatening. The incidence of breast cancer ranks first among female malignant tumors. Statistical data show that the incidence of breast cancer in Chinese women is high and tends to be younger (Amer et al., 2020).

Breast gland is the target organ of many endocrine hormones, among which estrone and estradiol are directly related to the occurrence of breast cancer. There are many risk factors for breast cancer, such as early menarche, late menopause age, infertility or late first childbearing age, lack of breast feeding (Breast cancer and breastfeeding, 2002), high estrogen level (Zhou et al., 2014), etc., which are all risk factors for breast cancer. In addition, genetic factors are also high-risk factors for breast cancer. The risk of breast cancer in first-degree relatives is 2–3 times higher than that in the general population (Gage et al., 2012). Some genetic mutations also increase the risk of breast cancer, which has been shown to be associated with estrogen exposure (Cavalieri et al., 2006; Sahu and Pattanayak, 2020). In addition, some physical factors, such as childhood radiation therapy for breast cancer, are also risk factors for breast cancer. Recent studies have found that environmental pollution can also increase the risk of breast cancer (Clapp et al., 2008).

The prognosis of breast cancer is closely related to the stage of disease development, and the earlier the disease is detected, the greater the chance of survival within 5 years. Prognosis is critical for treatment decisions, as minimally invasive treatments such as lumpectomy, radiation or hormone therapy are often offered to patients with a good prognosis. More aggressive treatments, such as more extensive mastectomy and one or more chemotherapy agents, are usually offered to patients with poor prognosis.

Although some progress has been made in the treatment and drug research of breast cancer (Zhuang et al., 2020), studies have shown that due to the lack of understanding of the pathogenesis of this complex disease, there is currently no effective treatment, and the recurrence and death of breast cancer patients are still not effectively controlled (Wapnir et al., 2018). In recent years, with the development of immunology, molecular biology and genomics technology, bioinformatics method has become an important means to study the pathogenesis of diseases, and the identification of valuable biomarkers has become a research focus (Sahu and Pattanayak, 2020; Tanaka et al., 2020; Kanathezath et al., 2021). Therefore, based on bioinformatics analysis methods, pathway and network analysis methods (Yu et al., 2021a) were used to systematically analyze breast cancer related genes, aiming to accurately understand the pathogenesis of breast cancer incidence and development, identify the most important pathogenic genes related to breast cancer, and provide valuable information for clinical diagnosis, treatment and control (Yu et al., 2018; Cao et al., 2021).

Previously, the use of bioinformatics methods to explore differentially expressed genes has made significant progress in pancreatic cancer (Xu et al., 2020), gastric cancer (Cheng et al., 2020; Gu et al., 2021; Yuan et al., 2021), colorectal cancer (Cheng et al., 2021; Zhang and Zhang, 2021), prostate cancer (Rode et al., 2021) and other diseases (Ao et al., 2021; Yu et al., 2021b; Laudisi et al., 2021; Zhang et al., 2021; Zhao et al., 2021). The occurrence and development of breast cancer is a process involving and synergistic action of multiple genes in multiple stages (Kretschmer et al., 2011). Therefore, the differential change of gene expression profile has always been a hot topic in breast cancer research (Manjili et al., 2012). At present, some studies have screened differentially expressed genes in breast cancer patients. Studies have shown that Grb14 is highly expressed in 23.1% of breast cancers, and this high expression is associated with better overall and disease-free survival, and can be used as a better independent prognostic factor (Huang et al., 2013). In adipose tissue of mammary gland, the overexpression of leptin can promote cell proliferation and cancer (Jardé et al., 2011). Abnormal growth factor signaling between stromal cells and epithelial cells can promote malignant cell growth (Wiseman and Werb, 2002; Haslam and Woodward, 2003).

In recent years, many researchers devote themselves to the research of breast cancer and have made great breakthroughs. Staub et al. have shown that patients with a low expression module co-expressed with the WIPF1 gene generally have a good prognosis in three tumor types including colorectal cancer, glioma and breast cancer (Staub et al., 2009). Protease has also been studied as a biomarker for prognosis and diagnosis of breast cancer (Jardé et al., 2011). Song et al. also demonstrated that the expression of transforming acid curly spiral egg (TACC3) gene is an independent prognostic factor in breast cancer patients (Song et al., 2018).

This study aims to identify diagnostic markers of breast cancer by differential expression genes screening and pathway networks constructing. The gene expression profile was used to screen the genes related to breast cancer, pre-screen the diagnostic markers of breast cancer, and analyze the function of these genes. The potential breast cancer markers were further explored through the construction of KEGG pathway network, and reliable breast cancer diagnostic markers were screened by combining with

differentially expressed genes. The diagnostic prediction model was constructed using these markers, and the predictive power of the model was verified. In addition, these markers are documented to provide targets and references for clinicians as well as biological experimentalists.

METHODS

Data Acquisition and Processing

In this study, expression profile data of breast cancer were obtained from The Cancer Genome Atlas (TCGA) database (Cancer Genome Atlas Research et al., 2013; Li et al., 2020). The TCGA database is one of the most ambitious and valuable cancer genomics projects currently under way. It is a joint initiative of the National Institute of Cancer Research and the National Human Genome Research Institute. It had the molecular signatures of more than 20,000 primary tumors and matched normal samples covering 33 cancer types. It contains clinical data on a variety of human cancers, genomic mutations, mRNA expression, miRNA expression, methylation, and other data, which is a very important source of information for cancer researchers (Tang et al., 2018; Tian and Wang, 2021; Zhu et al., 2021). These data have improved our ability to diagnose, treat and prevent cancer and will continue to be publicly available to anyone in the research community.

The level 3 gene expression data of RNA-SeqV2 in breast cancer were downloaded from the TCGA database and included 1218 samples, including 1104 breast cancer samples and 114 normal control samples, containing 20,530 genes. The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization center. This dataset shows the gene-level transcription estimates, as in $\log2(x+1)$ transformed RSEM normalized count. Genes are mapped onto the human genome coordinates using UCSC Xena HUGO probeMap.

Screening of Differentially Expressed Genes

In our study, the potential related genes of breast cancer were obtained by screening the differentially expressed genes. Differentially expressed genes were screened using edgeR software package. EdgeR software package is a Bioconductor software package for differential expression analysis of RNA-seq expression profiles with biological replication, which is used to identify differential expressions or differential markers using read counts from different technology platforms (including RNA-Seq, ChIP-seq, ATAC-seq, Bisulfite-seq, SAGE, etc.). It implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests (Robinson et al., 2010).

The screening principle of differentially expressed genes was based on the Fold Change (FC) and Pvalue of breast cancer genes between breast cancer patients and the control group. Set filter conditions when Pvalue <0.05, and the Fold Change of gene expression value was satisfied with FC > 2, or FC < 0.5, the genes that satisfy both conditions were identified as differentially expressed gene.

These differentially expressed genes were displayed using heat map and volcano plot. The heat map was plotted using gplots R package and the volcano plot was plotted using ggplot2 R package.

Biological Functions and Pathways Enrichment Analysis of Differentially Expressed Genes

In this study, using DAVID (database for Annotation, Visualization and Integrated Discovery) (Huang et al., 2009a; Huang et al., 2009b), a database used for annotation, visualization and integration of discoveries, we conduct a GO (Gene Ontology) biological functions enrichment analysis and a KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways enrichment analysis towards the list of differentially expressed genes, with p controlled within 0.05, which could find out the biological characteristics and pathways related.

DAVID is a bioinformatic database that combines information tools to provide a structured and complete description of biological functions for a large number of genes or proteins and help users obtain useful biological information.

Construction of KEGG Pathway Gene Interaction Network and Extraction of Key Genes

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database tool for understanding the advanced functions and availability of biological systems (such as cells, organisms, and ecosystems) based on information at the genome and molecular levels (Kanehisa and Goto, 2000). It is a computer representation of a biological system formed by molecular blocks of genes, proteins and chemicals that are integrated into molecular wiring diagrams of an information system of interactions, reactions and relationships. It also contains disease and drug information as disturbances to biological systems.

In this study, the KGML which were organized as XML files of significant pathways obtained from the previous enrichment analysis was downloaded from the KEGG website. Extract relation, entry and group in these XML files using "XML" R package. The entry element contains information about a node of the pathway. Relationship between two proteins (gene products) or two KOs (ortholog groups) or protein and compound, which is indicated by an arrow or a line connecting two nodes in the KEGG pathways. Group stands for complex of KOs group. The types associated between nodes include ECrel, PPrel, GErel, PCrel and maplink. Respectively represent enzyme-enzyme relation, protein-protein interaction, gene expression interaction, protein-compound interaction and link to another map. We integrate the real protein-protein interaction by connecting these three types of files and construct the KEGG path gene interaction network, and analyze the topology properties of the network. The network was analyzed and visualized using Cytoscape software (Shannon et al., 2003). Cytoscape is an open-source software platform that allows

users to visualize molecular interaction networks and biological pathways and integrate these networks with annotated information, gene expression profiles and other data.

The size of a node in the network is expressed by the degree of the node. The gene with a large degree of node is called Hub gene in the network. The larger its value is, the more edges are connected to the node, and it may play a more important role in maintaining the overall structure of the network. Its change may affect more genes that interact with it. Therefore, in this study, genes with a degree greater than 30 were extracted from the network as key genes affecting breast cancer.

Literature Mining Confirms the Genes Screened

Next, we take the intersection of the obtained differentially expressed genes and the Hub genes obtained in the previous step to obtain a narrow and reliable biomarker list for breast cancer diagnosis.

To test whether the biomarkers screened in our study are indeed associated with breast cancer, we use PubMed (www.pubmed.gov) to conduct a literature review, and analyze whether the genes are indeed related to breast cancer in previous reports, so as to prove that the screening of tumor-related genes using our methods is effective. PubMed (White, 2020) is a widely used search engine, built and maintained by the National Biotechnology Information Center (NCBI) of the National Library of Medicine (NLM), which can provide more than 28 million academic biomedical publications.

This method is simple and feasible. The selected genes and breast cancer will be searched as the keywords in the literature database, and then consult the literature to see if there is a strong or weak relationship between the screened genes and the occurrence and development of breast cancer.

The Construction of Diagnosis and Prediction Model of Breast Cancer

According to the expression profile data of breast cancer, the corresponding expression values of 23 breast cancer diagnostic markers obtained in the previous step were found, and the expression profile data of these 23 genes were obtained. To verify the accuracy of the model, cancer patients and normal control samples in this dataset were randomly divided into two sets, one for training set and another for test set. The principle of division was to ensure that the proportion of cancer patients and normal control samples in each set was the same. The training set included 609 samples, including 552 samples from breast cancer patients and 57 normal control samples. The validation dataset consisted of 609 samples, including 552 samples from breast cancer patients and 57 normal controls.

Next, we use the training set samples to build the support vector machine model. Support Vector machines (SVM) is a kind of dichotomous classification model. Its basic model is a linear classifier, which defines the maximum interval in the feature space, which is the biggest difference between it and perceptron. The SVM learning strategy is to maximize the interval, which can be formalized as a problem for solving convex quadratic programming.

The purpose of the SVM model is to find the maximum distance between each sample point and the hyperplane, that is, to find the hyperplane with the largest interval. Any hyperplane can be described by the following linear equation:

$$w^T x + b = 0 \tag{1}$$

Now let's start to calculate the interval. In fact, the interval is equal to the projection of the difference of two heterogeneous support vectors on w:

$$\gamma = \frac{\left(\vec{x}_+ - \vec{x}_-\right) \cdot \vec{w}^T}{||w||} = \frac{\vec{x}_+ \cdot \vec{w}^T - \vec{x}_- \cdot \vec{w}^T}{||w||} \tag{2}$$

x satisfy:

$$\begin{cases} 1^* \left(w^T x_+ + b\right) = 1, y_i = +1 \\ -1^* \left(w^T x_- + b\right) = 1, y_t = -1 \end{cases} \tag{3}$$

Get:

$$\gamma = \frac{1 - b + (1 + b)}{\|\mathrm{w}\|} = \frac{2}{\|\mathrm{w}\|} \tag{4}$$

After the interval is solved, the idea of SVM is to maximize the interval, so:

$$\max_{v,b} \frac{2}{\|\mathrm{w}\|}, \text{s.t. } y_i \left(w^T x_i + b\right) \geq 1 \, (i = 1, 2, \ldots, m) \tag{5}$$

max and min to obtain the optimization problem:

$$\min_{w,b} \frac{1}{2} \|\mathrm{w}\|^2, \text{ s.t. } y_i \left(w^T x_i + b\right) \geq 1 \, (i = 1, 2, \ldots, m) \tag{6}$$

The classification model was tested internally using a tenfold cross validation method, and the model was tested externally using test sets to verify the accuracy of the model in classifying new patients. The performance of the model was also measured using sensitivity and specificity (Cheng et al., 2018; Tao et al., 2020; Zhao et al., 2020).

RESULTS

Acquisition of Differentially Expressed Genes in Breast Cancer

By screening for differentially expressed genes, we obtained 1060 genes that were differentially expressed between breast cancer patients and normal control samples. These genes can be seen as potentially related to breast cancer. Among them, 516 were down-regulated genes and 544 were up-regulated genes. They were displayed using a volcano plot (**Figure 1A**), where the red dots represent the genes that are down-regulated in breast cancer, the blue dots represent the genes that are up-regulated in breast cancer, and the green dots represent the genes that are not significantly different.

In addition, we also used a heat map to show the expression levels of these differentially expressed genes in cancer and normal samples (**Figure 1B**). The rows in the figure represent

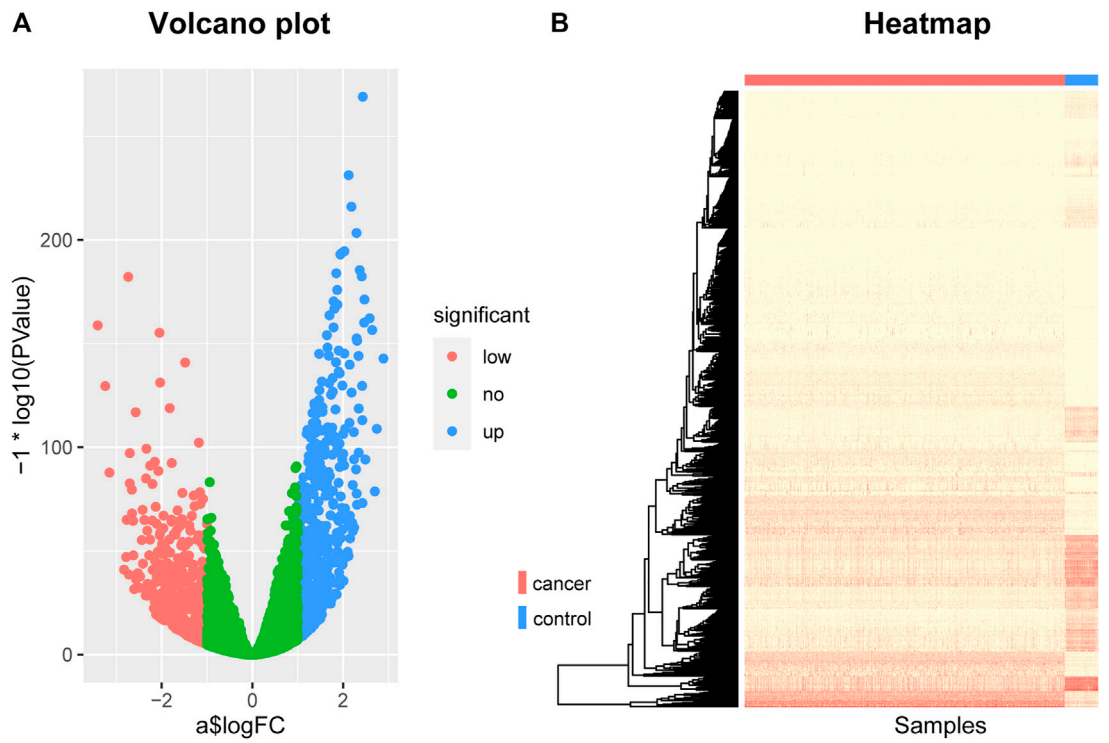


FIGURE 1 | The Volcano plot and Heat map of differentially expressed genes. **(A)** The Volcano plot of differentially expressed genes, where the red dots represent the genes that are down-regulated in breast cancer, the blue dots represent the genes that are up-regulated in breast cancer, and the green dots represent the genes that are not significantly different. **(B)** The Heat map of differentially expressed genes. The rows represent differentially expressed genes and the columns represent patients, where the red bar represents breast cancer patients and the blue bar represents normal control samples.

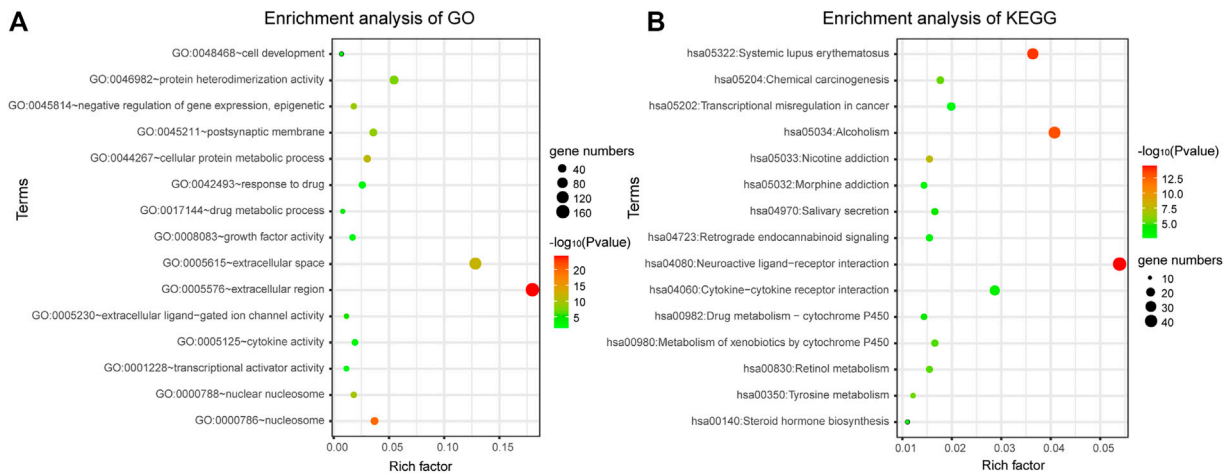


FIGURE 2 | Gene enrichment analysis. **(A)** Gene enrichment analysis of Gene Ontology. **(B)** Gene enrichment analysis of KEGG.

differentially expressed genes and the columns represent patients, where the red bar represents breast cancer patients and the blue bar represents normal control samples. As can be seen from this figure, there are indeed significant differences in these genes between breast cancer patients and normal control samples, which can clearly distinguish cancer patients from normal control samples. This suggests that the genes we screened are indeed associated with breast cancer.

TABLE 1 | KEGG pathways enriched by differentially expressed genes in breast cancer.

Term	Count	*p* value
hsa04080: Neuroactive ligand-receptor interaction	49	2.82E-15
hsa05322: Systemic lupus erythematosus	33	1.72E-14
hsa05034: Alcoholism	37	7.54E-14
hsa05033: Nicotine addiction	14	2.33E-08
hsa00350: Tyrosine metabolism	11	3.84E-06
hsa05204: Chemical carcinogenesis	16	5.16E-06
hsa00830: Retinol metabolism	14	8.86E-06
hsa00980: Metabolism of xenobiotics by cytochrome P450	15	9.57E-06
hsa04970: Salivary secretion	15	5.62E-05
hsa00982: Drug metabolism - cytochrome P450	13	8.52E-05
hsa04060: Cytokine-cytokine receptor interaction	26	2.69E-04
hsa04723: Retrograde endocannabinoid signaling	14	0.0011
hsa05032: Morphine addiction	13	0.0014
hsa00140: Steroid hormone biosynthesis	10	0.0017
hsa04727: GABAergic synapse	12	0.0025
hsa05202: Transcriptional misregulation in cancer	18	0.0028
hsa04972: Pancreatic secretion	12	0.0050
hsa04974: Protein digestion and absorption	11	0.0097
hsa04975: Fat digestion and absorption	7	0.0104
hsa00591: Linoleic acid metabolism	6	0.0117
hsa03320: PPAR signaling pathway	9	0.0150
hsa04024: cAMP signaling pathway	18	0.0152
hsa04976: Bile secretion	9	0.0176
hsa04724: Glutamatergic synapse	12	0.0217
hsa04913: Ovarian steroidogenesis	7	0.0297
hsa04950: Maturity onset diabetes of the young	5	0.0347
hsa00010: Glycolysis/Gluconeogenesis	8	0.0422
hsa00910: Nitrogen metabolism	4	0.0462

Pathways and Biological Functions Differentially Expressed Genes Involved

DAVID bioinformatics tool was employed to complete the Gene Ontology and KEGG pathway enrichment analysis. From the result of enrichment, we can see that the differentially expressed genes are involved in a number of cancer-related biological functions, such as cellular protein metabolic process, negative regulation of gene expression, epigenetic, response to drug, cell development, transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding et al. (**Figure 2A**). Through the bubble diagram, it can be seen intuitively that the GO function and KEGG pathway enriched by the 1060 differentially expressed genes. It is significant that these genes annotate the transcriptional misregulation in cancer pathway, suggesting that these genes may be closely related to cancer development and may serve as potential biomarkers for breast cancer.

In addition, a total of 28 significantly enriched pathways were identified (**Table 1**; **Figure 2B**), including Neuroactive ligand-receptor interaction, Systemic lupus erythematosus, Chemical carcinogenesis, Metabolism of xenobiotics by cytochrome, Drug metabolism—cytochrome P450, Transcriptional misregulation in cancer, cAMP signaling pathway et al. This suggests that the occurrence and development of breast cancer is a complex physiological process, with abnormal functions in a variety of pathways.

Acquisition of Potential Diagnostic Markers for Breast Cancer

Next, we downloaded the pathways with significant enrichment of differentially expressed genes obtained above from the KEGG website. Then a network is constructed by integrating the real gene interaction information from the KEGG pathways. The network contains 1277 nodes and 7345 edges, and the size of the nodes is represented by the degree (**Figure 3A**). Of the 1277 nodes (genes), 175 were differentially expressed genes (a total of 1060) previously screened between breast cancer and normal control samples, and the remaining 1102 genes were obtained by pathway enrichment analysis. If the key genes (nodes with large degree) are removed, the stability of the KEGG pathway gene interaction network will be seriously threatened, and the network topology will be lax. Therefore, although the nodes with large degree only account for a small part of the nodes in the network, which conforms to the power law distribution of the degree distribution of the biomolecular network (**Figure 3B**), they are essential key nodes.

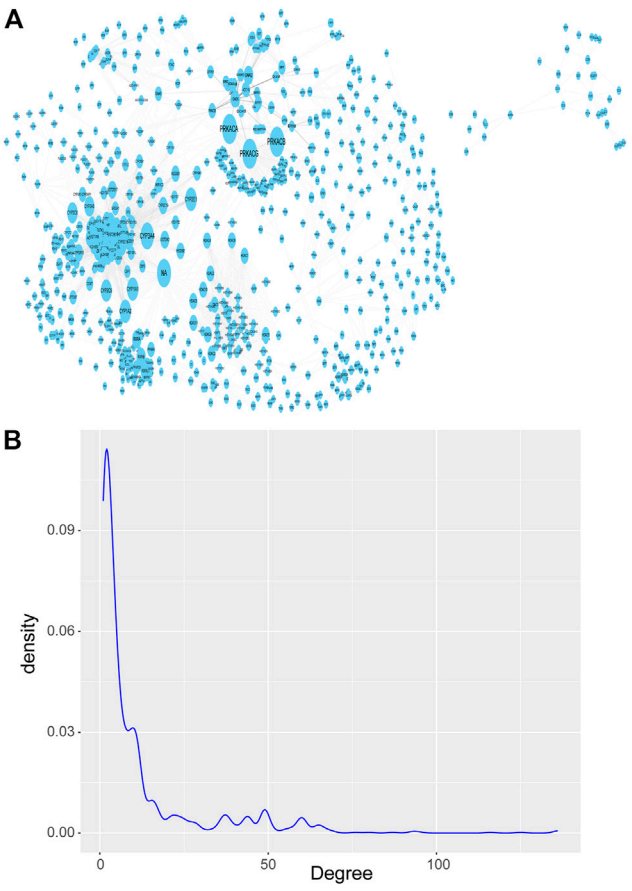


FIGURE 3 | KEGG pathways network of gene-gene interaction. **(A)** KEGG pathways network, which contains 1277 nodes and 7345 edges, and the size of the nodes is represented by the degree. **(B)** Probability density distribution of nodes degree in the network, which conforms to the power law distribution.

TABLE 2 | The candidate diagnostic markers for breast cancer.

Gene	Degree	Pvalue	Gene	Degree	*p* value
ADCY8	37	1.41E-32	CYP2C19[a]	56	1.74E-10
ADH1A[a]	65	1.50E-167	CYP3A4[a]	116	1.66E-66
ADH1C[a]	65	3.22E-87	CYP3A7[a]	43	1.59E-21
ADH4	65	4.01E-152	GNG13	44	1.61E-88
ADH6	65	1.43E-19	GNGT1	44	3.49E-22
ADH7	65	1.45E-28	GSTA1[a]	49	2.20E-66
AKR1C4[a]	25	4.83E-10	HSD3B1	52	3.17E-17
ALDH3A1[a]	38	2.24E-70	HSD3B2	52	4.44E-53
CYP1A2[a]	93	3.34E-23	RXRG[a]	64	1.15E-66
CYP2A13	58	1.23E-09	UGT1A7	60	9.17E-43
CYP2B6[a]	62	3.57E-25	UGT2B28	60	2.46E-44
CYP2C18[a]	55	6.14E-17	—	—	—

[a]*These genes have been documented to be associated with breast cancer.*

The nodes with large degree in the network are called Hub nodes. These nodes are generally the key nodes in the network, because their changes may affect more genes that interact with them. We tend to select the top 10% nodes in the network as Hub nodes. However, under this threshold, there are many nodes with the same degree and their degree is 38. Therefore, in order to ensure less omission of breast cancer diagnostic markers, we select nodes with a degree greater than 30 on the basis of this threshold. So, we filtered the degree attribute according to the attribute table exported after topology analysis and set the filter condition as degree >30, 154 genes, known as Hub genes, were identified as candidate genes potentially associated with breast cancer.

In this study, the common gene set of the previously obtained differentially expressed genes in breast cancer and key genes in the pathway network were selected, including 23 genes (**Table 2**). As candidate markers for the diagnosis of breast cancer, to ensure the accuracy of these markers.

Breast Cancer Related Genes Were Confirmed by Literature Review

Next, to verify that our method of screening for biomarkers for breast cancer is reliable. We selected the genes for literature mining and verification. To see if there is any evidence in the literature that these genes are indeed involved in the development and progression of breast cancer.

After the intersection of the two sets of genetic data sets, 23 diagnostic markers of breast cancer were obtained. After literature mining in PubMed literature retrieval system, we found that some of these 23 genes have been confirmed to be related to breast cancer, including 12 genes, ADH1A, ADH1C, AKR1C4, ALDH3A1, CYP1A2, CYP2B6, CYP2C18, CYP2C19, CYP3A4, CYP3A7, GSTA1, RXRG. For example, THE TT genotype of CYP3A4 polymorphism is associated with increased risk of breast cancer (Liu et al., 2019). RXRG protein is an independent predictor of breast cancer specific survival and distant metastasis-free survival (Joseph et al., 2019). ADH1A has been found to be a potential biomarker for diagnosis, treatment and prognosis of breast cancer (Wu and Yu, 2021). It is worth noting that the remaining genes, such as GNGT1 and UGT1A7, have not been documented to be associated with breast cancer. However, the literature has linked these genes to other cancers. Therefore, the prediction of the correlation between these genes and breast cancer will provide clinicians and biological experimentalists with targets and references for future research directions.

Construction and Prediction of Breast Cancer Diagnostic Model

According to the expression profile data of breast cancer, the corresponding expression profile data of 23 diagnostic markers of breast cancer obtained in the previous step were obtained. Cancer patients and normal control samples were divided into training sets and test sets. The proportion of breast cancer patients and normal control samples in both datasets was also ensured to be the same. The training set included 609 samples, including 552 samples from breast cancer patients and 57 from normal control samples; the validation dataset also consisted of 609 samples, including 552 samples from breast cancer patients and 57 from normal controls.

The training set was used to construct the classification model of support vector machine. The accuracy of the model was evaluated using a tenfold cross validation method. The confusion matrix was shown in **Figure 4A**. According to the matrix, 600 out of 609 samples were classified correctly, and the classification accuracy was 98.5%. The sensitivity and specificity of the model were 99.1 and 93%, respectively. The ROC curve (receiver operating characteristic curve) of the model is shown in **Figure 4C**, and the AUC reaches 0.960.

Next, the established model is used to predict the test set to test the prediction ability of the model. The confusion matrix is shown in **Figure 4B**, 593 out of 609 samples in test set are correctly classified, and the classification accuracy is 97.4%. The sensitivity and specificity of the model were 98.9 and 82.5%, respectively. Compared to the training set, there was a decrease in specificity and a certain increase in the misclassified normal control samples. The ROC curve of the model is shown in **Figure 4D**, and the AUC reaches 0.907. In other words, for new patients, once we have data on the expression levels of these 23 genes, we can use the classification model constructed in this study to predict whether they are likely to develop breast cancer.

These results indicate that the diagnostic prediction model constructed in this study can effectively distinguish between breast cancer patients and normal control population, and these 23 genes can be used as reliable biomarkers for breast cancer diagnosis, but further experiments are still necessary. Especially for genes that have not been reported in the literature. In addition, we analyzed the biological functions of these 23 diagnostic markers and found that they were involved in many cancer-related biological processes and pathways, such as chemical carcinogenesis, drug metabolism, xenobiotic metabolic process, metabolic pathways, oxygen binding etc.

A

Training dataset	Real cancer	Real normal	
Predict cancer	547	4	
Predict normal	5	53	Totals
Totals	552	57	609
Correct	547	53	600
Sensitivity(%)	99.1		
Specificity(%)		93	

B

Validation dataset	Real cancer	Real normal	
Predict cancer	546	10	
Predict normal	6	47	Totals
Totals	552	57	609
Correct	546	47	593
Sensitivity(%)	98.9		
Specificity(%)		82.5	

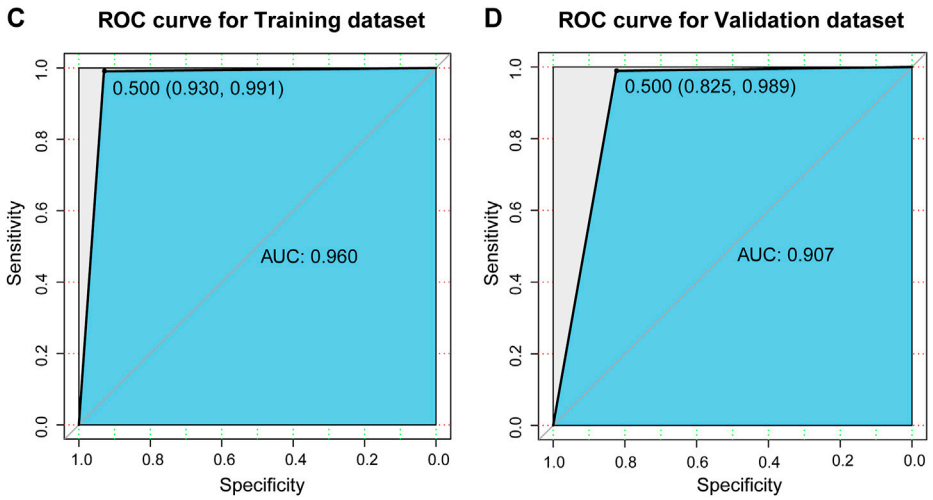


FIGURE 4 | Evaluation of classification models. **(A)** The confusion matrix of Training dataset. **(B)** The confusion matrix of Validation dataset. **(C)** ROC curve of Training dataset. **(D)** ROC curve of Validation dataset.

DISCUSSION

Breast cancer is a heterogeneous cancer with the highest incidence among women in the world, which is a serious threat to women's health. The occurrence of breast cancer is a complex biological process involved and regulated by multiple genes, and the difference of gene expression levels in tumor cells of different patients determines the different treatment and prognosis of patients (Saad et al., 2010). Therefore, to explore the characteristic changes of breast cancer from the level of genes and the discovery of biomarkers of breast cancer diagnose will play an important role in guiding the treatment of breast cancer.

Advances in high-throughput sequencing technology have made it easier for researchers to understand how diseases occur and develop at the genome-wide level. RNA-Seq refers to transcriptome sequencing technology, which is a high-throughput sequencing technology to reflect the expression level of mRNA, small RNA and noncoding RNA or some of them by determining their sequences with high-throughput sequencing technology. TCGA is an open and free database resource that integrates multiple cancer data types. A large sample size of RNA-Seq data for breast cancer can be obtained. In this study, the Level 3 gene expression data of RNA-SeqV2 in 1104 breast cancer tumor samples and 110 normal control samples adjacent to cancer were downloaded from the TCGA database, and the genetic characteristics of breast cancer were analyzed at the whole genome level using these data. Identify genes that are differentially expressed in breast cancer and molecular markers for cancer diagnosis.

In this study, a total of 1060 differentially expressed genes were screened from the tumor and normal samples of breast cancer by using edgeR R package, of which 544 genes were up-regulated and 516 were down-regulated in the cancer samples. Through the enrichment analysis of GO function, it was found that these genes were mainly enriched in biological processes such as cellular protein metabolic process, negative regulation of gene expression, epigenetic, response to drug, cell development, transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding et al. In addition, KEGG pathway enrichment analysis of these genes showed that they were significantly enriched in 28 pathways such as Neuroactive ligand-receptor interaction, Systemic lupus erythematosus,

Chemical carcinogenesis, Metabolism of xenobiotics by cytochrome, Drug metabolism - cytochrome P450, Transcriptional misregulation in cancer, cAMP signaling pathway et al. By integrating the interaction information of genes in these pathways, a KEGG pathway network was constructed, and the hub nodes in the network were extracted, and 154 candidate genes potentially related to breast cancer were obtained. By integrating with the list of differentially expressed genes, 23 potential diagnostic markers of breast cancer were finally obtained. Some of these genes have been shown to play important roles in the development of breast cancer, which confirms the reliability of the results of this study. However, the remaining genes have not been studied in relation to the risk of breast cancer. They may be new diagnostic factors or risk genes for breast cancer, so further analysis and experimental confirmation of these genes are necessary.

Using gene expression profile of the 23 diagnostic marker genes, we constructed the breast cancer diagnosis prediction model based on SVM classifier, and analyzed the model prediction ability. The results proved that the model can effectively distinguish breast cancer patients and normal people, and further identifying these 23 genes can be used as diagnostic markers of breast cancer. It provides targets and reference for clinical doctors and biological experimentalists to treat breast cancer. It is worth noting that many of the 23 genes come from the same gene family, and further research is needed to determine whether there is redundant information between them. In addition, the methods in this paper can also be applied to other patient data to guide the diagnosis and treatment of cancer patients.

In this study, bioinformatics research methods were used to systematically and comprehensively analyze the differentially expressed genes related to breast cancer, and explore the pathogenesis, therapeutic targets and target prediction of breast cancer.

In conclusion, the use of computer technology and mathematical modeling and other methods can effectively analyze the differential expression genes related to breast cancer, so as to understand the pathogenesis of breast cancer and predict the prevention and treatment targets. It can also effectively solve the problems in biomedical science, and further provide meaningful information for future experimental research.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary material.

AUTHOR CONTRIBUTIONS

GW conceived and designed the experiments. SZ and HJ conducted all the data processing work described in the section of methods and performed the analysis. SZ prepared and edited the manuscript. BG checked and proofread the entire manuscript. WY participated in the revision of the manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (62002057, 61771165, and 62172129); Fundamental Research Funds for the Central Universities (2572020BH02); National Key R and D Program of China (2021YFC2100100).

REFERENCES

Amer, A., Nagah, A., Tian, T., and Zhang, X. (2020). Mutation Mechanisms of Breast Cancer Among the Female Population in China. *Curr. Bioinformatics* 15 (3), 253–259. doi:10.2174/1574893615666191220141548

Ao, C., Yu, L., and Zou, Q. (2021). Prediction of Bio-Sequence Modifications and the Associations with Diseases. *Brief. Funct. Genomics* 20 (1), 1–18. doi:10.1093/bfgp/elaa023

Breast Cancer and Breastfeeding: Collaborative Reanalysis of Individual Data from 47 Epidemiological Studies in 30 Countries, Including 50302 Women with Breast Cancer and 96973 Women without the Disease. *Lancet* 2002, 360(9328): 187–195.doi:10.1016/S0140-6736(02)09454-0

Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Cao, C., Wang, J., Kwok, D., Cui, F., Zhang, Z., Zhao, D., et al. (2021). webTWAS: a Resource for Disease Candidate Susceptibility Genes Identified by Transcriptome-wide Association Study. *Nucleic Acids Res.* [Epub ahead of print]. doi:10.1093/nar/gkab195710.1093/nar/gkab957

Cavalieri, E., Chakravarti, D., Guttenplan, J., Hart, E., Ingle, J., Jankowiak, R., et al. (2006). Catechol Estrogen Quinones as Initiators of Breast and Other Human Cancers: Implications for Biomarkers of Susceptibility and Cancer Prevention. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1766 (1), 63–78. doi:10.1016/j.bbcan.2006.03.001

Cheng, L., Qi, C., Yang, H., Lu, M., Cai, Y., Fu, T., et al. (2021). gutMGene: a Comprehensive Database for Target Genes of Gut Microbes and Microbial Metabolites. *Nucleic Acids Res.* [Epub ahead of Print] doi:10.1093/nar/gkab786

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics* 34 (11), 1953–1956. doi:10.1093/bioinformatics/bty002

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a Comprehensive Database for Dysbiosis of the Gut Microbiota in Disorders and Interventions. *Nucleic Acids Res.* 48 (D1), D554–D560. doi:10.1093/nar/gkz843

Clapp, R. W., Jacobs, M. M., and Loechler, E. L. (2008). Environmental and Occupational Causes of Cancer: New Evidence 2005-2007. *Rev. Environ. Health* 23 (1), 1–38. doi:10.1515/reveh.2008.23.1.1

Gage, M., Wattendorf, D., and Henry, L. R. (2012). Translational Advances Regarding Hereditary Breast Cancer Syndromes. *J. Surg. Oncol.* 105 (5), 444–451. doi:10.1002/jso.21856

Gu, Y., Gao, Y., Tang, X., Xia, H., and Shi, K. (2021). Bioinformatics Analysis Identifies CPZ as a Tumor Immunology Biomarker for Gastric Cancer. *Curr. Bioinformatics* 16 (1), 98–105. doi:10.2174/1574893615999200707145643

Haslam, S. Z., and Woodward, T. L. (2003). Host Microenvironment in Breast Cancer Development: Epithelial-Cell-Stromal-Cell Interactions and Steroid Hormone Action in normal and Cancerous Mammary Gland. *Breast Cancer Res.* 5 (4), 208–215. doi:10.1186/bcr615

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res.* 37 (1), 1–13. doi:10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4 (1), 44–57. doi:10.1038/nprot.2008.211

Huang, O., Jiang, M., Zhang, X., Xie, Z., Chen, X., Wu, J., et al. (2013). Grb14 as an Independent Good Prognosis Factor for Breast Cancer Patients Treated with Neoadjuvant Chemotherapy. *Jpn. J. Clin. Oncol.* 43 (11), 1064–1072. doi:10.1093/jjco/hyt130

Jardé, T., Perrier, S., Vasson, M.-P., and Caldefie-Chézet, F. (2011). Molecular Mechanisms of Leptin and Adiponectin in Breast Cancer. *Eur. J. Cancer* 47 (1), 33–43. doi:10.1016/j.ejca.2010.09.005

Joseph, C., Al-Izzi, S., Alsaleem, M., Kurozumi, S., Toss, M. S., Arshad, M., et al. (2019). Retinoid X Receptor Gamma (RXRG) Is an Independent Prognostic Biomarker in ER-Positive Invasive Breast Cancer. *Br. J Cancer* 121 (9), 776–785. doi:10.1038/s41416-019-0589-0

Kanathezath, A., Chembra, V., Padingare Variyath, S. K., and Nair, G. G. (2021). Identification of Biomarkers and Functional Modules from Genomic Data in Stage-wise Breast Cancer. *Curr. Bioinformatics* 16 (5), 722–733. doi:10.2174/1574893615999200922123104

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Kretschmer, C., Sterner-Kock, A., Siedentopf, F., Schoenegg, W., Schlag, P. M., and Kemmner, W. (2011). Identification of Early Molecular Markers for Breast Cancer. *Mol. Cancer* 10 (1), 15. doi:10.1186/1476-4598-10-15

Laudisi, F., Stolfi, C., Bevivino, G., Maresca, C., Franze, E., Troncone, E., et al. (2021). GATA6 Deficiency Leads to Epithelial Barrier Dysfunction and Enhances Susceptibility to Gut Inflammation. *J. Crohns Colitis* [Epub ahead of print]. doi:10.1093/ecco-jcc/jjab145

Li, J., Liu, L., Cui, Q., and Zhou, Y. (2020). Comparisons of MicroRNA Set Enrichment Analysis Tools on Cancer De-regulated miRNAs from TCGA Expression Datasets. *Curr. Bioinformatics* 15 (10), 1104–1112.

Liu, X., Huang, X., Zhang, S., Niu, F., Ouyang, Y., Shou, Z., et al. (2019). Correlations between CYP3A4 Polymorphism and Susceptibility to Breast Cancer in Chinese Han Population. *Int. J. Clin. Oncol.* 24 (2), 179–188. doi:10.1007/s10147-018-1346-8

Manjili, M. H., Najarian, K., and Wang, X.-Y. (2012). Signatures of Tumor-Immune Interactions as Biomarkers for Breast Cancer Prognosis. *Future Oncol.* 8 (6), 703–711. doi:10.2217/fon.12.57

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* 26 (1), 139–140. doi:10.1093/bioinformatics/btp616

Rode, M. P., Silva, A. H., Cisilotto, J., Rosolen, D., and Creczynski-Pasa, T. B. (2021). miR-425-5p as an Exosomal Biomarker for Metastatic Prostate Cancer. *Cell Signal.* 87, 110113. doi:10.1016/j.cellsig.2021.110113

Saad, E. D., Katz, A., and Buyse, M. (2010). Overall Survival and post-progression Survival in Advanced Breast Cancer: a Review of Recent Randomized Clinical Trials. *J. Clin. Oncol.* 28 (11), 1958–1962. doi:10.1200/jco.2009.25.5414

Sahu, R., and Pattanayak, S. P. (2020). Strategic Developments & Future Perspective on Gene Therapy for Breast Cancer: Role of mTOR and Brk/PTK6 as Molecular Targets. *Curr. Gene Ther.* 20 (4), 237–258. doi:10.2174/1566523220999200731002408

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Song, H., Liu, C., Shen, N., Yi, P., Dong, F., Li, X., et al. (2018). Overexpression of TACC3 in Breast Cancer Associates with Poor Prognosis. *Appl. Immunohistochem. Mol. Morphol.* 26 (2), 113–119. doi:10.1097/pai.0000000000000392

Staub, E., Groene, J., Heinze, M., Mennerich, D., Roepcke, S., Klaman, I., et al. (2009). An Expression Module of WIPF1-Coexpressed Genes Identifies Patients with Favorable Prognosis in Three Tumor Types. *J. Mol. Med.* 87 (6), 633–644. doi:10.1007/s00109-009-0467-y

Tanaka, E., Uchida, D., Shiraha, H., Kato, H., Ohyama, A., Iwamuro, M., et al. (2020). Promising Gene Therapy Using an Adenovirus Vector Carrying REIC/Dkk-3 Gene for the Treatment of Biliary Cancer. *Curr. Gene Ther.* 20 (1), 64–70. doi:10.2174/1566523220666200309125709

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor Origin Detection with Tissue-specific miRNA and DNA Methylation Markers. *Bioinformatics* 34 (3), 398–406. doi:10.1093/bioinformatics/btx622

Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD. *Comput. Math. Methods Med.* 2020, 8926750. doi:10.1155/2020/8926750

Tian, L., and Wang, S.-L. (2021). Exploring miRNA Sponge Networks of Breast Cancer by Combining miRNA-disease-lncRNA and miRNA-target Networks. *Curr. Bioinformatics* 16 (3), 385–394. doi:10.2174/1574893615999200711171530

Wapnir, I. L., Price, K. N., Anderson, S. J., Robidoux, A., Martín, M., Nortier, J. W. R., et al. (2018). Efficacy of Chemotherapy for ER-Negative and ER-Positive Isolated Locoregional Recurrence of Breast Cancer: Final Analysis of the CALOR Trial. *J. Clin. Oncol.* 36 (11), 1073–1079. doi:10.1200/jco.2017.76.5719

White, J. (2020). PubMed 2.0. *Med. Reference Serv. Q.* 39 (4), 382–387. doi:10.1080/02763869.2020.1826228

Wiseman, B. S., and Werb, Z. (2002). Stromal Effects on Mammary Gland Development and Breast Cancer. *Science* 296 (5570), 1046–1049. doi:10.1126/science.1067431

Wu, X., and Yu, L. (2021). *EPSOL: Sequence-Based Protein Solubility Prediction Using Multidimensional Embedding*. Oxford, England: Bioinformatics.

Xu, J., Liao, K., Fu, Z., and Xiong, Z. (2020). Screening Differentially Expressed Genes of Pancreatic Cancer between Mongolian and Han People Using Bioinformatics Technology. *BMC Cancer* 20 (1), 298. doi:10.1186/s12885-020-06722-7

Yu, L., Xia, M., and An, Q. (2021). A Network Embedding Framework Based on Integrating Multiplex Network for Drug Combination Prediction. *Brief. Bioinformatics* [Epub ahead of print]. doi:10.1093/bib/bbab364

Yu, L., Wang, M., Yang, Y., Xu, F., Zhang, X., Xie, F., et al. (2021). Predicting Therapeutic Drugs for Hepatocellular Carcinoma Based on Tissue-specific Pathways. *Plos Comput. Biol.* 17 (2), e1008696. doi:10.1371/journal.pcbi.1008696

Yu, L., Zhao, J., and Gao, L. (2018). Predicting Potential Drugs for Breast Cancer Based on miRNA and Tissue Specificity. *Int. J. Biol. Sci.* 14 (8), 971–982. doi:10.7150/ijbs.23350

Yuan, G., Ding, W., Sun, B., Zhu, L., Gao, Y., and Chen, L. (2021). Upregulated circRNA_102231 Promotes Gastric Cancer Progression and its Clinical Significance. *Bioengineered* 12 (1), 4936–4945. doi:10.1080/21655979.2021.1960769

Zhang, T., Liu, N., Wei, W., Zhang, Z., and Li, H. (2021). Integrated Analysis of Weighted Gene Coexpression Network Analysis Identifying Six Genes as Novel Biomarkers for Alzheimer's Disease. *Oxid Med. Cel Longev* 2021, 9918498. doi:10.1155/2021/9918498

Zhang, Z., and Zhang, X. (2021). Identification of m6A-Related Biomarkers Associated with Prognosis of Colorectal Cancer. *Med. Sci. Monit.* 27, e932370. doi:10.12659/MSM.932370

Zhao, H., Zhang, Q., Chen, H., Rezanur Rahman, M., and Md Faruquee, H. (2021). Integrated Multi-Omics Approach Identified Molecular Mechanism and Pathogenetic Processes of COVID-19 that Affect Patient with Parkinson's Disorder. *Saudi J. Biol. Sci.* 28, 6939–6945. doi:10.1016/j.sjbs.2021.07.074

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an Ensemble Classifier-Based Feature Selection for Differential Expression Analysis on Expression Profiles. *BMC Bioinformatics* 21 (1), 43. doi:10.1186/s12859-020-3388-y

Zhou, Z., Qiao, J. X., Shetty, A., Wu, G., Huang, Y., Davidson, N. E., et al. (2014). RETRACTED ARTICLE: Regulation of Estrogen Receptor Signaling in Breast

Carcinogenesis and Breast Cancer Therapy. *Cell. Mol. Life Sci.* 71 (8), 1549. doi:10.1007/s00018-013-1376-3

Zhu, Z., Han, X., and Cheng, L. (2021). Identification of Gene Signature Associated with Type 2 Diabetes Mellitus by Integrating Mutation and Expression Data. *Curr. Gene Ther.* 21 [Epub ahead of print]. doi:10.2174/1566523221666210707140839

Zhuang, J., Dai, S., Zhang, L., Gao, P., Han, Y., Tian, G., et al. (2020). Identifying Breast Cancer-Induced Gene Perturbations and its Application in Guiding Drug Repurposing. *Curr. Bioinformatics* 15 (9), 1075–1089. doi:10.2174/1574893615666200203104214

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhang, Jiang, Gao, Yang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 19 January 2022
doi: 10.3389/fcell.2021.820342

WAFNRLTG: A Novel Model for Predicting LncRNA Target Genes Based on Weighted Average Fusion Network Representation Learning Method

Jianwei Li[1,2]*, *Zhenwu Yang*[1], *Duanyang Wang*[1] and *Zhiguang Li*[1]

[1]School of Artificial Intelligence, Institute of Computational Medicine, Hebei University of Technology, Tianjin, China, [2]Hebei Province Key Laboratory of Big Data Calculation, Hebei University of Technology, Tianjin, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Li Xing,
University of Saskatchewan, Canada
Bingqiang Liu,
Shandong University, China

***Correspondence:**
Jianwei Li
lijianwei@hebut.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 22 November 2021
Accepted: 14 December 2021
Published: 19 January 2022

Citation:
Li J, Yang Z, Wang D and Li Z (2022) WAFNRLTG: A Novel Model for Predicting LncRNA Target Genes Based on Weighted Average Fusion Network Representation Learning Method. Front. Cell Dev. Biol. 9:820342. doi: 10.3389/fcell.2021.820342

Long non-coding RNAs (lncRNAs) do not encode proteins, yet they have been well established to be involved in complex regulatory functions, and lncRNA regulatory dysfunction can lead to a variety of human complex diseases. LncRNAs mostly exert their functions by regulating the expressions of target genes, and accurate prediction of potential lncRNA target genes would be helpful to further understanding the functional annotations of lncRNAs. Considering the limitations in traditional computational methods for predicting lncRNA target genes, a novel model which was named Weighted Average Fusion Network Representation learning for predicting LncRNA Target Genes (WAFNRLTG) was proposed. First, a novel heterogeneous network was constructed by integrating lncRNA sequence similarity network, mRNA sequence similarity network, lncRNA-mRNA interaction network, lncRNA-miRNA interaction network and mRNA-miRNA interaction network. Next, four popular network representation learning methods were utilized to gain the representation vectors of lncRNA and mRNA nodes. Then, the representations of lncRNAs and target genes in the heterogeneous network were obtained with the weighted average fusion network representation learning method. Finally, we merged the representations of lncRNAs and related target genes to form lncRNA-gene pairs, trained the XGBoost classifier and predicted potential lncRNA target genes. In five-cross validations on the training and independent datasets, the experimental results demonstrated that WAFNRLTG obtained better AUC scores (0.9410, 0.9350) and AUPR scores (0.9391, 0.9350). Moreover, case studies of three common lncRNAs were performed for predicting their potential lncRNA target genes and the results confirmed the effectiveness of WAFNRLTG. The source codes and all data of WAFNRLTG can be freely downloaded at https://github.com/HGDYZW/WAFNRLTG.

Keywords: lncRNA target genes prediction, weighted average fusion network representation learning, heterogeneous network, machine learning, XGBoost

INTRODUCTION

Long non-coding RNAs (lncRNAs) are important components of non-coding RNAs whose transcript lengths exceed 200 nucleotides (Ponting et al., 2009). LncRNAs generally exhibit low cross-species conservation, low expression levels and high tissue specificity (Mercer et al., 2008; Pauli et al., 2012), and do not have the functions of protein coding (Carninci and Hayashizaki, 2007). LncRNAs can interfere with the expressions of downstream genes through base complementary pairing, and participate in most biological processes, including cell proliferation, differentiation, chromatin remodeling, epigenetic regulation, genomic splicing, transcription, translation and other aspects (Lander et al., 2001; Guttman et al., 2009; Mercer et al., 2009; Wapinski and Chang, 2011). Due to the important role of lncRNAs in biological processes, their regulatory dysfunctions are commonly associated with a variety of human diseases, especially cancers (Gupta et al., 2010; Zhang et al., 2016). Recent studies have found that lncRNAs regulate many key biological processes by interacting with their target genes. For example, the binding of lncRNA BACE1-AS with its target gene BACE1 increases the stability of BACE1, which regulates BACE1 profile and subsequently affects BACE1 protein expression (Faghihi et al., 2008). In addition, lncRNAs can also be used as the competing endogenous RNA to indirectly regulate mRNA through the shared miRNAs. For example, lncRNA HULC can competitively regulate PRKACB by sharing the common binding site of miR-372, and induce the phosphorylation of CREB in liver cancer (Qi et al., 2015). Since lncRNAs have an important role in biological processes, it prompted researchers to develop computational methods to identify lncRNA regulatory functions. Currently, these computational methods are mainly classified into two categories based on their aims, in terms of lncRNA related diseases identification and lncRNA target genes identification.

In computational models aimed on diseases identification, they can be further divided into two groups: machine learning methods and biological network methods. In recent years, machine learning has been widely applied to predict lncRNA-disease associations. These methods extract the biological features of lncRNAs and diseases and then use machine learning classifiers to infer lncRNA-related diseases. Chen et al. (Chen and Yan, 2013) developed a novel model, LRLSLDA, which predicted potential disease-related lncRNAs in a semi-supervised learning framework. In addition, LRLSLDA requires only human lncRNA expression profiles and known lncRNA-disease associations without negative samples to produce reliable results. Lan et al. (2017) proposed LDAP model to predict lncRNA-disease associations by using a bagging SVM classifier based on lncRNA similarity and disease similarity. Yao et al. (2020) implemented a random forest and feature selection-based lncRNA-disease association prediction model, RFLDA. RFLDA integrated experimentally supported miRNA-disease associations, lncRNA-disease associations, disease semantic similarity, lncRNA functional similarity and lncRNA-miRNA interactions as input features. RFLDA selected the most useful features to train the prediction model by feature selection based on the importance scores of random forest variables. Based on the hypothesis that similar diseases are more likely to be associated with similar lncRNAs, a number of biological network-based lncRNA-disease association prediction methods have been proposed. Sun et al. (2014) proposed a global network-based computational method named RWRlncD by integrating disease similarity network, lncRNAs functional similarity network and known lncRNA-disease associations. Zhou et al. (2015) proposed RWRHLD method to predict lncRNA-disease associations, which integrated three networks into a heterogeneous network and implemented a random walk on it. Deng et al. (2021) came up with a method, LDAH2V, for inferring lncRNA-disease associations by integrating lncRNA-disease associations, miRNA co-expression profiles, miRNA-disease associations, lncRNA-miRNA associations and lncRNA functional similarity. LDAH2V is a generic network-based link prediction model that can be applied to any number of entity networks.

The theoretical foundation for lncRNAs target genes prediction is the assumption that highly similar lncRNAs tend to have similar interaction. Many studies have shown that lncRNAs indirectly regulate gene expressions via adjusting expressions of miRNAs (Jones-Rhoades and Bartel, 2004). Therefore, exploring the interaction of lncRNA-miRNA would contribute to understand the complex functions of lncRNAs. Due to the rapid development of RNA sequencing technology, lncRNA-related and miRNA-related biological data are increasing rapidly. Predicting the interactions between lncRNAs and miRNAs through traditional experimental methods is very time-consuming and labor-intensive. Recently, many computational methods for predicting lncRNA-miRNA interactions have been proposed. For example, Wong et al. (2020) proposed the LNRLMI model, which constructed a bipartite network to predict potential lncRNA-miRNA interactions by combining the known interaction network and the similarity of the expression profile of lncRNA-miRNA. Zhou et al. (2019) proposed a GEEL model that constructed the lncRNA-miRNA interaction network based on the sequence features and known interactions of lncRNA and miRNA, and then used five different graph embedding methods to obtain the node representation of lncRNA/miRNA. Based on the embedding results, GEEL used individual graph embedding method-based model as basic predictors and build an ensemble model to predict the potential interactions between lncRNAs and miRNAs. Yang et al. (2020) proposed the lncMirNet model which predicted lncRNA-miRNA interactions based on hybrid sequence features. Based on these, Zhao et al. (2020) developed a method, named DeepLGP, for prioritizing lncRNA target genes via encoding gene and lncRNA features. These features were used by the convolutional neural network and were combined as the features of lncRNA-gene pairs. Finally, the CNN model was used to classify lncRNA-gene pairs into true or false pairs.

In this study, we developed a Weighted Average Fusion Network Representation Learning method-based model to improve the performance of lncRNA Target Genes prediction (WAFNRLTG). First, we constructed a heterogeneous network, which integrated two similar networks and three interaction networks. Next, the network representation learning method

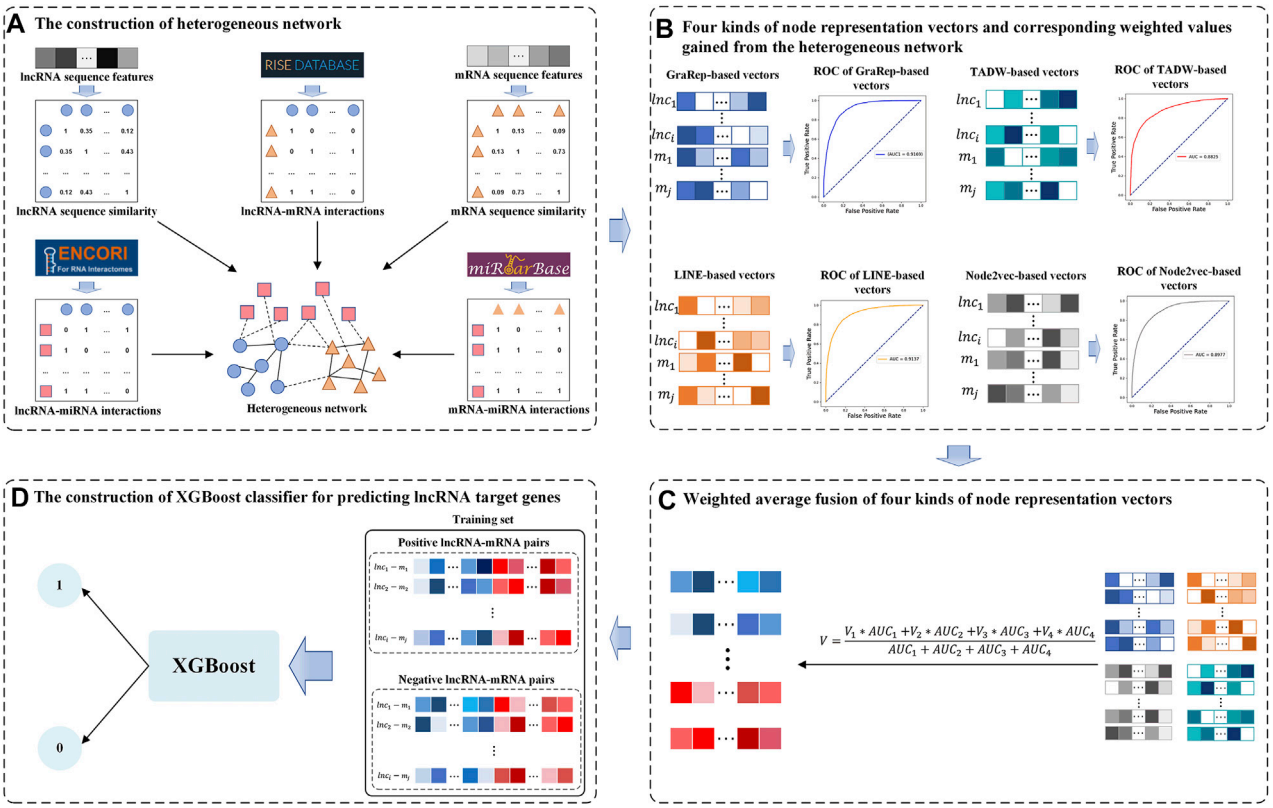


FIGURE 1 | Flow chart of WAFNRLTG. **(A)** Two sequence similar networks (lncRNA-lncRNA, mRNA-mRNA) are constructed based on the sequence features of lncRNA/mRNA and three interaction networks (lncRNA-mRNA, lncRNA-miRNA, mRNA-miRNA) are constructed based on various data sources. The heterogeneous network is established with the above five networks. **(B)** Four kinds of network representation learning methods are employed to gain node representation vectors of lncRNAs and mRNAs from the heterogeneous network. Based on these vectors, the AUC scores of the corresponding models are gained as weighted values. **(C)** The new node representation vectors are gained via fusing the four kinds of representation vectors with the weighted average method **(D)** The representation vectors of lncRNAs and mRNAs are combined to train the XGBoost classifier for predicting the target genes of lncRNAs.

was utilized to gain the representation vectors of lncRNA and mRNA nodes. Four popular network representation learning methods (GraRep, LINE, TADW, and Node2vec) were adopted in our model. Then, the weighted average method was further employed to fuse the corresponding representation vectors according to the AUC scores. The novel representation vectors were obtained which integrated different network structure information and improved the generalization ability of the model. Finally, the representation vectors of lncRNAs and the representation vectors of mRNAs were merged to form the lncRNA-gene pairs, and XGBoost classifier was built based on the merged representations of lncRNA-miRNA pairs. With five-fold cross-validations on training and independent dataset, the experimental results demonstrated that WAFNRLTG obtained AUC scores (0.9410, 0.9350) and AUPR scores (0.9391, 0.9350), and outperformed the individual network representation learning method-based models. Furthermore, three case studies were performed to evaluate the capability of WAFNRLTG. The results can be verified by the existing PubMed literatures. In conclusion, WAFNRLTG is an effective tool for predicting the potential lncRNA target genes. The flow chart of WAFNRLTG is shown in **Figure 1**.

MATERIALS AND METHODS

Datasets

In this paper, there were 5,435 validated lncRNA-gene interactions used as positive samples which were obtained from the RISE database (Hubbard et al., 2002). Negative samples were randomly selected from all unknown lncRNA-gene interactions. Because the number of unknown lncRNA-gene interactions are far more than the number of positive samples, a total of 5,435 negative samples were generated with the same number of positive samples. Eventually, we obtained 5,435 positive samples and 5,435 negative samples as a new dataset. Subsequently, five-sixth of the new dataset were randomly selected as training samples to train the classifier, and the remaining samples were used as the independent samples to evaluate each classifier.

LncRNA and mRNA Sequence Similarity Networks

The corresponding sequences of lncRNAs and mRNAs were downloaded from the Ensembl database (Hubbard et al., 2002) according to their corresponding Ensembl IDs. Linear

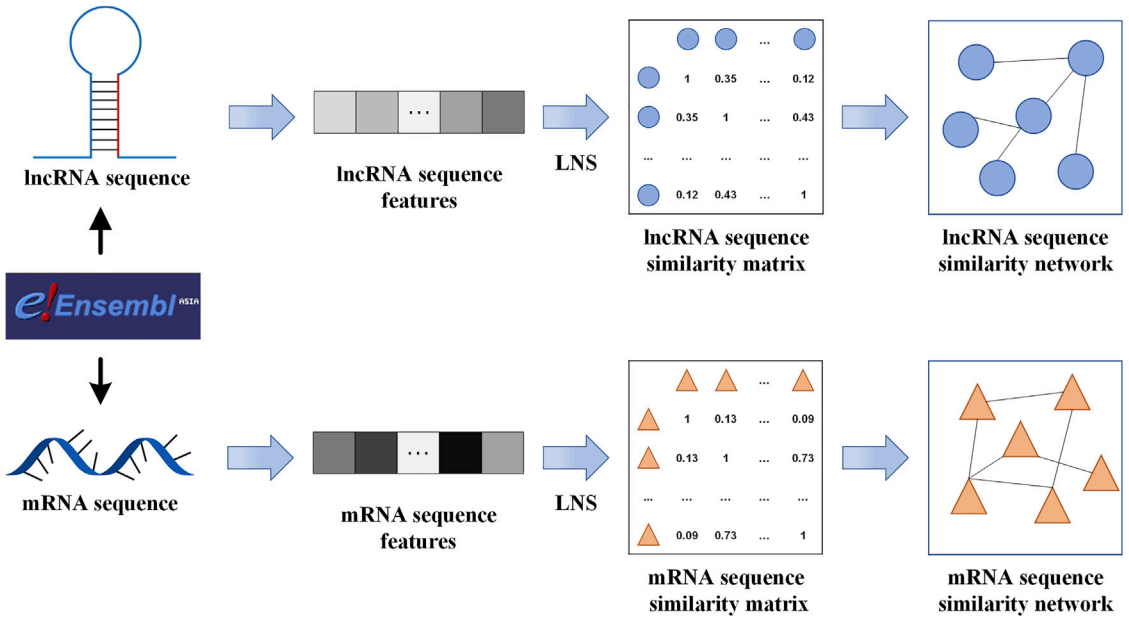


FIGURE 2 | The flowchart of constructing the lncRNA and mRNA sequence similarity network.

neighborhood similarity measure (LNS) and two lncRNA (miRNA) sequence features [k-mer (Gupta et al., 2008) and CTD (Tong and Liu, 2019)] were employed to calculate the lncRNA similarities and miRNA similarities respectively. For each RNA sequence, its k-mer frequency distribution is usually defined as the occurrence frequency of corresponding k-length contiguous subsequences. LNS is a recently proposed similarity calculation method and has been widely used in the field of bioinformatics. The 30-dimensional CTD (composition, transition, and distribution) features are used to represented RNA structure information. In order to construct lncRNA and mRNA sequence similarity networks, the 3-mers features and CTD features of a lncRNA/mRNA sequence were merged into union vectors respectively. For the lncRNA sequence similarity network, the union vectors were used to construct the lncRNA similarity matrix by LNS (Zhang et al., 2017). For example, given a specified lncRNA, its top ten lncRNAs with similarity weights greater than 0 were considered to be linked with it. Based on this strategy, closely homologous lncRNAs were linked with this lncRNA for establishing the lncRNA sequence similarity network. For the mRNA sequence similarity network, its construction procedure was the same as those of the lncRNA sequence similarity network. After this, we gained a lncRNA sequence similarity network involving 2,249 lncRNAs and mRNA sequence similarity network involving 3,785 mRNAs. The flowchart of constructing the lncRNA and mRNA sequence similarity network is shown in **Figure 2**.

LncRNA-mRNA Interaction Network

In this study, experimentally validated lncRNA-mRNA interaction data was downloaded from the RISE database (Gong et al., 2018), and it included 10,941 lncRNA-mRNA interactions. After removing the redundant data and nonhuman data, 5,435 associations involving 2,249 lncRNAs and 3,785 mRNAs were finally obtained. Therefore, the lncRNA-mRNA interaction network in our model was constructed based on these 5,435 interactions.

LncRNA-miRNA Interaction Network

First, the known lncRNA-miRNA interactions were downloaded from the ENCORI database (Li et al., 2014). Then, the duplicate interactions were removed and the only interactions between the lncRNAs which were from the lncRNA-mRNA interactions and miRNAs were preserved. In the end, the lncRNA-miRNA interaction network was constructed with 6,053 lnRNA-miRNA interactions between 2,249 lncRNAs and 636 miRNAs.

MRNA-miRNA Interaction Network

We downloaded the known mRNA-miRNA interactions from the miTarbese database (Huang et al., 2020). Then, the duplicate interactions were removed and the only interactions between the mRNAs which were from the lncRNA-mRNA interactions and the miRNAs which were from the lncRNA-miRNA interactions were retained. Ultimately, the constructed mRNA-miRNA network in our study contained 1983 mRNA-miRNA interactions between 305 mRNAs and 636 miRNAs.

Network Representation Learning of the Heterogeneous Network

Recently, many Network Representation Learning (NRL) methods have been proposed (Zhang et al., 2020), of which main purpose is to find a proper mapping function to map large-scale, high-dimensional, sparse vectors into a low-dimensional, dense

semantic space, while keep the proximity of these low-dimensional vector representations to the original network. NRL has attracted the attention of scholars in the fields of data mining of biological information data. The low-dimensional representation learned from the network representation is applied to downstream network analysis tasks, such as node classification (Tang et al., 2016a), link prediction, association mining (Zhao et al., 2019), information recommendation (Han et al., 2018) and network visualization (Tang et al., 2016b).

In our study, four state-of-the-art network representation learning methods [GraRep (Cao et al., 2015), LINE (Tang et al., 2015), TADW (Yang et al., 2015) and Node2vec (Grover and Leskovec, 2016)] were used to learn the representation vectors of lncRNA nodes and mRNA nodes for making full use of the various useful information in the heterogeneous network.

LINE maintains both first-order and second-order proximity during learning node representations. Given an undirected edge (i, j), the joint probability of node v_i and node v_j is as follows:

$$p_1(v_i, v_j) = \frac{1}{1 + \exp\left(-\vec{u}_i^T \cdot \vec{u}_j\right)} \tag{1}$$

where $\vec{u}_j \in R^d$ is the low-dimensional vector representation of node v_i. The empirical probability of the distribution $p(\cdot, \cdot)$ on the space V×V is:

$$\hat{p}_i(i, j) = \frac{w_{ij}}{W} \tag{2}$$

where $W = \sum_{(i,j)\in E} w_{ij}$ and w_{ij} is the weight of edge (i, j).

After optimizing the model by minimizing the KL scatter of the two distributions, the objective function is defined as follows:

$$O_1 = -\sum_{(i,j)\in E} w_{ij} \log p_1(v_i, v_j) \tag{3}$$

The second-order similarity scenario assumes that nodes sharing a large number of connections with other nodes are similar to each other, and each node is considered as a specific context, then nodes with similar distribution on the context are similar. Here, two vectors $\vec{u}_i$ and $\vec{u}_i^{'}$ are introduced, where $\vec{u}_i$ is the representation of v_i when it is treated as a vertex and $\vec{u}_i^{'}$ is the representation of v_i when it is treated as a specific "context". For an undirected edge (i, j), the probability of generating the context v_j from v_i is:

$$p_2(v_j|v_i) = \frac{\exp\left(\vec{u}_j^{'T} \cdot \vec{u}_i\right)}{\sum_{k=1}^{|V|} \exp\left(\vec{u}_k^{'T} \cdot \vec{u}_i\right)} \tag{4}$$

where $|V|$ denotes the number of nodes or contexts. The empirical distribution $\hat{p}_2(\cdot|v_i)$ is defined as follows:

$$\hat{p}_2(v_j, v_i) = \frac{w_{ij}}{d_i} \tag{5}$$

where w_{ij} is the weight of edge (i, j) and d_i is the out-degree of node v_i. d_i is used in LINE as the importance of nodes λ_i. By using the KL scatter while ignoring some constants, the objective function is obtained as follows:

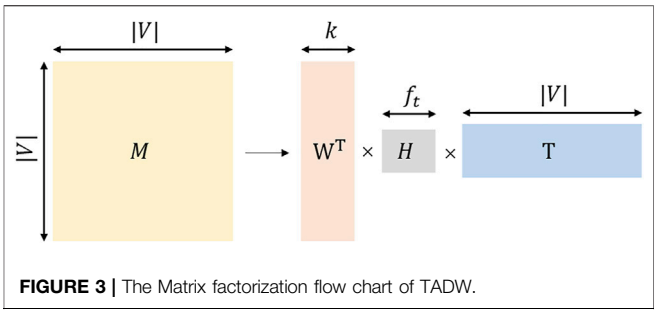


FIGURE 3 | The Matrix factorization flow chart of TADW.

$$O_2 = -\sum_{(i,j)\in E} w_{ij} \log p_2(v_j, v_i) \tag{6}$$

LINE also adopted negative sampling to optimize the model, while using the Alias method to accelerate the sampling process.

GraRep extends LINE by learning the k-order relational vector representations of the network nodes separately through matrix factorization and combines the k-order relational vector representations as the final representation. For a network G, the degree matrix D of the network is defined using the adjacency matrix S. The first-order transfer probability matrix is defined as follows:

$$A = D^{-1}S \tag{7}$$

where $A_{i,j}$ denotes the probability of transferring from v_i to v_j by one step.

Then using Skip-Gram and NCE (noise contrastive estimation) methods, for a transfer of order k, the model can be reduced to the decomposition problem with matrix $Y_{i,j}^k$.

For a transfer of order k, the model is then reduced to a decomposition problem of matrix $Y_{i,j}^k$ by using the Skip-Gram and NCE (noise contrastive estimation) methods.

$$Y_{i,j}^k = W_i^k \cdot C_j^k = \log\left(\frac{A_{i,j}^k}{\sum_t A_{t,j}^k}\right) - \log(\beta) \tag{8}$$

where $\beta = \lambda/N$, λ is the number of negative samples and N is the number of edges in network G.

Node2vec designs a biased random wander over a scalable node neighborhood and explores different node neighborhoods by using breadth first search (BFS) and depth first search (DFS), and then inputs the resulting node sequences as sentences into the skip-gram model to learn node representations. For a random wandering sequence $v_{i-w}, \ldots, v_{i-1}, v_i, v_{i+1}, \ldots, v_{i+m}$, centered on v_i with windows size w. Node2vec uses the Skip-Gram algorithm to optimize the model.

$$Pr(\{v_{i-w}, \ldots, v_{i+w}\} \backslash v_i | \Phi(v_i)) = \prod_{j=i-w, j\neq i}^{i+w} \backslash Pr(v_j | \Phi(v_i)) \tag{9}$$

TADW utilizes the inductive matrix decomposition method based on integrating the network topology information and

textual information. The Matrix factorization flow chart of TADW is shown in **Figure 3**.where V denotes the set of nodes, $W \in R^{k\times|V|}$, $M \in R^{|V|\times|V|}$, $H \in R^{k\times f_t}$, and $T \in R^{f_t\times|V|}$, and T is the text feature matrix. Therefore, the loss function of TADW as follows:

$$\min_{W,H} \left\|M - W^T HT\right\|_F^2 + \frac{\lambda}{2}\left(\|W\|_F^2 + \|H\|_F^2\right) \quad (10)$$

These four state-of-the-art network representation learning methods can capture valuable information of the structure and intrinsic properties of the heterogeneous network. The learned representations of lncRNAs and mRNAs were further utilized to construct WAFNRLTG.

Weighted Average Fusion Node Representation

The processes of fusing the node representations derived from the network representation learning methods (Grarep, LINE, TADW, and Node2vec) are described below. First, low-dimensional representations of lncRNAs and mRNAs were obtained from the heterogeneous network. Thus, the four representation vectors of nodes were obtained with different methods were V_1, V_2, V_3 and V_4, and the corresponding prediction models were constructed based on these vectors. The AUC score of each model was calculated as the weighted value of the corresponding methods, they were labeled as AUC_1, AUC_2, AUC_3 and AUC_4. To make full use of the information in node representation vectors and improve the generalization ability of the model, the new representation vector V was obtained by the weighted average fusion of the four representation vectors.

$$V = \frac{V_1 * AUC_1 + V_2 * AUC_2 + V_3 * AUC_3 + V_4 * AUC_4}{AUC_1 + AUC_2 + AUC_3 + AUC_4} \quad (11)$$

RESULTS

Parameter and Experimental Settings

In this study, four network representation learning methods (GraRep, LINE, TADW, and Node2vec) were employed to obtain node representation vectors from the heterogeneous network.

First, we took the dimensions of the node representation vectors as the common parameter of these four methods. Tuning dimensions d, and the d -dimensional features that produced the best AUC were selected. The experimental results under node representation vectors of different dimensions are shown in **Table 1**.

The other parameters of the network representation learning method are discussed in the following. GraRep has a parameter: k-step k, which indicates the k-step transfer matrix. Node2vec has four tunable parameters: number-walks n, indicates the number of random walks from each node; walk-length l, indicates the length of a random walk from each node; p and q control the probability of the random walk to the next node. The combination of number-walks n and walk-length l were considered, and the rest of the parameter was set as defaults. TADW has a parameter: λ, which controls the weight of the regularization term. In LINE, a parameter: Order o, denotes the order of proximity was considered. Above all, we adjusted different parameter values and adopt the optimal values which produced the best AUC scores. The parameter settings for various of the network representation learning method are shown in **Table 2**. For more details, please see **Supplementary File S1**.

TABLE 1 | AUC scores of models with different dimension node representation vectors.

Method	$d = 16$	$d = 32$	$d = 64$	$d = 128$	$d = 256$
GraRep	0.8725	0.8991	0.9095	0.9147	0.8957
LINE	0.8375	0.8939	0.9177	0.9123	0.8828
TADW	0.7974	0.8466	0.8803	0.8397	0.9002
Node2vec	0.8709	0.8978	0.9131	0.9099	0.9115

TABLE 2 | Parameters settings for network representation learning methods.

Method	Optimal parameter value
GraRep	d: 128, k: 4
LINE	d: 64, o: 3
TADW	d: 256, λ:0.1
Node2vec	d: 64, n: 20, l:80

Comparison With Four Network Representation Learning Methods

In this section, we first constructed four models based on four representation network representation learning methods (Grarep, LINE, TADW, and Node2vec) and evaluated their effectiveness. In order to make full use of four network representation learning methods, the AUC scores of each method were adopted as weight values, and the four representation vectors were fused by the weighted average method. According to these constructed representation vectors, XGBoost classifier (Chen and Guestrin, 2016) was selected into WAFNRLTG.

Based on these four representations from the above network representation learning models, we subsequently fused them by the weighted average method to improve performance and generalization of WAFNRLTG. For assessing the prediction performance of WAFNRLTG, five-fold cross-validations experiments were used to evaluate the classification performances and the four network representation learning models. In our study, seven commonly metrics, namely Sensitivity (SN), Specificity (SP), Precision (PREC), Accuracy (ACC), Matthews correlation coefficient (MCC), AUC, and AUPR, were employed as evaluation metrics. ROC (receiver operating characteristic curve, ROC) and PR (Precision-Recall curve, PR) curves were plotted for showing the different performance of each model. The results of WAFNRLTG and

TABLE 3 | Performances of the five network representation learning models.

Method	Acc	Sen	Spec	Prec	MCC	AUC	AUPR
GraRep	0.8107	0.7920	0.8295	0.8229	0.6220	0.9147	0.9097
LINE	0.8358	0.8275	0.8441	0.8413	0.6718	0.9177	0.9158
TADW	0.8223	0.7798	0.8648	0.8523	0.6471	0.9002	0.9053
Node2vec	0.8305	0.8328	0.8282	0.8290	0.6611	0.9131	0.9136
WAFNRLTG	0.8683	0.8638	0.8728	0.8716	0.7362	0.9410	0.9391

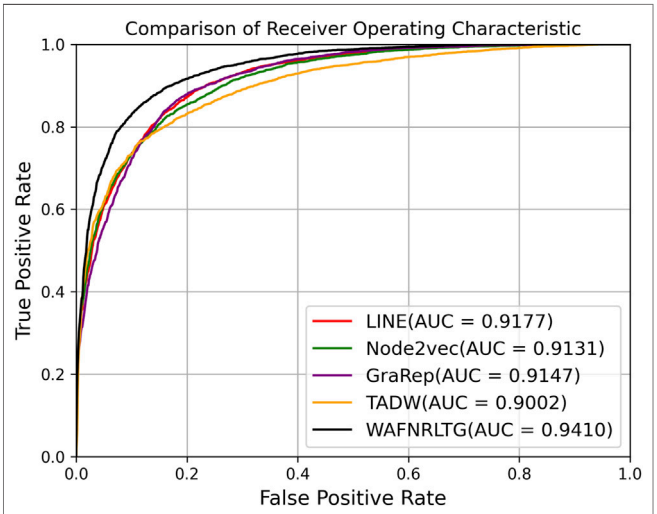


FIGURE 4 | ROC curves of five models in five-fold cross-validations.

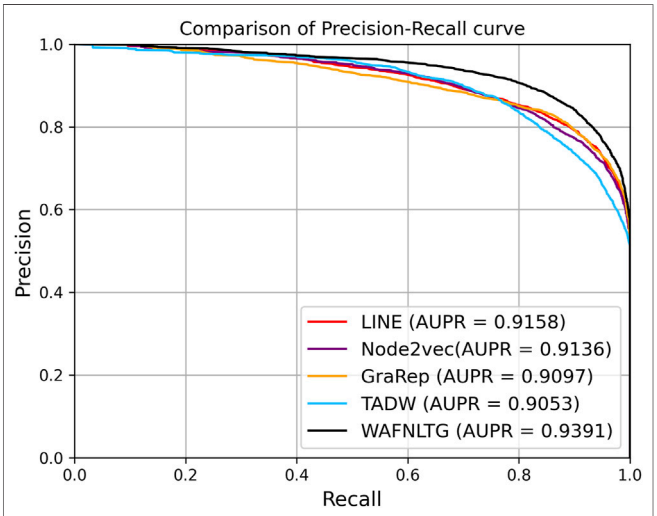


FIGURE 5 | PR curves of five models in five-fold cross-validations.

the four network representation learning models are shown in **Table 3**. It can been seen from the table that WAFNRLTG achieves AUC score of 0.9410 and AUPR score of 9,391, which outperforms GraRep (AUC score: 0.9147; AUPR score: 0.9097), LINE (AUC score: 9,177; AUPR score: 0.9158), TADW (AUC score: 0.9002; AUPR score: 0.9053) and Node2vec (AUC score: 0.9131; AUPR score: 0.9136). ROC and PR curves are plotted for the five models to further display their different performances, **Figure 4** shows five ROC curves and **Figure 5** shows five PR curves.

For comparative analysis, we also adopted the other two fusion methods (concatenate and average) to fuse the four node representation vectors similar to the weighted average method. The experimental results of the three fusion methods are recorded in **Table 4**. As shown in **Table 4**, the weighted average fusion method outperforms other methods.

Moreover, the information extracted from heterogeneous networks by WAFNRLTG are brought to the subsequent work for predicting of lncRNA target genes.

In order to evaluate the generalization ability of WAFNRLTG, we applied it on the training dataset and the independent dataset. Experimental results are shown in **Table 5**. As exhibited from **Table 5**, the results on the independent dataset are comparable to the results on the training dataset. The experimental results on the independent dataset demonstrate that WAFNRLTG is a robust and reliable model for predicting potential lncRNA target genes.

Comparison Among Different Classifiers

After the acquisition of weighted averages fused representation vectors, we compared five different machine learning methods and selected the most appropriate one. These five classification methods include K Nearest Neighbor (KNN), AdaBoost (Freund and Schapire, 1997), Support Vector Machine (SVM), Gradient Boosting Decision Tree (GBDT) and XGBoost, which are well known to perform well on a variety of tasks. They were evaluated by five-fold cross-validations. **Figure 6** shows their prediction performances. The AUC scores of KNN, AdaBoost, SVM, GBDT and XGBoost are 0.9082, 0.8717, 0.8959, 0.8462, and 0.9394, respectively. The detailed results of these five classifiers are shown in **Table 6**. From the analysis of above results, XGBoost model achieved the best performance.

Case Studies

The main objective of this study is to screen potential target genes of lncRNAs and guide relevant researchers to explore novel target genes. For further evaluating the performance of WAFNRLTG model in practical applications, we selected three common lncRNAs (MALAT1, PVT1, and NEAT1) as case studies. The general processes of each of case studies were as following. First, all lncRNA-mRNA interactions from our dataset were utilized to construct the WAFNRLTG model. Then, the interactions between above three lncRNAs and the other mRNAs were adopted as the dataset in WAFNRLTG model. The model

TABLE 4 | The experimental results of the three fusion methods.

Method	ACC	SEN	SPEC	PREC	MCC	AUC	AUPR
Concatenate	0.86	0.8479	0.8722	0.8691	0.7203	0.9391	0.9307
Average	0.8515	0.8412	0.8618	0.8589	0.7033	0.9296	0.9277
Weighted average	0.8683	0.8638	0.8728	0.8716	0.7362	0.9410	0.9391

TABLE 5 | Performances of WAFNRLTG on the independent dataset and the training dataset.

Dataset	Acc	Sen	Spec	Prec	MCC	AUC	AUPR
independent dataset	0.8666	0.8567	0.8765	0.874	0.7333	0.9358	0.9350
training dataset	0.8683	0.8638	0.8728	0.8716	0.7362	0.9410	0.9391

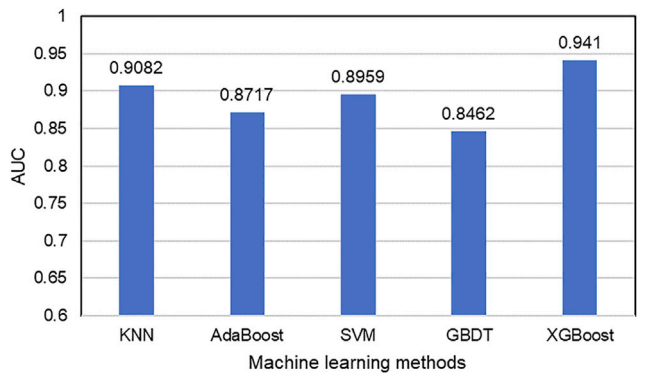


FIGURE 6 | Comparisons of AUC scores of five classifiers.

TABLE 6 | The detailed experimental results of the five classifiers.

Classifier	ACC	SEN	SPEC	PREC	MCC	AUC	AUPR
GBDT	0.8235	0.8037	0.8433	0.8369	0.6475	0.9082	0.9039
KNN	0.7197	0.7604	0.679	0.7037	0.4412	0.8012	0.8005
RF	0.785	0.7761	0.794	0.7907	0.5703	0.8718	0.8745
SVM	0.8063	0.784	0.8284	0.8204	0.6131	0.8959	0.8961
XGBoost	0.8683	0.8638	0.8728	0.8716	0.7362	0.9410	0.9391

outputted the predicting scores of the lncRNA-mRNA pairs. Finally, the top ten mRNAs were selected for literature mining on PubMed. We found that four of the top ten mRNAs corresponding to MALAT1 and NEAT1 were validated in the literature, and three of the top ten mRNAs corresponding to PVT1 were validated in the literatures.

MALAT1 is one of the first lncRNA discovered that are associated with human diseases. Many studies have demonstrated that the abnormal expression of MALAT1 is closely related to cancer pathophysiology, and has the potential to be translated clinically. MALAT1 regulates cancer processes by interacting with molecules, such as proteins, RNAs and DNAs, and further alters different signal pathways. To

TABLE 7 | The predicting target genes in the top ten for MALAT1.

Rank	Score	Target genes	PMID
1	0.9952035	MYC	33312756
2	0.953083	MT-ND4L	Not found
3	0.952141	LRRC2	Not found
4	0.947306	TET2	33165916
5	0.92575	ECT2	27313681
6	0.923501	OR1M1	Not found
7	0.917459	ATXN2L	Not found
8	0.917261	TP73AS1	32714991
9	0.915255	NFIC	Not found
10	0.908282	FASTK	Not found

demonstrate the ability of our model for predicting potential lncRNA target genes, we predicted the top ten mRNAs that interact with MALAT1, as shown in **Table 7**. With literature mining, we found that four mRNAs interact with MALAT1 and they can be used as the target genes of MALAT1. For example, abnormal expression of MALAT1 leads to reduced expression of TET2 thus causing neuronal damage, so there may be a targeting relationship between MALAT1 and TET2 (Li et al., 2021). The complete case study results of MALAT1 are available in **Supplementary File S2**.

NEAT1 expression is upregulated in many human malignancies, such as lung, esophageal and gastric cancers. In order to demonstrate that WAFNRLTG is effective in predicting potential target genes of lncRNAs, we predicted the top 10 mRNAs associated with NEAT1, and the results are shown in **Table 8**. After literature mining, four of them were shown to be target genes of NEAT1. For example, it was found that MYC-regulated NEAT1 promoted diffuse large B-cell lymphoma (DLBCL) proliferation via the miR-34b-5p-GLI1 pathway, which could provide a novel therapeutic target for DLBCL (Qian et al., 2020). The complete case study results of NEAT1 are available in **Supplementary File S3**.

Plasmacytoma variant translocation 1 (PVT1) is a newly discovered long non-coding RNA which preforms regulating functions as an oncogenic molecule in different cancers. In order to understand the functions of PVT1, WAFNRLTG

TABLE 8 | The predicting target genes in the top ten for NEAT1.

Rank	Score	Target genes	PMID
1	0.975445	SUSD6	Not found
2	0.961867	CPN2	Not found
3	0.952019	PEX26	Not found
4	0.951388	WNT9A	Not found
5	0.942409	SRP19	Not found
6	0.931132	MYC	32206038
7	0.926551	OAS3	33138195
8	0.925127	CPSF6	22960638
9	0.924981	QSOX2	Not found
10	0.910639	TET2	33987091

TABLE 9 | The predicting target genes in the top ten for PVT1.

Rank	Score	Target genes	PMID
1	0.971728	NANOG	34230224
2	0.9714675	GID4	Not found
3	0.970239	WNT9B	Not found
4	0.969168	TRPM3	Not found
5	0.966868	COL5A2	33750300
6	0.966181	RCC2	Not found
7	0.963472	RBM7	Not found
8	0.961825	WNT3A	32727463
9	0.957906	FAM101B	Not found
10	0.953643	CCDC115	Not found

predicted its target genes. We conducted a literature survey of the top ten mRNAs predicted by WAFNRLTG to interact with PVT1 and found that three mRNAs were verified to be its target genes, as shown in **Table 9**. For example, the abnormal expression of PVT1 affects the expression of NANOG and thus makes difference in the development of glioma (Gong et al., 2021). The complete case study results of PVT1 are available in **Supplementary File S4**.

DISCUSSION

LncRNAs and its target genes are involved in a variety of biological processes and are closely associated with serious human diseases. Predicting the potential lncRNA target genes can decipher complex biological mechanisms and reveal the functions of lncRNAs. In this paper, we firstly collected and processed multiple data from the multiple open databases, including lncRNA-mRNA interactions, lncRNA-miRNA interactions and mRNA-miRNA interactions. The lncRNA sequence similarity network and mRNA sequence similarity network were constructed based on sequence features. Then, we proposed a novel model, WAFNRLTG, to infer potential lncRNA target genes by integrating above data. In five-fold cross-validations on training and independent dataset, the experimental results demonstrated that WAFNRLTG obtained AUC scores (0.9410, 0.9350) and AUPR scores (0.9391, 0.9350). Three common lncRNAs (MALAT1, NEAT1, and PVT1) were introduced to WAFNRLTG model. Several target genes in the predicted results were found according to experimental PubMed literatures.

Although WAFNRLTG has achieved satisfactory results in predicting lncRNA target genes, it is still necessary to point out there are still limitations in our model. For example, the negative samples randomly selected from unknown lncRNA-mRNA interactions may have the chance of becoming positive samples. This would have an impact on the accuracy of WAFNRLTG. The information extracted from the heterogeneous network is not comprehensive, some valuable information may be neglected. With the accumulation of biological data, more interaction records would be introduced to enrich the heterogeneous network and improve the prediction ability of WAFNRLTG.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JL conceived, designed the study, ZY and DW developed the algorithm and performed the statistical analysis, ZY and ZL wrote the codes. ZY drafted the original manuscript, JL revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China under Grants No. 81672113, 62072154, 61806072.

ACKNOWLEDGMENTS

We thank members in our groups for their valuable discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.820342/full#supplementary-material

Supplementary File S1 | The experimental details of WAFNRLTG.

Supplementary File S2 | The prediction of MALAT1 target genes results of WAFNRLTG.

Supplementary File S3 | The prediction of NEAT1 target genes results of WAFNRLTG.

Supplementary File S4 | The prediction of PVT1 target genes results of WAFNRLTG.

REFERENCES

Cao, S., Lu, W., and Xu, Q. (2015). "GraRep: Learning Graph Representations with Global Structural Information," in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (Melbourne, Australia: Association for Computing Machinery), 891–900. doi:10.1145/2806416.2806512

Carninci, P., and Hayashizaki, Y. (2007). Noncoding RNA Transcription beyond Annotated Genes. *Curr. Opin. Genet. Development* 17 (2), 139–144. doi:10.1016/j.gde.2007.02.008

Chen, T., and Guestrin, C. (2016). "XGBoost," in *Proceedings Of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining)* (San Francisco: Association for Computing Machinery), 785–794. doi:10.1145/2939672.2939785

Chen, X., and Yan, G.-Y. (2013). Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics* 29 (20), 2617–2624. doi:10.1093/bioinformatics/btt426

Deng, L., Li, W., and Zhang, J. (2021). LDAH2V: Exploring Meta-Paths across Multiple Networks for lncRNA-Disease Association Prediction. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18 (4), 1572–1581. doi:10.1109/TCBB.2019.2946257

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., et al. (2008). Expression of a Noncoding RNA Is Elevated in Alzheimer's Disease and Drives Rapid Feed-Forward Regulation of β-secretase. *Nat. Med.* 14 (7), 723–730. doi:10.1038/nm1784

Freund, Y., and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Computer Syst. Sci.* 55 (1), 119–139. doi:10.1006/jcss.1997.1504

Gong, J., Shao, D., Xu, K., Lu, Z., Lu, Z. J., Yang, Y. T., et al. (2018). RISE: a Database of RNA Interactome from Sequencing Experiments. *Nucleic Acids Res.* 46 (D1), D194–D201. doi:10.1093/nar/gkx864

Gong, R., Li, Z.-Q., Fu, K., Ma, C., Wang, W., and Chen, J.-C. (2021). Long Noncoding RNA PVT1 Promotes Stemness and Temozolomide Resistance through miR-365/ELF4/SOX2 Axis in Glioma. *Exp. Neurobiol.* 30 (3), 244–255. doi:10.5607/en20060

Grover, A., and Leskovec, J. (2016). "node2vec," in *Proceedings Of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco: Association for Computing Machinery), 855–864. doi:10.1145/2939672.2939754

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010). Long Non-coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis. *Nature* 464 (7291), 1071–1076. doi:10.1038/nature08975

Gupta, S., Dennis, J., Thurman, R. E., Kingston, R., Stamatoyannopoulos, J. A., and Noble, W. S. (2008). Predicting Human Nucleosome Occupancy from Primary Sequence. *Plos Comput. Biol.* 4 (8), e1000134. doi:10.1371/journal.pcbi.1000134

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., et al. (2009). Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-coding RNAs in Mammals. *Nature* 458 (7235), 223–227. doi:10.1038/nature07672

Han, X., Shi, C., Zheng, L., Yu, P. S., Li, J., and Lu, Y. (2018). "Representation Learning with Depth and Breadth for Recommendation Using Multi-View Data," in *Web and Big Data* (Springer International Publishing), 181–188. doi:10.1007/978-3-319-96890-2_15

Huang, H.-Y., Lin, Y.-C. -D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2020). miRTarBase 2020: Updates to the Experimentally Validated microRNA-Target Interaction Database. *Nucleic Acids Res.* 48 (D1), D148–D154. doi:10.1093/nar/gkz896

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., et al. (2002). The Ensembl Genome Database Project. *Nucleic Acids Res.* 30 (1), 38–41. doi:10.1093/nar/30.1.38

Jones-Rhoades, M. W., and Bartel, D. P. (2004). Computational Identification of Plant microRNAs and Their Targets, Including a Stress-Induced miRNA. *Mol. Cell* 14 (6), 787–799. doi:10.1016/j.molcel.2004.05.027

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F.-X., Pan, Y., et al. (2017). LDAP: a Web Server for lncRNA-Disease Association Prediction. *Bioinformatics* 33 (3), btw639–460. doi:10.1093/bioinformatics/btw639

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial Sequencing and Analysis of the Human Genome. *Nature* 409 (6822), 860–921. doi:10.1038/35057062

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucl. Acids Res.* 42 (Database issue), D92–D97. doi:10.1093/nar/gkt1248

Li, L., Miao, M., Chen, J., Liu, Z., Li, W., Qiu, Y., et al. (2021). Role of Ten Eleven Translocation-2 (Tet2) in Modulating Neuronal Morphology and Cognition in a Mouse Model of Alzheimer's Disease. *J. Neurochem.* 157 (4), 993–1012. doi:10.1111/jnc.15234

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long Non-coding RNAs: Insights into Functions. *Nat. Rev. Genet.* 10 (3), 155–159. doi:10.1038/nrg2521

Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., and Mattick, J. S. (2008). Specific Expression of Long Noncoding RNAs in the Mouse Brain. *Proc. Natl. Acad. Sci.* 105 (2), 716–721. doi:10.1073/pnas.0706729105

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., et al. (2012). Systematic Identification of Long Noncoding RNAs Expressed during Zebrafish Embryogenesis. *Genome Res.* 22 (3), 577–591. doi:10.1101/gr.133009.111

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. *Cell* 136 (4), 629–641. doi:10.1016/j.cell.2009.02.006

Qi, X., Zhang, D.-H., Wu, N., Xiao, J.-H., Wang, X., and Ma, W. (2015). ceRNA in Cancer: Possible Functions and Clinical Implications. *J. Med. Genet.* 52 (10), 710–718. doi:10.1136/jmedgenet-2015-103334

Qian, C.-S., Li, L.-J., Huang, H.-W., Yang, H.-F., and Wu, D.-P. (2020). MYC-regulated lncRNA NEAT1 Promotes B Cell Proliferation and Lymphomagenesis via the miR-34b-5p-GLI1 Pathway in Diffuse Large B-Cell Lymphoma. *Cancer Cell Int* 20, 87. doi:10.1186/s12935-020-1158-6

Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., et al. (2014). Inferring Novel lncRNA-Disease Associations Based on a Random Walk Model of a lncRNA Functional Similarity Network. *Mol. Biosyst.* 10 (8), 2074–2081. doi:10.1039/C3MB70608G

Tang, J., Aggarwal, C., and Liu, H. (2016a). "Node Classification in Signed Social Networks," in *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)* (Miami: Society for Industrial and Applied Mathematics), 54–62. doi:10.1137/1.9781611974348.7

Tang, J., Liu, J., Zhang, M., and Mei, Q. (2016b). "Visualizing Large-Scale and High-Dimensional Data," in *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec: International World Wide Web Conferences Steering Committee), 287–297. doi:10.1145/2872427.2883041

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "LINE: Large-scale Information Network Embedding," in Proceedings of the 24th international conference on world wide web (Florence, Italy: International World Wide Web Conferences Steering Committee), 1067–1077. doi:10.1145/2736277.2741093

Tong, X., and Liu, S. (2019). CPPred: Coding Potential Prediction Based on the Global Description of RNA Sequence. *Nucleic Acids Res.* 47 (8), e43. doi:10.1093/nar/gkz087

Wapinski, O., and Chang, H. Y. (2011). Long Noncoding RNAs and Human Disease. *Trends Cell Biol.* 21 (6), 354–361. doi:10.1016/j.tcb.2011.04.001

Wong, L., Huang, Y. A., You, Z. H., Chen, Z. H., and Cao, M. Y. (2020). LNRLMI: Linear Neighbour Representation for Predicting lncRNA-miRNA Interactions. *J. Cell Mol Med* 24 (1), 79–87. doi:10.1111/jcmm.14583

Yang, C., Liu, Z., Zhao, D., Sun, M., and Chang, E. (2015). "Network Representation Learning with Rich Text Information," in *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina: AAAI Press), 2111–2117.

Yang, S., Wang, Y., Lin, Y., Shao, D., He, K., and Huang, L. (2020). LncMirNet: Predicting LncRNA-miRNA Interaction Based on Deep Learning of Ribonucleic Acid Sequences. *Molecules* 25 (19), 4372. doi:10.3390/molecules25194372

Yao, D., Zhan, X., Zhan, X., Kwoh, C. K., Li, P., and Wang, J. (2020). A Random forest Based Computational Model for Predicting Novel lncRNA-Disease Associations. *BMC Bioinformatics* 21 (1), 126. doi:10.1186/s12859-020-3458-1

Zhang, D., Yin, J., Zhu, X., and Zhang, C. (2020). Network Representation Learning: A Survey. *IEEE Trans. Big Data* 6 (1), 3–28. doi:10.1109/TBDATA.2018.2850013

Zhang, K., Luo, Z., Zhang, Y., Zhang, L., Wu, L., Liu, L., et al. (2016). Circulating lncRNA H19 in Plasma as a Novel Biomarker for Breast Cancer. *Cbm* 17 (2), 187–194. doi:10.1016/j.molcel.2004.05.02710.3233/cbm-160630

Zhang, W., Yue, X., Liu, F., Chen, Y., Tu, S., and Zhang, X. (2017). A Unified Frame of Predicting Side Effects of Drugs by Using Linear Neighborhood Similarity. *BMC Syst. Biol.* 11 (6), 101–134. doi:10.1186/s12918-017-0477-2

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428

Zhao, Z., Li, C., Zhang, X., Chiclana, F., and Viedma, E. H. (2019). An Incremental Method to Detect Communities in Dynamic Evolving Social Networks. *Knowledge-Based Syst.* 163, 404–415. doi:10.1016/j.knosys.2018.09.002

Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., et al. (2015). Prioritizing Candidate Disease-Related Long Non-coding RNAs by Walking on the Heterogeneous lncRNA and Disease Network. *Mol. Biosyst.* 11 (3), 760–769. doi:10.1039/c4mb00511b

Zhou, S., Yue, X., Xu, X., Liu, S., Zhang, W., and Niu, Y. (2019). "LncRNA-miRNA Interaction Prediction from the Heterogeneous Network through Graph Embedding Ensemble Learning," in Proceeding of the 2019IEEE International Conference on Bioinformatics and Biomedicine (BIBM)), San Diego, CA, USA, 8-21 Nov. 2019 (IEEE), 622–627. doi:10.1109/BIBM47256.2019.8983044

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Li, Yang, Wang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 26 January 2022
doi: 10.3389/fcell.2022.825272

Identification of Novel Lung Cancer Driver Genes Connecting Different Omics Levels With a Heat Diffusion Algorithm

Fei Yuan[1†], Xiaoyu Cao[2†], Yu-Hang Zhang[3†], Lei Chen[4], Tao Huang[5,6], ZhanDong Li[7]* and Yu-Dong Cai[8]**

[1]Department of Science and Technology, Binzhou Medical University Hospital, Binzhou, China, [2]Department of Neurology, Binzhou Medical University Hospital, Binzhou, China, [3]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [4]College of Information Engineering, Shanghai Maritime University, Shanghai, China, [5]Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [6]CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [7]College of Food Engineering, Jilin Engineering Normal University, Changchun, China, [8]School of Life Sciences, Shanghai University, Shanghai, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Juntao Li,
Henan Normal University, China
Meijing Li,
Shanghai Maritime University, China

***Correspondence:**
Tao Huang
tohuangtao@126.com
ZhanDong Li
lizd591@jlenu.edu.cn
Yu-Dong Cai
cai_yud@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *30 November 2021*
Accepted: *06 January 2022*
Published: *26 January 2022*

Citation:
Yuan F, Cao X, Zhang Y-H, Chen L, Huang T, Li Z and Cai Y-D (2022) Identification of Novel Lung Cancer Driver Genes Connecting Different Omics Levels With a Heat Diffusion Algorithm. Front. Cell Dev. Biol. 10:825272. doi: 10.3389/fcell.2022.825272

Cancer driver gene is a type of gene with abnormal alterations that initiate or promote tumorigenesis. Driver genes can be used to reveal the fundamental pathological mechanisms of tumorigenesis. These genes may have pathological changes at different omics levels. Thus, identifying cancer driver genes involving two or more omics levels is essential. In this study, a computational investigation was conducted on lung cancer driver genes. Four omics levels, namely, epigenomics, genomics, transcriptomics, and post-transcriptomics, were involved. From the driver genes at each level, the Laplacian heat diffusion algorithm was executed on a protein–protein interaction network for discovering latent driver genes at this level. A following screen procedure was performed to extract essential driver genes, which contained three tests: permutation, association, and function tests, which can exclude false-positive genes and screen essential ones. Finally, the intersection operation was performed to obtain novel driver genes involving two omic levels. The analyses on obtained genes indicated that they were associated with fundamental pathological mechanisms of lung cancer at two corresponding omics levels.

Keywords: lung cancer, driver gene, epigenomics, genomics, transcriptomics, posttranscriptomics, heat diffusion algorithm, protein-protein interaction network

INTRODUCTION

Driver gene is a commonly used description in oncology to describe genes with abnormal alterations that initiate or promote tumorigenesis (Pao and Girard, 2011; Tokheim et al., 2016). Identifying driver genes can help us reveal the fundamental pathological mechanisms of tumorigenesis (Pao and Girard, 2011). During tumorigenesis, genes may have pathological changes at different omics levels, including but not restricted to genomics (as DNA sequence alterations), epigenomics (as methylation status or other DNA modification status alterations), transcriptomics, and proteomics (Cancer Genome Atlas Research Network, 2014; Huang et al., 2015; Li et al., 2018). Using current biological

techniques is time consuming but can detect all pathological alterations associated with tumorigenesis. However, not all alterations are associated with driver genes. Genes that are mutated or abnormally regulated during tumorigenesis but not associated with the initiation or progression of tumors are summarized as passenger genes (Pon and Marra, 2015). Most of the genes altered during tumorigenesis are actually passenger genes, which cannot help us understand or reveal the pathological mechanisms of cancer. Therefore, the identification of driver genes is one of the major research directions in oncology.

Based on traditional experiments, only hot spot genes, which have high mutation rate in sporadic cancer cases or have clear and obvious family hereditary histories, can be screened as candidates for cancer driver genes (Korenjak and Zavadil, 2019; Sears and Mazzone, 2020). For such pre-identified genes, driver potential needs to be validated using *in vitro* and *in vivo* experiments, which are quite time consuming and expensive (Chu et al., 2018; Sears and Mazzone, 2020), making it impossible and unreasonable for whole-omics wide screening. Therefore, to overcome the restrictions, scholars have introduced computational methods for discovering and pre-selecting candidate driver genes at whole omics level. Nowadays, identifying cancer driver genes using computational methods are not only effective but also reliable. At the beginning, computational methods are applied only for data at one omics level. With the development of algorithms and computational workflows, the integration of two or more omics levels of data to identify core cancer drivers has been realized (Turanli et al., 2018; Olivier et al., 2019). In 2016, Chen et al. (Chen et al., 2016) proposed shortest path-based method to identify novel lung cancer driver genes involving two omics levels. Later, Yuan and Lu employed another network algorithm, random walk with restart (RWR), to conduct the same investigation (Yuan and Lu, 2017). Several possible lung cancer driver genes involving two omics levels were proposed. This study continued the above two previous studies.

Cancer is one of the major threatening diseases for human health in the 21st century. As introduced above, identifying cancer biomarkers is one of the most effective way to explore the pathological mechanisms of tumorigenesis. However, revealing the potential biomarkers of all cancer subjects is difficult due to the limitation of data availability. In this study, we focused on one of the most common and deadly cancer subtype, namely, lung cancer. According to the statistics from GLOBOCAN estimates, in 2020, the general cumulative rates (0–75 years old) for lung cancer have been up to 3.78% in males and 1.77% in females, both of which are in the top of all cancer subtypes (rank 1 in males and rank 2 in females, following breast cancer) (Siegel et al., 2021). In 2021, more than 235,000 of new lung cancer cases and more than 130,000 of deaths from lung cancer are predicted in the United States alone (Siegel et al., 2021), confirming that lung cancer is a growing threat to humans in the 21st century.

Here, we integrated four levels of lung cancer omics data including epigenomics (methylation), genomics (gene variations), transcriptomics (gene expression), and post-transcriptomics regulation (microRNAs). Driver genes identified at each individual omics level were regarded as candidates for lung cancer drivers. Based on the candidates at each level, the Laplacian heat diffusion (LHD) (Carlin et al., 2017) algorithm was executed on a protein–protein interaction (PPI) network to identify raw driver genes. These genes were further filtered by a screen procedure, including permutation, association, and function tests, to exclude false-positive genes and select essential ones. Finally, intersection operation was carried out to identify driver genes involving at least two omics levels. Multiple genes were screened as potential multi-omics lung cancer drivers, and several genes were validated by recent publications via literature mining. The recognition of effective novel driver genes associated with lung cancer provided a novel approach for exploring the cancer pathological mechanisms of lung cancer and identifying clinical biomarkers.

MATERIALS AND METHODS

Datasets

The driver genes at four omics levels, namely, epigenomics, genomics, transcriptomics, and post-transcriptomics regulation, were retrieved from a previous study (Chen et al., 2016). Genes were extracted from DNA methylation, somatic mutation, gene expression, and microRNA expression data, respectively, which were collected in TCGA (https://tcga-data.nci.nih.gov/docs/publications/luad_2014/) (Cancer Genome Atlas Research Network, 2014). The detailed data clean procedures can be found in Ref. (Chen et al., 2016). Finally, we obtained 153 driver genes at epigenomics level, 197 driver genes at genomics level, 1,373 driver genes at transcriptomics level, and 825 driver genes at post-transcriptomics level. Basing on these genes, we aimed to identify driver genes involving two levels.

Network Construction

In this study, we used a powerful network algorithm, LHD algorithm (Carlin et al., 2017), to discover driver genes involving two omics levels. Thus, a network is necessary. PPI information is widely used to tackle protein- and gene-related problems (Ng et al., 2010; Chen et al., 2018; Zhang et al., 2019; Zhang and Chen, 2020; Liu et al., 2021; Pan et al., 2021a; Pan et al., 2021b; Zhu et al., 2021). Here, the human PPI information reported in STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, http://www.string-db.org/, Version 10.0) (Szklarczyk et al., 2014) was adopted to construct a PPI network. To this end, we downloaded the file "9606. protein.links. v10. txt.gz", containing 4,274,001 PPIs that involve 19,247 human proteins. Each PPI consists of two proteins, encoded by Ensembl IDs, and one confidence score with range between 1 and 999. Such score indicates the strength of the PPI. In fact, such score is obtained by considering several aspects of proteins, including close neighborhood in (prokaryotic) genomes, gene fusion, occurrence across species, gene coexpression, literature description, etc. As elaborated in the website of STRING, PPIs

in STRING are derived from genomic context predictions, high-throughput lab experiments, (conserved) co-expression, automated textmining, previous knowledge in databases. Thus, they can widely measure the associations of proteins. This is the great advantage compared with the PPIs reported in other public databases. The constructed PPI network defined 19,247 proteins as nodes, and two nodes were connected by an edge if and only if they can constitute a PPI. Evidently, each edge in the network represented a PPI. To further indicate the difference of PPIs, we assigned each edge with a weight, which was defined as the confidence score of the corresponding PPI. For easy description, such PPI network was denoted by N.

LHD-Based Method

Based on the PPI network N, an LHD-based method was designed to discover driver genes involving two omics levels. The method consisted of two stages. In the first stage, novel driver genes at each of four omics levels were identified by applying the LHD algorithm on the network N and performing a screen procedure. In the second stage, the novel driver genes at each level were refined to discover latent driver genes involving two omics levels.

LHD Algorithm

The LHD algorithm is a network diffusion algorithm (Carlin et al., 2017) that can deliver heat values on seed nodes to others in the network. Given a network N, let A be its adjacent matrix and D be the diagonal matrix, storing the degree of all vertices in the network. The Laplacian matrix L was defined as D-A, i.e., $L = D$-A. In addition, let S be the seed node set. The heat values on seed nodes are stored in a vector, denoted by $H(t_0)$, and its length is equal to the number of nodes in the network N. Evidently, each component corresponds to one node, indicating the heat value of the node. In $H(t_0)$, components corresponding to seed nodes are set to $1/|S|$, and others are set to zero. With the pass of time, heat values on seed nodes are transmitted to other nodes in the following manner:

$$H(t) = H(t_0) \times e^{-Lt}, \tag{1}$$

where t represents the time passed, and $H(t)$ indicates the distribution of heat values at time t. Generally, as the time passes by, $H(t)$ becomes stable. In reality, we tried several values of t and compared two consecutive vectors. If they were close enough, then the LHD algorithm was stopped. The final heat value distribution vector was selected as the outcome of the algorithm. Basing on this vector, we can extract the heat value of each node in the network. Nodes with high heat values were deemed to have strong associations with the seed nodes. By setting a proper threshold, important nodes can be selected.

For driver genes at one omics level, their encoding proteins were first obtained and fed into the LHD algorithm as seed nodes. The LHD algorithm was then performed on the PPI network N. According to the outcomes of the algorithm, nodes (proteins) assigned high heat values were selected as the raw driver genes at this level.

Screen Procedure

Some raw driver genes at each omics level can be obtained by executing the LHD algorithm. However, false-positive genes were inevitably included. A screen procedure was designed to control these genes.

Permutation Test

Heat value is very important to determine the selection of nodes (proteins). However, this value on some nodes (protein) was highly related to the structure of network N. Some nodes (proteins) more easily received high heat values, regardless of which nodes were seed nodes. Thus, the significance of heat value on each raw driver gene selected by the LHD algorithm should be further measured. For raw driver genes at each level, a permutation test was performed. In detail, we randomly generated 500 gene sets, which contained the same number of driver genes at this level. For each generated gene set, the LHD algorithm was executed on N with genes in this set as seed nodes. Finally, each raw driver gene was assigned a heat value. After all the randomly produced sets were considered, each raw driver gene was assigned 500 heat values. These values can be used to measure the significance of actual heat value that was obtained by driver genes at this level. A p-value was computed for each raw driver gene g as follows:

$$p - value(g) = \frac{Heat_{>}}{500}, \tag{2}$$

where $Heat_{>}$ denotes the number of randomly produced gene sets, on which the heat value of g is larger than its actual heat value. If a novel driver gene had a high p-value, then the heat values on several randomly generated gene sets were higher than its actual heat value, indicating that this heat value had no statistical significance. As such, this gene was not special for the driver genes at this level. Thus, we should select novel driver genes with low p-values. Given that 0.05 is always used as the cutoff to measure statistical significance, this study adopted it to filter novel driver genes at each level.

Association Test

By the permutation test, some false positive genes produced by the PPI network were excluded. To further select essential genes among remaining genes, we designed an association test. Several studies have reported that interacting proteins are more likely to share common functions. If the strength of the interaction was considered, then two proteins in a strong interaction were more likely to share common functions than those in a weak interaction. As mentioned in *Network construction*, the confidence score in STRING can measure the strength of an interaction and thus can be used to design the association test. For formulation, the confidence score on the interaction of proteins p_1 and p_2 was denoted by $S(p_1, p_2)$. A measurement, namely maximum association score (MAS), was calculated for each driver gene g by

$$MAS(g) = \max\left\{S\left(g, g'\right)\middle| g' \text{ is a driver gene}\right\} \tag{3}$$

Clearly, the gene with a high MAS was more important and should be kept. We can set a high threshold of MAS to select essential driver genes.

Function Test

To further extract essential driver genes at each level, we designed a third test, namely, function test. This test was based on two types of functional terms: 1) gene ontology and 2) KEGG pathway. In general, driver genes at some level may be annotated by some common functional terms. If the latent one exhibited similar functional terms to some driver genes, then it had a high probability to be a novel driver gene. We first used enrichment theory (Subramanian et al., 2005) to measure the relationship between one gene and all functional terms. In detail, for one gene g and one functional term f, let G be the set consisting of g and its interacting genes and F be the set containing genes annotated by f. The enrichment score between g and f was defined as the $-\log_{10}$ of the hypergeometric test p value of G and F, which can be computed by

$$\text{Enrichment score}(g, f) = -\log_{10}\left(\sum_{k=m}^{n}\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right) \quad (4)$$

where N stands for total number of human genes, M and n represent the number of genes in F and G, respectively, m denotes the number of common genes in F and G. The obtained values were collected in a vector, denoted by $ES(g)$ for gene g. The linkage between two genes was measured according to their vectors by using the following equation:

$$\Psi(g_1, g_2) = \frac{ES(g_1)\cdot ES(g_2)}{\|ES(g_1)\|\cdot\|ES(g_2)\|}, \quad (5)$$

where g_1 and g_2 represent two genes, $ES(g_1)\cdot ES(g_2)$ denote the dot product of two vectors, and $\|ES(g_1)\|$ indicates the module of the vector. Evidently, higher values indicated stronger linkage of two genes. For each latent driver gene g, the maximum function score (MFS) was computed as follows:

$$MFS(g) = \max\left\{\Psi(g, g')\middle| g' \text{ is a driver gene}\right\} \quad (6)$$

Similar to MAS, genes with high MFS values were more likely to be novel driver genes. By setting a proper threshold, essential genes can be obtained.

Intersection Operation

Based on the LHD algorithm and a screen procedure, some essential novel driver genes were obtained at each omics level. By taking the intersection operation, we identified some driver genes involving two levels. As four levels were considered in this study, we finally obtained six driver gene sets. Genes in each set were deemed to be driver genes involving two omics levels.

RESULTS

In this study, an LHD-based method was proposed to identify novel driver genes involving two omics levels. The entire procedures are illustrated in **Figure 1**. The detailed results at each procedure are presented in this section.

Latent Driver Genes at Each Level

In the first stage of LHD-based method, some latent driver genes were identified for each omics level. For epigenomics level, 153 validated driver genes were fed into the LHD algorithm, which was executed on the PPI network N. Each node was assigned a heat value. We selected the nodes with heat value no less than 10^{-5} and obtained 13,101 nodes. The heat values of the selected nodes are provided in **Supplementary Table S1**. A screen procedure was then performed to filter essential candidates. First, a permutation test was adopted to determine the statistical significance of heat value on each selected node, resulting in a p-value for each node (**Supplementary Table S1**). A total of 311 nodes with p-value less than 0.05 were selected. Second, an association test was executed to test the importance of the 311 remaining nodes, assigning an MAS to each node (**Supplementary Table S1**). The threshold of MAS was set as 400, resulting in 228 nodes. Finally, function test was used to evaluate each remaining node, which was assigned with an MFS (**Supplementary Table S1**). The threshold of MFS was set as 0.3. A total of 199 nodes were obtained, and their corresponding genes were selected as the latent driver genes at epigenomics level.

For the three other levels, the same procedures with common thresholds for all measurements were performed. All measurements on each node at the three levels are provided in **Supplementary Tables S2–S4**. The numbers of remaining nodes at each stage of LHD-based method are listed in **Table 1**. As a result, 84, 174, and 39 latent driver genes were accessed at levels of genomics, transcriptomics, and post-transcriptomics regulation, respectively.

For latent driver genes at each omics level, we further investigated their associations with validated genes mentioned in *Datasets*. For each omics level, the PPIs between latent and validated genes were extracted. For each latent gene, we counted three values, which were defined as the number of validated genes that can interact with the latent gene with medium confidence (≥400), high confidence (≥700) and highest confidence (≥900). These values of all latent genes at four omics levels were indicated by four boxplots, as shown in **Figure 2**. It can be observed that each latent gene can interact with at least one validated gene with medium confidence and several latent genes can interact with validated genes with high or highest confidence. This fact indicated that the relationships between latent and validated genes were quite close, increasing the probabilities of latent genes to be actual driver genes.

Driver Genes Involving Two Levels

According to LHD-based method, we conducted the intersection operation of latent driver genes at two levels. Some driver genes

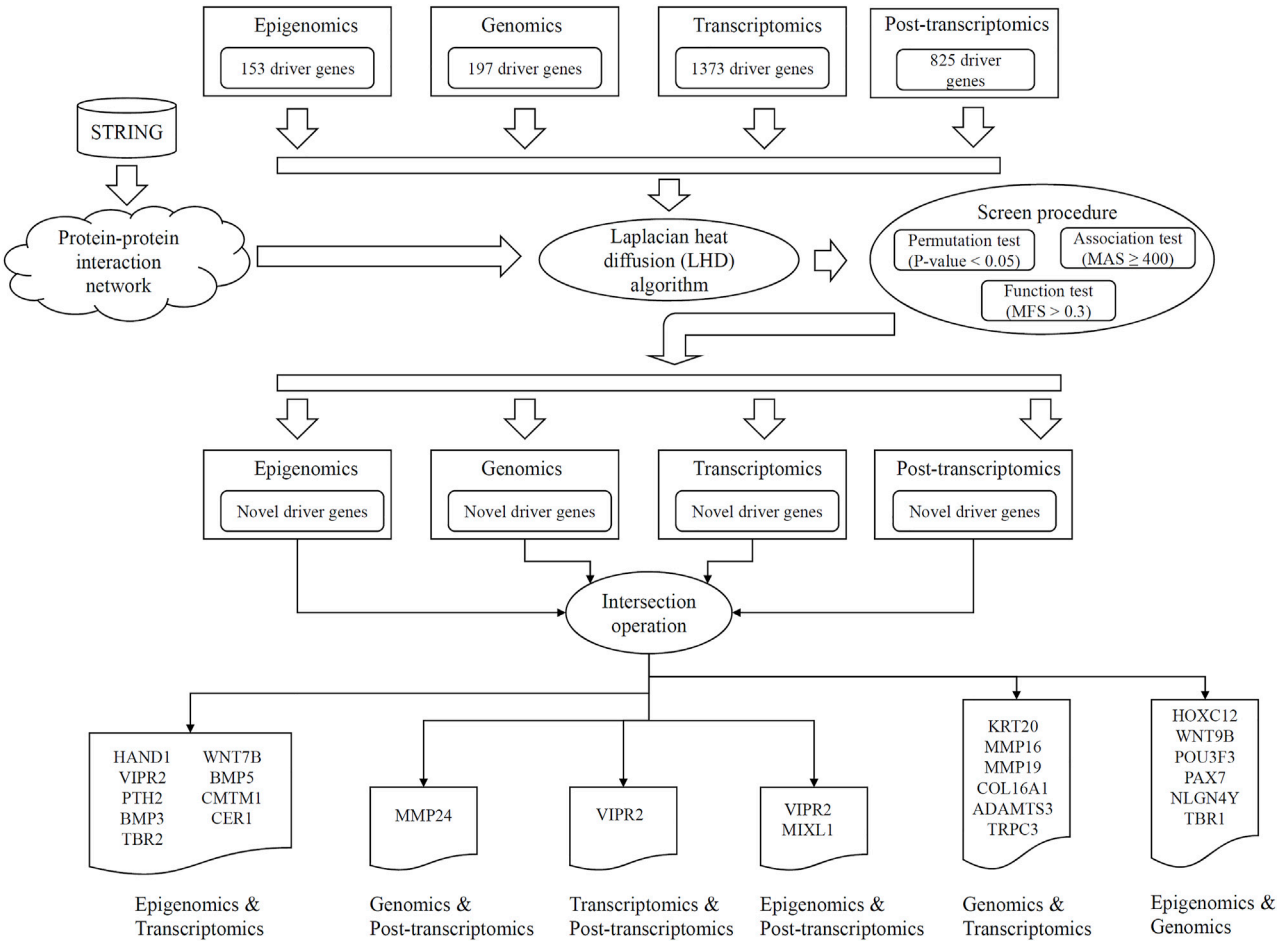


FIGURE 1 | Entire procedures of LHD-based method for identification of multi-omics lung cancer driver genes. Based on driver genes at one omics level, the Laplacian heat diffusion (LHD) algorithm is executed on a protein–protein interaction network reported in STRING to identify raw driver genes. These genes are filtered by a screen procedure. The intersection operation is conducted to identify driver genes involving any two omics levels. Six groups are accessed, each of which contains driver genes involving two omics levels.

TABLE 1 | Remaining latent driver genes at each stage of LHD-based method.

Level	LHD-algorithm	Permutation test	Association test	Function test
Epigenomics	13,101	311	228	199
Genomics	14,321	226	114	84
Transcriptomics	16,256	224	184	174
Post-transcriptomics	17,007	59	44	39

involving two levels were obtained. The number of driver genes involving two levels is listed in **Table 2**, and the detailed genes for any two levels are also listed in **Table 2**. At least one driver gene was accessed for any two levels. *Discussion* presents an extensive discussion on some of the obtained genes.

Comparison With Previous Results

Two previous studies (Chen et al., 2016; Yuan and Lu, 2017) reported some novel driver genes involving two omics levels. The comparison of driver genes indicated that the driver genes reported in the present study were completely different from those in previous two studies. The previous two studies adopted the shortest path-based and RWR-based methods, respectively, to discover novel driver genes, which had quite different principles and procedures; as such, the difference in the driver genes reported between the present study and previous studies was considered reasonable. On the other hand, each method has its limitations, and some driver genes may be omitted. The driver genes reported in this study can be essential supplements for the previous studies.

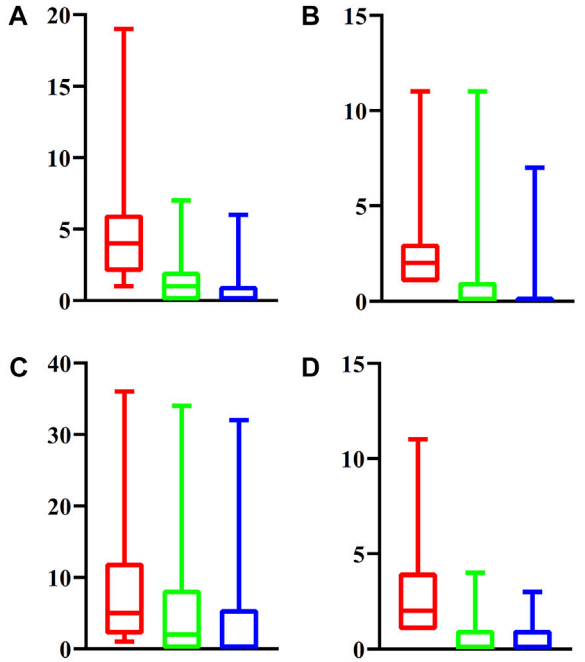


FIGURE 2 | Boxplot to show the associations between latent driver genes and validated ones at each of four omic levels. **(A)** Epigenomics; **(B)** Genomics; **(C)** Transcriptomics; **(D)** Post-transcriptomics. The *Y*-axis represents the number of validated genes that can interact with latent genes with different strength. The red (green, blue, respectively) box denotes number of validated genes that can interact with latent genes with medium (high, highest, respectively) confidence.

DISCUSSION

Novel driver genes involving two of the four omics levels (epigenomics, genomics, transcriptomics, post-transcriptomics levels) were identified using LHD-based method. Six groups of latent multi-omics driver genes for lung cancer were screened. According to recent publications, several identified multi-omics driver genes involving two of the four omics levels can be confirmed to be associated with fundamental lung cancer tumorigenesis-associated pathological mechanisms. The detailed discussion on the genes identified in each group can be seen below.

Shared Genes Between Epigenomics and Post-transcriptomics Regulation

Two genes were identified to be lung cancer drivers at epigenomics (methylation) and post-transcriptomics regulation (microRNA) levels. The first gene is *VIPR2* (ENSP00000262178). Early in 2017, researchers from Sungkyunkwan University confirmed that *VIPR2* is associated with lung cancer at the DNA methylation level (Um et al., 2017). As for the microRNA level, no direct reports confirmed the correlation between lung cancer and *VIPR2*. However, such gene has been shown to be associated with pancreatic cancer tumorigenesis at the microRNA level (Naderi et al., 2014), indicating the specific role of the gene during tumorigenesis. Therefore, this gene could have regulatory effects on lung cancer at the microRNA level. For the next gene involving two omics levels, *MIXL1* (ENSP00000355775), an epigenome analyses in 2019 on circulating tumor cells associated with metastasis revealed that circulating tumor cells in lung cancer has typical epigenomic alterations in *MIXL1* (Gkountela et al., 2019). As for its regulatory effects at post-transcriptomics microRNA level, effective microRNAs from famous microRNA family, let-7 family, has shown to be associated with the dedifferentiation transformation of lung cells during embryonic development or malignant transformation (Navarro and Monzo, 2010). The gene *MIXL1* is regulated by microRNAs from let-7 family during the initiation and proliferation of lung cancer stem cells (Navarro and Monzo, 2010). Therefore, such gene is a potential biomarker presenting abnormal microRNA level regulation during lung tumorigenesis.

Shared Genes Between Epigenomics and Genomics

Six genes were shown to be essential lung cancer drivers involving methylation and genomics levels. The first gene *HOXC12* (ENSP00000243103) has been reported to be triggered by its pathological methylated and inactivated promoter during lung tumorigenesis, validating its specific driver role at the methylation level (Guerrero-Preston et al., 2014). At the genomics level, researchers from Iran in 2019 reported that a SNP in gene *HOXC12* is associated with risk of multiple cancer subtypes, implying the specific lung cancer driver potentials of this gene (Hajjari and Rahnama, 2019). Genes such as *WNT9B* (ENSP00000290015) (Xu et al., 2019), *POU3F3*

TABLE 2 | Driver genes involving two omics levels[a].

Level	Epigenomics	Genomics	Transcriptomics	Post-transcriptomics
Epigenomics	-	6	9	2
Genomics	*HOXC12*, *WNT9B*, *POU3F3*, *PAX7*, *NLGN4Y*, *TBR1*	-	6	1
Transcriptomics	*HAND1*, *VIPR2*, *PTH2*, *BMP3*, *TBR2*, *WNT7B*, *BMP5*, *CMTM1*, *CER1*	*KRT20*, *MMP16*, *MMP19*, *COL16A1*, *ADAMTS3*, *TRPC3*	-	1
Post-transcriptomics	*VIPR2*, *MIXL1*	*MMP24*	*VIPR2*	-

[a]*Numbers in the upper triangle part of this table represent the numbers of driver genes involving two omics levels.*

(ENSP00000355001) (Zeng et al., 2020), and *PAX7* (ENSP00000364524) (Rácz et al., 2000) are lung cancer biomarkers at different omics levels. As for their respective contribution at epigenomics and genomics levels, *WNT9B* (Lan et al., 2006; Farkas et al., 2014), *POU3F3* (Li et al., 2014; Kumar et al., 2016), and *PAX7* (Starzyńska et al., 2020) have all been shown to be associated with lung cancer at epigenomics and genomics levels independently. For the two remaining genes, *NLGN4Y* (ENSP00000342535) and *TBR1* (ENSP00000374205), in 2019, researchers from University of Southampton summarized *NLGN4Y* as a multi-omics level driver for lung cancer at least at epigenomics and transcriptomics levels (Jeyananthan and Niranjan, 2019). As for the genomics level, variants in *NLGN4Y* can regulate cell proliferation in multiple pathogenesis, though not directly reported in lung cancer (Nardello et al., 2021). The potential regulatory effects of such gene on cell proliferation indicated that it may also be associated with lung cancer at the genomic level. Similar experimental works support the driver role of the *TBR1* gene at genomics and epigenomics levels (Ischenko et al., 2014; Serth et al., 2020). Therefore, all predicted genes at epigenomic (methylation) and genomics levels have been validated to be potential lung cancer drivers.

Shared Genes Between Epigenomics and Transcriptomics

Nine genes have been shown to be associated with lung cancer and are potential lung cancer drivers at epigenomics and transcriptomics levels. Among the nine genes, *VIPR2* (ENSP00000262178) has already been discussed above and shown to be regulated at epigenomics and post-transcriptomics level. According to the same supporting paper mentioned above, this gene can be regulated at the transcriptomics level (Um et al., 2017). Other genes, such as *TBR2* (ENSP00000295743) and *WNT7B* (ENSP00000341032), which are the homologues of *TBR1* and *WNT9B*, respectively, are also reasonable to be speculated as candidate regulators at epigenomics and transcriptomics levels. Considering the limitation of the manuscript's length, we selected three candidates for detailed discussion: *HAND1*, *BMP3*, and *CMTM1*. According to recent publications, *HAND1* is a typical DNA methylation biomarker of small-cell lung cancers (Kalari et al., 2013). At the transcriptomics level, *HAND1* has been identified as a biomarker in a newly reported single-cell study (Yin et al., 2019). Apart from *HAND1*, *BMP3* together with *BMP5* has been considered potential multi-omics level lung cancer biomarkers according to recent publications. In 2015, researchers from Huazhong University of Science and Technology conducted transcriptomics level analyses on human lung squamous cell carcinoma, confirming the driver role of *BMP3* and *BMP5* at the transcriptomics levels (Deng et al., 2015). At the epigenomics level (methylation), an earlier study on colorectal cancer confirmed that the methylation alteration on *BMP3* or *BMP5* may trigger the malignant transformation of normal cells (Loh et al., 2008). Therefore, the two genes could be potential lung cancer biomarkers at multi-omics levels. The *CMTM1* gene has also been reported by two independent studies to be associated with lung cancer at transcriptomics (Hou et al., 2020) and methylation (Shao et al., 2007) levels, respectively. Other genes, such as *PTH2* (Kim et al., 1998), *TBR2* (Kaowinn et al., 2017), *WNT7B* (Kirikoshi and Katoh, 2002), and *CER1* (Semenova et al., 2016), either have similar effects shared with their homologues or have been independently reported to be associated with lung cancer at different omics levels.

Shared Genes Between Genomics and Post-transcriptomics Regulation

Only one gene has been predicted to regulate lung cancer-associated pathological effects at genomics and post-transcriptomics levels. According to recent publications, *MMP24* (ENSP00000246186) is associated with lung cancer at different omics levels (Fontenele et al., 2015; Wang et al., 2019; Wang et al., 2020). At genomics and post-transcriptomics levels, variants on *MMP24* have been shown to be associated with lung cancer via a GWAS study in 2015 (Fontenele et al., 2015). In 2019, researchers from Tumor Hospital of Wuwei confirmed that microRNA-133a regulates *MMP24* and further contributes to the tumorigenesis of lung cancer (Wang et al., 2019).

Shared Genes Between Transcriptomics and Post-transcriptomics Regulation

For genes that have been identified to be potential lung cancer drivers at transcriptomics and post-transcriptomics levels, only one gene, namely, *VIPR2* (ENSP00000262178), was identified as multi-omics level regulator. As discussed above, this gene has shown to be effective at epigenomics, transcriptomics, and post-transcriptomics. Therefore, this gene could be a multi-omics regulator.

Shared Genes Between Genomics and Transcriptomics

Six genes have been identified to regulate lung tumorigenesis via genomics and transcriptomics levels. The first gene *KRT20* (ENSP00000167588) has been shown to define the invasive characteristics of cancer cells at the transcriptomics level in multiple cancer subtypes (Eckstein et al., 2018), including lung cancer (Mollaoglu et al., 2018; Maly et al., 2019). As for the genomics level, variants in *KRT20* are associated with lung cancer (Huang et al., 2015). Genes *MMP16* (ENSP00000286614) and *MMP19* (ENSP00000313437) belong to the matrix metalloproteinase family that is associated with multiple cancer subtypes, including lung cancer (Rudolph-Owen et al., 1998). As for their effects on genomics and transcriptomics levels, in the same publication, a summary of the effects of matrix metalloproteinases at multi-omics have been presented and demonstrated. *TRPC3* (ENSP00000368966), as another novel driver gene associated with lung cancer at genomics and transcriptomics levels, has been validated to be potential genomic and transcriptomic driver for lung cancer. In 2016, a researcher from Guangzhou Medical University validated the

effects of *TRPC3* variants on lung cancer risk (Zhang et al., 2016). In 2021, such gene has been reported to be associated with malignant transformation at the transcriptomics level (Lin et al., 2021). The two remaining genes are *COL16A1* (ENSP00000362776) and *ADAMTS3* (ENSP00000286657). In 2016, *COL16A1* has been shown to be associated with lung tumorigenesis via specific mutant patterns validated by *in vitro* A549 cell lines (Wang et al., 2016). As for the transcriptomics level, researchers from Medical University of South Carolina validated that at least in oral squamous cell carcinoma, the alteration of the gene expression of *COL16A1* may promote tumor growth via interacting with the RNA-binding protein CELF1 (House et al., 2015). As for the last predicted gene *ADAMTS3*, according to recent publications, this gene has been confirmed to participate in lung tumorigenesis at genomics and transcriptomics levels (Vanni et al., 2016; Hannen et al., 2019).

Overall, several identified multi-omics lung cancer drivers were validated to be associated with fundamental pathological mechanisms. Therefore, LHD-based method was effective and accurate to identify lung cancer-associated tumor drivers and may help reveal the potential mechanisms of lung tumorigenesis.

Clinical Applications on New Driver Genes

Based on recent publications, some identified multi-omics lung cancer drivers have been already applied for clinical use on the diagnosis or treatment against lung cancer. According to recent publications, *WNT9B*, as the multi-omics biomarker for lung cancer at both epigenomics and genomics levels has been shown to act as an effective drug target (Stewart et al., 2014). The application of such drug can significantly improve the survival probability for lung cancer (Stewart et al., 2014). The next widely reported gene is *MMP19*, as a new driver gene at both genomics and transcriptomics levels. It has been widely reported to be associated with tumorigenesis, including gastric cancer (Shen et al., 2020), glioma (Luo et al., 2018) and lung cancer (Yu et al., 2014). As for its clinical application for lung cancer diagnosis and treatment, *MMP19* has been reported to promote the metastasis of lung cancer and is associated with increased mortality of lung cancer as an active biomarker (Yu et al., 2014). *TRPC3*, as a driver gene at both genomics and transcriptomics levels have also been recognized as a promising clinical biomarker for lung cancer (Lastraioli et al., 2015), indicating the clinical significance of such multi-omics biomarker.

CONCLUSION

This study investigated lung cancer driver genes involving two omics levels. An LHD-based method was proposed to identify multi-omics lung cancer driver genes. Several genes, such as *HOXC12*, *HAND1*, *VIPR2*, *KRT20*, *MMP24*, and *VIPR2*, were discovered, and their special roles at different omics levels of lung cancer were confirmed. These findings may help improve the research progress on lung cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://tcga-data.nci.nih.gov/docs/publications/luad_2014/.

AUTHOR CONTRIBUTIONS

TH, ZL, and Y-DC designed the study. FY, XC, and LC performed the experiments. FY, XC, and Y-HZ analyzed the results. FY and Y-HZ wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA26040304, XDB38050200), National Key R and D Program of China (2018YFC0910403), the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002), the science and technology project of Binzhou Medical University (BY2016KYQD22), the medicine and health science technology development program of Shandong Province (2018WS541).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.825272/full#supplementary-material

REFERENCES

Cancer Genome Atlas Research Network (2014). Comprehensive Molecular Profiling of Lung Adenocarcinoma. *Nature* 511, 543–550. doi:10.1038/nature13385

Carlin, D. E., Demchak, B., Pratt, D., Sage, E., and Ideker, T. (2017). Network Propagation in the Cytoscape Cyberinfrastructure. *Plos Comput. Biol.* 13, e1005598. doi:10.1371/journal.pcbi.1005598

Chen, L., Huang, T., Zhang, Y.-H., Jiang, Y., Zheng, M., and Cai, Y.-D. (2016). Identification of Novel Candidate Drivers Connecting Different Dysfunctional Levels for Lung Adenocarcinoma Using Protein-Protein Interactions and a Shortest Path Approach. *Sci. Rep.* 6, 29849. doi:10.1038/srep29849

Chen, L., Zhang, Y.-H., Zhang, Z., Huang, T., and Cai, Y.-D. (2018). Inferring Novel Tumor Suppressor Genes with a Protein-Protein Interaction Network and Network Diffusion Algorithms. *Mol. Ther. Methods Clin. Dev.* 10, 57–67. doi:10.1016/j.omtm.2018.06.007

Chu, G. C. W., Lazare, K., and Sullivan, F. (2018). Serum and Blood Based Biomarkers for Lung Cancer Screening: a Systematic Review. *BMC Cancer* 18, 181. doi:10.1186/s12885-018-4024-3

Deng, T., Lin, D., Zhang, M., Zhao, Q., Li, W., Zhong, B., et al. (2015). Differential Expression of Bone Morphogenetic Protein 5 in Human Lung Squamous Cell Carcinoma and Adenocarcinoma. *Acta Biochim. Biophys. Sin.* 47, 557–563. doi:10.1093/abbs/gmv037

Eckstein, M., Wirtz, R., Gross-Weege, M., Breyer, J., Otto, W., Stoehr, R., et al. (2018). mRNA-Expression of KRT5 and KRT20 Defines Distinct Prognostic

Subgroups of Muscle-Invasive Urothelial Bladder Cancer Correlating with Histological Variants. *Int. J. Mol. Sci.* 19, 3396. doi:10.3390/ijms19113396

Farkas, S. A., Vodickova, L., Vodicka, P., Nilsson, T. K., and Vymetalkova, V. (2014). DNA Methylation Changes in Genes Frequently Mutated in Sporadic Colorectal Cancer and in the DNA Repair and Wnt/β-Catenin Signaling Pathway Genes. *Epigenomics* 6, 179–191. doi:10.2217/epi.14.7

Fontenele, E. G. P., Amaral de Moraes, M. E., Brasil d'Alva, C., Pinheiro, D. P., Aguiar Sales Pinheiro Landim, S., de Sousa Barros, F. A., et al. (2015). Association Study of GWAS-Derived Loci with Height in Brazilian Children: Importance of MAP3K3, MMP24 and IGF1R Polymorphisms for Height Variation. *Horm. Res. Paediatr.* 84, 248–253. doi:10.1159/000437324

Gkountela, S., Castro-Giner, F., Szczerba, B. M., Vetter, M., Landin, J., Scherrer, R., et al. (2019). Circulating Tumor Cell Clustering Shapes DNA Methylation to Enable Metastasis Seeding. *Cell* 176, 98–112. doi:10.1016/j.cell.2018.11.046

Guerrero-Preston, R., Michailidi, C., Marchionni, L., Pickering, C. R., Frederick, M. J., Myers, J. N., et al. (2014). Key Tumor Suppressor Genes Inactivated by "greater Promoter" Methylation and Somatic Mutations in Head and Neck Cancer. *Epigenetics* 9, 1031–1046. doi:10.4161/epi.29025

Hajjari, M., and Rahnama, S. (2019). Association between SNPs of Long Non-coding RNA HOTAIR and Risk of Different Cancers. *Front. Genet.* 10, 113. doi:10.3389/fgene.2019.00113

Hannen, R., Selmansberger, M., Hauswald, M., Pagenstecher, A., Nist, A., Stiewe, T., et al. (2019). Comparative Transcriptomic Analysis of Temozolomide Resistant Primary GBM Stem-like Cells and Recurrent GBM Identifies Up-Regulation of the Carbonic Anhydrase CA2 Gene as Resistance Factor. *Cancers* 11, 921. doi:10.3390/cancers11070921

Hou, X., He, S., Zhang, D., Yang, C., Shi, Y., and Zhang, K. (2020). Expression and Clinical Significance of CMTM6 in Nonsmall Cell Lung Cancer. *DNA Cel Biol.* 39, 2265–2271. doi:10.1089/dna.2020.5564

House, R. P., Talwar, S., Hazard, E. S., Hill, E. G., and Palanisamy, V. (2015). RNA-binding Protein CELF1 Promotes Tumor Growth and Alters Gene Expression in Oral Squamous Cell Carcinoma. *Oncotarget* 6, 43620–43634. doi:10.18632/oncotarget.6204

Huang, T., Yang, J., and Cai, Y. D. (2015). Novel Candidate Key Drivers in the Integrative Network of Genes, microRNAs, Methylations, and Copy Number Variations in Squamous Cell Lung Carcinoma. *Biomed. Res. Int.* 2015, 358125. doi:10.1155/2015/358125

Ischenko, I., Liu, J., Petrenko, O., and Hayman, M. J. (2014). Transforming Growth Factor-Beta Signaling Network Regulates Plasticity and Lineage Commitment of Lung Cancer Cells. *Cell Death Differ* 21, 1218–1228. doi:10.1038/cdd.2014.38

Jeyananthan, P., and Niranjan, M. (2019). "Classification and Regression Analysis of Lung Tumors from Multi-Level Gene Expression Data," in International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14-19 July 2019 (IEEE), 1–8. doi:10.1109/IJCNN.2019.8852282

Kalari, S., Jung, M., Kernstine, K. H., Takahashi, T., and Pfeifer, G. P. (2013). The DNA Methylation Landscape of Small Cell Lung Cancer Suggests a Differentiation Defect of Neuroendocrine Cells. *Oncogene* 32, 3559–3568. doi:10.1038/onc.2012.362

Kaowinn, S., Kim, J., Lee, J., Shin, D. H., Kang, C.-D., Kim, D.-K., et al. (2017). Cancer Upregulated Gene 2 Induces Epithelial-Mesenchymal Transition of Human Lung Cancer Cells via TGF-β Signaling. *Oncotarget* 8, 5092–5110. doi:10.18632/oncotarget.13867

Kim, S. K., Su, L.-K., Oh, Y., Kemp, B. L., Hong, W. K., and Mao, L. (1998). Alterations of PTEN/MMAC1, a Candidate Tumor Suppressor Gene, and its Homologue, PTH2, in Small Cell Lung Cancer Cell Lines. *Oncogene* 16, 89–93. doi:10.1038/sj.onc.1201512

Kirikoshi, H., and Katoh, M. (2002). Expression of WNT7A in Human normal Tissues and Cancer, and Regulation of WNT7A and WNT7B in Human Cancer. *Int. J. Oncol.* 21, 895–900. doi:10.3892/ijo.21.4.895

Korenjak, M., and Zavadil, J. (2019). Experimental Identification of Cancer Driver Alterations in the Era of pan-cancer Genomics. *Cancer Sci.* 110, 3622–3629. doi:10.1111/cas.14210

Kumar, S., Rathkolb, B., Kemter, E., Sabrautzki, S., Michel, D., Adler, T., et al. (2016). Generation and Standardized, Systemic Phenotypic Analysis of Pou3f3L423P Mutant Mice. *PLoS One* 11, e0150472. doi:10.1371/journal.pone.0150472

Lan, Y., Ryan, R. C., Zhang, Z., Bullard, S. A., Bush, J. O., Maltby, K. M., et al. (2006). Expression ofWnt9b and Activation of Canonical Wnt Signaling during Midfacial Morphogenesis in Mice. *Dev. Dyn.* 235, 1448–1454. doi:10.1002/dvdy.20723

Lastraioli, E., Iorio, J., and Arcangeli, A. (2015). Ion Channel Expression as Promising Cancer Biomarker. *Biochim. Biophys. Acta Biomembr.* 1848, 2685–2702. doi:10.1016/j.bbamem.2014.12.016

Li, W., Zheng, J., Deng, J., You, Y., Wu, H., Li, N., et al. (2014). Increased Levels of the Long Intergenic Non-protein Coding RNA POU3F3 Promote DNA Methylation in Esophageal Squamous Cell Carcinoma Cells. *Gastroenterology* 146, 1714–1726. doi:10.1053/j.gastro.2014.03.002

Li, Y., Gu, J., Xu, F., Zhu, Q., Ge, D., and Lu, C. (2018). Transcriptomic and Functional Network Features of Lung Squamous Cell Carcinoma through Integrative Analysis of GEO and TCGA Data. *Sci. Rep.* 8, 15834. doi:10.1038/s41598-018-34160-w

Lin, D.-C., Zheng, S.-Y., Zhang, Z.-G., Luo, J.-H., Zhu, Z.-L., Li, L., et al. (2021). TRPC3 Promotes Tumorigenesis of Gastric Cancer via the CNB2/GSK3β/NFATc2 Signaling Pathway. *Cancer Lett.* 519, 211–225. doi:10.1016/j.canlet.2021.07.038

Liu, H., Hu, B., Chen, L., and Lu, L. (2021). Identifying Protein Subcellular Location with Embedding Features Learned from Networks. *CP* 18, 646–660. doi:10.2174/1570164617999201124142950

Loh, K., Chia, J. A., Greco, S., Cozzi, S.-J., Buttenshaw, R. L., Bond, C. E., et al. (2008). Bone Morphogenic Protein 3 Inactivation Is an Early and Frequent Event in Colorectal Cancer Development. *Genes Chromosom. Cancer* 47, 449–460. doi:10.1002/gcc.20552

Luo, Q., Luo, H., Chen, X., Yan, P., Fu, H., Huang, H., et al. (2018). The Expression of MMP19 and its Clinical Significance in Glioma. *Int. J. Clin. Exp. Pathol.* 11, 5407–5412.

Maly, V., Maly, O., Kolostova, K., and Bobek, V. (2019). Circulating Tumor Cells in Diagnosis and Treatment of Lung Cancer. *In Vivo* 33, 1027–1037. doi:10.21873/invivo.11571

Mollaoglu, G., Jones, A., Wait, S. J., Mukhopadhyay, A., Jeong, S., Arya, R., et al. (2018). The Lineage-Defining Transcription Factors SOX2 and NKX2-1 Determine Lung Cancer Cell Fate and Shape the Tumor Immune Microenvironment. *Immunity* 49, 764–779. doi:10.1016/j.immuni.2018.09.020

Naderi, E., Mostafaei, M., Pourshams, A., and Mohamadkhani, A. (2014). Network of microRNAs-mRNAs Interactions in Pancreatic Cancer. *Biomed. Res. Int.* 2014, 534821. doi:10.1155/2014/534821

Nardello, R., Antona, V., Mangano, G. D., Salpietro, V., Mangano, S., and Fontana, A. (2021). A Paradigmatic Autistic Phenotype Associated with Loss of PCDH11Y and NLGN4Y Genes. *BMC Med. Genomics* 14, 98. doi:10.1186/s12920-021-00934-x

Navarro, A., and Monzó, M. (2010). MicroRNAs in Human Embryonic and Cancer Stem Cells. *Yonsei Med. J.* 51, 622–632. doi:10.3349/ymj.2010.51.5.622

Ng, K.-L., Ciou, J.-S., and Huang, C.-H. (2010). Prediction of Protein Functions Based on Function-Function Correlation Relations. *Comput. Biol. Med.* 40, 300–305. doi:10.1016/j.compbiomed.2010.01.001

Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D., and Cox, L. A. (2019). The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *Int. J. Mol. Sci.* 20, 4781. doi:10.3390/ijms20194781

Pan, X., Chen, L., Liu, M., Niu, Z., Huang, T., and Cai, Y. D. (2021a). Identifying Protein Subcellular Locations with Embeddings-Based Node2loc. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi:10.1109/tcbb.2021.3080386

Pan, X., Li, H., Zeng, T., Li, Z., Chen, L., Huang, T., et al. (2021b). Identification of Protein Subcellular Localization with Network and Functional Embeddings. *Front. Genet.* 11, 626500. doi:10.3389/fgene.2020.626500

Pao, W., and Girard, N. (2011). New Driver Mutations in Non-small-cell Lung Cancer. *Lancet Oncol.* 12, 175–180. doi:10.1016/s1470-2045(10)70087-5

Pon, J. R., and Marra, M. A. (2015). Driver and Passenger Mutations in Cancer. *Annu. Rev. Pathol. Mech. Dis.* 10, 25–50. doi:10.1146/annurev-pathol-012414-040312

Rácz, A., Brass, N., Höfer, M., Sybrecht, G. W., Remberger, K., and Meese, E. U. (2000). Gene Amplification at Chromosome 1pter-P33 Including the Genes PAX7 and ENO1 in Squamous Cell Lung Carcinoma. *Int. J. Oncol.* 17, 67–73. doi:10.3892/ijo.17.1.67

Rudolph-Owen, L. A., Chan, R., Muller, W. J., and Matrisian, L. M. (1998). The Matrix Metalloproteinase Matrilysin Influences Early-Stage Mammary Tumorigenesis. *Cancer Res.* 58, 5500–5506.

Sears, C. R., and Mazzone, P. J. (2020). Biomarkers in Lung Cancer. *Clin. Chest Med.* 41, 115–127. doi:10.1016/j.ccm.2019.10.004

Semenova, E. A., Kwon, M.-c., Monkhorst, K., Song, J.-Y., Bhaskaran, R., Krijgsman, O., et al. (2016). Transcription Factor NFIB Is a Driver of Small Cell Lung Cancer Progression in Mice and Marks Metastatic Disease in Patients. *Cel Rep.* 16, 631–643. doi:10.1016/j.celrep.2016.06.020

Serth, J., Peters, I., Dubrowinskaja, N., Reese, C., Albrecht, K., Klintschar, M., et al. (2020). Age-, Tumor-, and Metastatic Tissue-Associated DNA Hypermethylation of a T-Box Brain 1 Locus in Human Kidney Tissue. *Clin. Epigenet* 12, 33. doi:10.1186/s13148-020-0823-x

Shao, L., Cui, Y., Li, H., Liu, Y., Zhao, H., Wang, Y., et al. (2007). CMTM5 Exhibits Tumor Suppressor Activities and Is Frequently Silenced by Methylation in Carcinoma Cell Lines. *Clin. Cancer Res.* 13, 5756–5762. doi:10.1158/1078-0432.ccr-06-3082

Shen, D., Zhao, H., Zeng, P., Song, J., Yang, Y., Gu, X., et al. (2020). Circular RNA Hsa_circ_0005556 Accelerates Gastric Cancer Progression by Sponging miR-4270 to Increase MMP19 Expression. *J. Gastric Cancer* 20, 300. doi:10.5230/jgc.2020.20.e28

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer Statistics, 2021. *CA Cancer J. Clin.* 71, 7–33. doi:10.3322/caac.21654

Starzyńska, T., Karczmarski, J., Paziewska, A., Kulecka, M., Kuśnierz, K., Żeber-Lubecka, N., et al. (2020). Differences between Well-Differentiated Neuroendocrine Tumors and Ductal Adenocarcinomas of the Pancreas Assessed by Multi-Omics Profiling. *Int. J. Mol. Sci.* 21, 4470. doi:10.3390/ijms21124470

Stewart, D. J., Chang, D. W., Ye, Y., Spitz, M., Lu, C., Shu, X., et al. (2014). Wnt Signaling Pathway Pharmacogenetics in Non-small Cell Lung Cancer. *Pharmacogenomics J.* 14, 509–522. doi:10.1038/tpj.2014.21

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* 43, D447–D452. doi:10.1093/nar/gku1003

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., and Karchin, R. (2016). Evaluating the Evaluation of Cancer Driver Genes. *Proc. Natl. Acad. Sci. USA* 113, 14330–14335. doi:10.1073/pnas.1616440113

Turanli, B., Karagoz, K., Gulfidan, G., Sinha, R., Mardinoglu, A., and Arga, K. Y. (2018). A Network-Based Cancer Drug Discovery: From Integrated Multi-Omics Approaches to Precision Medicine. *Curr. Pharm. Des.* 24, 3778–3790. doi:10.2174/1381612824666181106095959

Um, S.-W., Kim, H. K., Kim, Y., Lee, B. B., Kim, D., Han, J., et al. (2017). Bronchial Biopsy Specimen as a Surrogate for DNA Methylation Analysis in Inoperable Lung Cancer. *Clin. Epigenet* 9, 131. doi:10.1186/s13148-017-0432-5

Vanni, I., Coco, S., Bonfiglio, S., Cittaro, D., Genova, C., Biello, F., et al. (2016). Whole Exome Sequencing of Independent Lung Adenocarcinoma, Lung Squamous Cell Carcinoma, and Malignant Peritoneal Mesothelioma. *Medicine (Baltimore)* 95, e5447. doi:10.1097/md.0000000000005447

Wang, Y., Yan, P., Liu, Z., Yang, X., Wang, Y., Shen, Z., et al. (2016). MEK Inhibitor Can Reverse the Resistance to Bevacizumab in A 549 Cells Harboring Kirsten Rat Sarcoma Oncogene Homolog Mutation. *Thorac. Cancer* 7, 279–287. doi:10.1111/1759-7714.12325

Wang, H., Zhang, Y., Zhang, Y., Liu, W., and Wang, J. (2019). Cryptotanshinone Inhibits Lung Cancer Invasion via microRNA-133a/matrix Metalloproteinase 14 Regulation. *Oncol. Lett.* 18, 2554–2559. doi:10.3892/ol.2019.10580

Wang, Y., Ding, X., Liu, B., Li, M., Chang, Y., Shen, H., et al. (2020). ETV4 Overexpression Promotes Progression of Non-small Cell Lung Cancer by Upregulating PXN and MMP1 Transcriptionally. *Mol. Carcinog* 59, 73–86. doi:10.1002/mc.23130

Xu, H., Jiao, X., Wu, Y., Li, S., Cao, L., and Dong, L. (2019). Exosomes Derived from PM2.5-treated Lung Cancer Cells Promote the Growth of Lung Cancer via the Wnt3a/β Catenin Pathway. *Oncol. Rep.* 41, 1180–1188. doi:10.3892/or.2018.6862

Yin, N., Liang, X., Liang, S., Liang, S., Yang, R., Hu, B., et al. (2019). Embryonic Stem Cell- and Transcriptomics-Based *In Vitro* Analyses Reveal that Bisphenols A, F and S Have Similar and Very Complex Potential Developmental Toxicities. *Ecotoxicol. Environ. Saf.* 176, 330–338. doi:10.1016/j.ecoenv.2019.03.115

Yu, G., Herazo-Maya, J. D., Nukui, T., Romkes, M., Parwani, A., Juan-Guardela, B. M., et al. (2014). Matrix Metalloproteinase-19 Promotes Metastatic BehaviorIn Vitroand Is Associated with Increased Mortality in Non-small Cell Lung Cancer. *Am. J. Respir. Crit. Care Med.* 190, 780–790. doi:10.1164/rccm.201310-1903oc

Yuan, F., and Lu, W. (2017). Prediction of Potential Drivers Connecting Different Dysfunctional Levels in Lung Adenocarcinoma via a Protein-Protein Interaction Network. *Biochim. Biophys. Acta Mol. Basis Dis.* 1864, 2284–2293. doi:10.1016/j.bbadis.2017.11.018

Zeng, Q., Dai, Y., Duan, C., Zeng, R., Zeng, Q., and Wei, C. (2020). Long Noncoding RNA POU3F3 Enhances Cancer Cell Proliferation, Migration and Invasion in Non-small Cell Lung Cancer (Adenocarcinoma) by Downregulating microRNA-30d-5p. *BMC Pulm. Med.* 20, 185. doi:10.1186/s12890-020-01218-3

Zhang, X., and Chen, L. (2020). Prediction of Membrane Protein Types by Fusing Protein-Protein Interaction and Protein Sequence Information. *Biochim. Biophys. Acta Proteins Proteomics* 1868, 140524. doi:10.1016/j.bbapap.2020.140524

Zhang, Z., Wang, J., He, J., Zeng, X., Chen, X., Xiong, M., et al. (2016). Identification of TRPCs Genetic Variants that Modify Risk for Lung Cancer Based on the Pathway and Two-Stage Study. *Meta Gene* 9, 191–196. doi:10.1016/j.mgene.2016.07.005

Zhang, X., Chen, L., Guo, Z.-H., and Liang, H. (2019). Identification of Human Membrane Protein Types by Incorporating Network Embedding Methods. *IEEE Access* 7, 140794–140805. doi:10.1109/access.2019.2944177

Zhu, Y., Hu, B., Chen, L., and Dai, Q. (2021). iMPTCE-Hnetwork: A Multilabel Classifier for Identifying Metabolic Pathway Types of Chemicals and Enzymes with a Heterogeneous Network. *Comput. Math. Methods Med.* 2021, 6683051. doi:10.1155/2021/6683051

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Yuan, Cao, Zhang, Chen, Huang, Li and Cai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 26 January 2022
doi: 10.3389/fcell.2021.720321

Interpreting Functional Impact of Genetic Variations by Network QTL for Genotype–Phenotype Association Study

Kai Yuan[1], *Tao Zeng*[1,2,3]* *and Luonan Chen*[1,4,5,6]*

[1]Key Laboratory of Systems Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai, China, [2]Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [3]Guangzhou Laboratory, Guangzhou, China, [4]Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou, China, [5]School of Life Science and Technology, ShanghaiTech University, Shanghai, China, [6]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Jianxin Wang,
Central South University, China
Bingqiang Liu,
Shandong University, China

***Correspondence:**
Tao Zeng
zengtao@sibs.ac.cn
Luonan Chen
lnchen@sibs.ac.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *04 June 2021*
Accepted: *13 December 2021*
Published: *26 January 2022*

Citation:
Yuan K, Zeng T and Chen L (2022) Interpreting Functional Impact of Genetic Variations by Network QTL for Genotype–Phenotype Association Study. Front. Cell Dev. Biol. 9:720321. doi: 10.3389/fcell.2021.720321

An enormous challenge in the post-genome era is to annotate and resolve the consequences of genetic variation on diverse phenotypes. The genome-wide association study (GWAS) is a well-known method to identify potential genetic loci for complex traits from huge genetic variations, following which it is crucial to identify expression quantitative trait loci (eQTL). However, the conventional eQTL methods usually disregard the systematical role of single-nucleotide polymorphisms (SNPs) or genes, thereby overlooking many network-associated phenotypic determinates. Such a problem motivates us to recognize the network-based quantitative trait loci (QTL), i.e., network QTL (nQTL), which is to detect the cascade association as genotype → network → phenotype rather than conventional genotype → expression → phenotype in eQTL. Specifically, we develop the nQTL framework on the theory and approach of single-sample networks, which can identify not only network traits (e.g., the gene subnetwork associated with genotype) for analyzing complex biological processes but also network signatures (e.g., the interactive gene biomarker candidates screened from network traits) for characterizing targeted phenotype and corresponding subtypes. Our results show that the nQTL framework can efficiently capture associations between SNPs and network traits (i.e., edge traits) in various simulated data scenarios, compared with traditional eQTL methods. Furthermore, we have carried out nQTL analysis on diverse biological and biomedical datasets. Our analysis is effective in detecting network traits for various biological problems and can discover many network signatures for discriminating phenotypes, which can help interpret the influence of nQTL on disease subtyping, disease prognosis, drug response, and pathogen factor association. Particularly, in contrast to the conventional approaches, the nQTL

Abbreviations: CCLE, Cancer Cell Line Encyclopedia; eQTL, expression quantitative trait loci; GWAS, genome-wide association study; nQTL, network quantitative trait loci; NSCLC, non-small-cell lung cancer; PBMC, peripheral blood mononuclear cell; PLSDA, partial least squares discriminant analysis; PPI, protein–protein interaction; SCLC, small-cell lung cancer; SNP, single-nucleotide polymorphisms; SSN, single-sample network; TF, transcriptional factor.

framework could also identify many network traits from human bulk expression data, validated by matched single-cell RNA-seq data in an independent or unsupervised manner. All these results strongly support that nQTL and its detection framework can simultaneously explore the global genotype–network–phenotype associations and the underlying network traits or network signatures with functional impact and importance.

Keywords: genetic variation, expression quantitative trait loci, network quantitative trait loci, data integration, single-sample network, single cell, network trait, network signature

INTRODUCTION

An enormous challenge in the post-genome era is to annotate and resolve the consequences of diverse genetic variations (Lynch and Hsiao, 2019; Strober et al., 2019; Young et al., 2019), particularly within the context of human diseases (Gibbs et al., 2010; Liang et al., 2013; Zeggini et al., 2019). The genome-wide association study (GWAS) is a well-known method to identify potential genetic loci for complex diseases from a large number of genetic variations (Gamazon et al., 2010); however, the target genes and underlying mechanisms remain massively unknown due to absent functional information for the vast majority of loci (Zhu et al., 2016; Mason et al., 2018). As a crucial mechanism of genetic variants that affect gene expression (Kang et al., 2012; Peters et al., 2016; Strunz et al., 2018), expression quantitative trait loci (eQTL) indicate genomic loci that contribute to variations in gene expression levels, which reveals the connection between single-nucleotide polymorphisms (SNPs) and genes on functions rather than on sequences (Li et al., 2016), supplying detailed functional explanations of GWAS outcomes (Michaelson et al., 2009; Holloway et al., 2011; Peterson et al., 2016; Joehanes et al., 2017; Son et al., 2017; Guo et al., 2018).

To date, the eQTL method can be carried out across different genetic populations (Chen et al., 2020) in a cell-type-specific manner (Westra et al., 2015; Kasela et al., 2017; Vuckovic et al., 2020), which can generate specific molecular hypotheses (Wu et al., 2008; Thibodeau et al., 2015), e.g., the lineage-specific regulators (Liu et al., 2015; O'Brien et al., 2018) or additional pathway members (Kabakchiev and Silverberg, 2013; Zhang et al., 2018). However, the efficiency issues of eQTL methods are still widely focused in many methodology studies, such as how to illuminate the full structure of the eQTL data (Huang et al., 2009); how to distinguish true causal polymorphisms or causal factors (Suthram et al., 2008; Lee et al., 2009; Chipman and Singh, 2011; Wang and Zhang, 2011); how to implement multiple-comparison adjustment (Chen et al., 2008) or confounding factor removal (Ju et al., 2017; Yuan et al., 2017); how to detect group-wise and individual associations between SNPs and expression traits (Cheng et al., 2015; 2016); and how to calculate fast for the computationally intensive part of the eQTL identification algorithm (Shabalin, 2012). Particularly, an urgent task nowadays is to capture the functional impact of detected eQTLs. Although many of the conventional eQTL methods tend to use the network concept or model to interpret the biological or biomedical significance of their discovery (Sun et al., 2007; Verbeke et al., 2013; Cheng et al., 2014; Ho et al., 2014; Zhang and Kim, 2014; De Maeyer et al., 2016), most of them derive the associations between SNP and gene groups rather than between SNP and gene-pair/edge groups (networks); i.e., they usually disregard the systematical role of those discovered SNPs or genes, thereby overlooking many network-associated phenotypic determinates.

Such a problem motivates us to investigate network QTL (nQTL), which indicates the cascade association of genotype → network → phenotype rather than traditional genotype → expression → phenotype in eQTL. Of note, conventional SNP–gene–gene triplets (Fraser et al., 2010; Budach et al., 2016) would directly contain two SNP–gene pairs as eQTLs. By contrast, nQTL consists of associated genotype on the SNP variant and network on gene co-expression. Several recent works have combined the eQTL and gene co-expression (module) (Kugler et al., 2013; Saha et al., 2017; Kolberg et al., 2020; Cui et al., 2021). Dissimilar to them, the SNP–gene involved in nQTL is not strictly limited to eQTL; thus, the identification of nQTL would be more general to model the genotype–phenotype association on the gene network level, which might be disregarded on the gene level.

In this work, we propose a new nQTL analysis framework to study the associations from genotypes to networks and further to phenotypes at a system level (**Figure 1A**). The first stage is to convert the transcriptome data of individual samples from genes' expression data into gene-pairs' correlation-like data based on the single-sample network (SSN) theory and method (Zhang et al., 2014; Hu et al., 2018) (see **Figure 1B** and step 1 of the nQTL framework in Methods); then the second stage is to apply the MatrixEQTL (Shabalin, 2012) approach to capture the edge/network traits, i.e., associations between SNPs and edges/gene-pairs across samples (see **Figure 1C** and steps (ii) and (iii) of the nQTL framework in Methods), in contrast with expression/gene traits identified by conventional eQTL methods; the third stage is to extract the edge/network signatures (i.e., the network signature generally includes a set of edge signatures where one edge signature is an edge/gene-pair corresponding to network traits), anchoring with hot SNPs (see **Figure 1D** and steps (iv) and (v) of the nQTL framework in Methods); in particular, an edge trait (a gene-pair trait) of nQTL in a specific sample includes additional association information from other samples (same type or phenotype), different from traditional co-expression gene-pair (with the information of only this specific sample) due to our edge-like correlation scheme. And the last stage is to infer the links/associations between network signatures and targeted (interested) phenotypes/factors, which construct the complete cascade associations of genotype → network → phenotype (see **Figure 1E** and steps (vi) and (vii) of the nQTL framework in Methods). The simulation studies demonstrate the efficiency of the

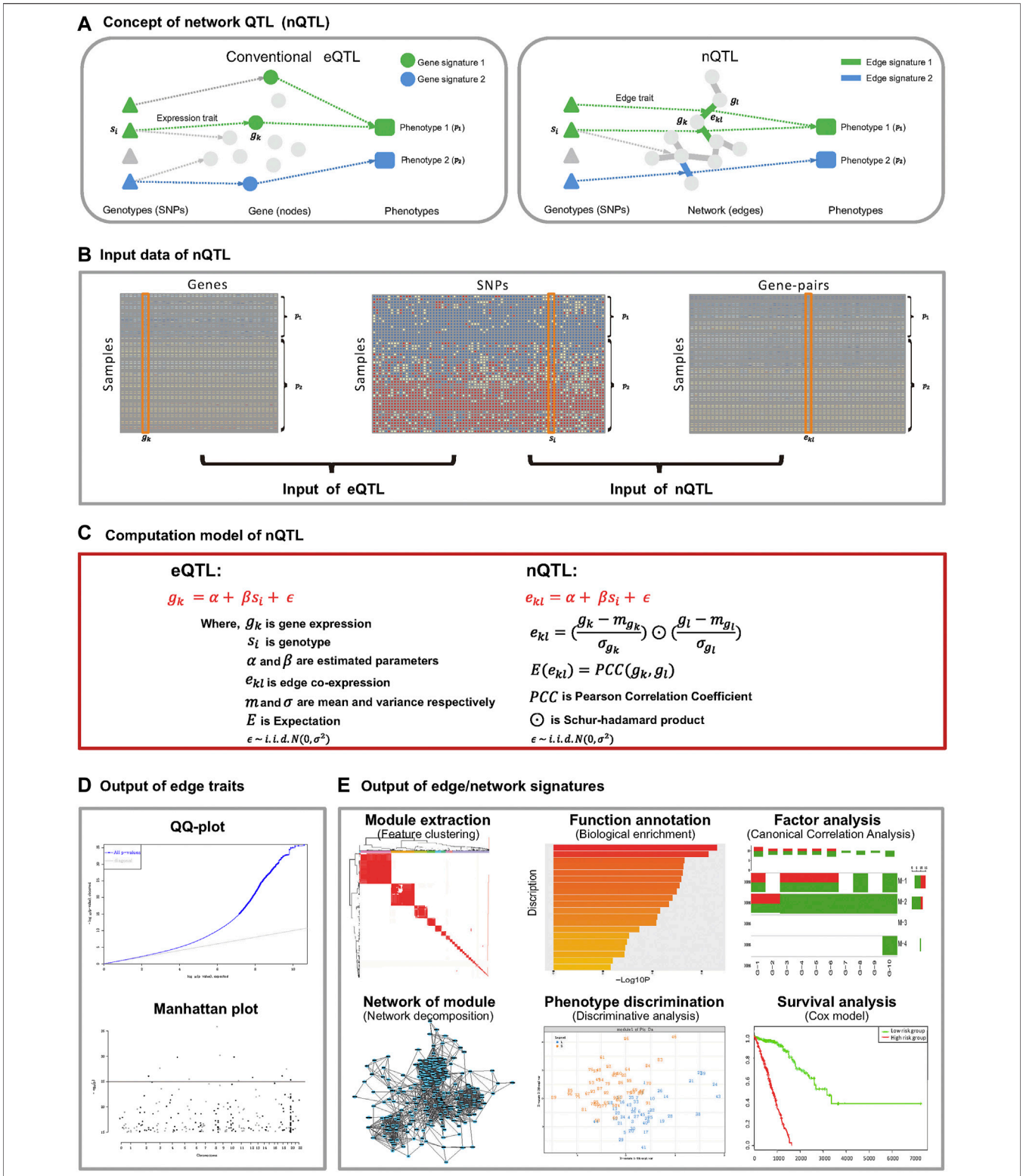


FIGURE 1 | Flow chart of the nQTL framework. It includes upstream nQTL identification and downstream biological and biomedical significance analyses of nQTL gene, network, and module. **(A)** Concept of nQTL compared to conventional eQTL. **(B)** Organization of input data for eQTL and nQTL analyses. **(C)** Computational model for eQTL and nQTL analyses. **(D)** The output of nQTL analysis as a network trait, which is different from the conventional expression/gene trait of eQTL analysis. **(E)** Downstream output of network signatures in the nQTL analysis framework, including nQTL module identification, biological enrichment of module, correlation between module and relevant factors, network structure of module, discriminative model, and Cox model based on the nQTL module.

nQTL framework to capture network traits, compared with traditional eQTL methods. And several case studies by the nQTL framework have been carried out for various real datasets. The first case study on lung cancer data detected lung-cancer-associated network signatures, whose efficiency was validated in an independent survival analysis. The second case study on Cancer Cell Line Encyclopedia (CCLE) tumor cell line data supports the efficiency of the nQTL framework again by providing additional evidence of network signature interpretation on both disease and drug treatment phenotypes. Importantly, the third case study on both healthy human bulk and single-cell RNA-seq data identified immune-associated network traits, network signatures, and *a posteriori* targeted phenotypes (e.g., inflammatory factors) with the corresponding genotypes at the single cell level. Collectively, these results show that the nQTL framework as a novel computational tool can detect the global genotype–network–phenotype associations, which reveals the underlying network traits and network signatures, indicating the significantly functional impacts of nQTL on biology and biomedicine.

MATERIALS AND METHODS

Edges in nQTL

For nQTL analysis, the input data in concept consists of two data matrices. One is the sequence variant data matrix (e.g., SNP data of each sample), and the other one is the network data matrix (e.g., co-expression data of individual samples), where a row represents a gene-pair/edge, a column represents the matched sample as same as that in the sequence variant data matrix, and every element represents the co-expression level of one gene-pair/edge in one sample. Notably, the sample-specific co-expression is a key to achieve nQTL analysis. Recent studies have supplied a few methods to estimate such network measurements in one sample. Considering the generality of nQTL analysis, the single-sample Pearson correlation coefficient (PCC) is adopted in this work by edge-like transformation (Zeng et al., 2014) and can be viewed as a correlation-like expression at an individual sample level because the summation of such information for all single-samples is actually the PCC in terms of mathematical representation. This kind of measurement is originally developed to explain the heterogeneity of cancer samples (Zhang et al., 2014; Yu et al., 2015; Zhang et al., 2015) and has also been applied in some other fields recently (Hu et al., 2018; Lu et al., 2019). Given the original gene expression matrix, the *PCC* between genes i and j in population samples can be calculated as follows:

$$PCC(i,j) = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_{ik}-\mu_i}{\sigma_i}\cdot\frac{x_{jk}-\mu_j}{\sigma_j}\right) = \frac{1}{n}\sum_{k=1}^{n} e_{ijk}$$

where there are n samples; for sample k, its expressions on genes i and j are x_{ik} and x_{jk}, respectively; the average expression and variance for gene i (or gene j) on population samples are μ_i and σ_i (or μ_j and σ_j), respectively. The strength of each edge is obtained by transforming the expressions of two genes (i.e., x_{ik} and x_{jk}) to the edge-like correlation of the gene-pair (i.e., e_{ijk}) in sample k. Clearly, as indicated by the above equation, the mean of the edge-like correlation of a gene-pair in all samples is just the PCC (i.e., $PCC(i, j)$) on all n samples, so that this measurement has an equivalent numerical meaning for any sample. Repeating this data transformation for candidate gene-pairs (e.g., all matched gene-pairs in numeric or prior selected gene-pairs in a biological context), we obtain the network data matrix. Finally, MatrixEQTL can be applied to infer the association between SNPs and gene-pairs (i.e., network trait) directly and efficiently. And as mentioned above, many other alternative sample-specific network construction methods could be applied in the nQTL framework (Liu et al., 2016; Yu et al., 2017; Dai et al., 2019; Kuijjer et al., 2019; Tanaka et al., 2020), although they would hold by particular mathematical/biological hypothesis different from conventional QTL analysis. For example, the SSN approach would associate genotype with sample-specific network changes due to its different quantifications on Pearson correlation change of gene-pairs (Liu et al., 2016; Yu et al., 2017), and the cell-specific network (CSN) approach could link genotype to nonlinear correlation of gene-pairs on the basis of mutual information (Dai et al., 2019).

nQTL Framework

Based on our nQTL concept and model, the whole nQTL framework is designed and implemented to carry out the association cascade analysis among genotype → network → phenotype. It includes several calculation steps (where the front three steps are identifying nQTL and the remaining steps are evaluating the biological or biomedical significance corresponding to genes, networks, and modules involved in nQTL):

(i) Preparing the datasets for nQTL analysis, e.g., genotype dataset, molecular phenotype dataset, physiological phenotype dataset, and some optional datasets (e.g., covariates). The network (phenotype) dataset, especially, can be produced from the molecular phenotype dataset (**Figure 1**), and it is a new data matrix whose row represents the gene-pair, column represents the sample, and matrix element indicates the single-sample co-expression level of one gene-pair on one sample. As required in such a study, all types of data should have values on the same samples; e.g., the samples and their ID labels should be matched in all datasets;

(ii) Carrying out the nQTL analysis for edge traits as introduced above, where an edge trait is the significant association between a SNP and an edge/gene-pair across samples and an expression/gene trait is the significant association between a SNP and a gene across samples;

(iii) Filtering the nQTL or eQTL associations (i.e., edge traits or expression traits) according to hot SNPs and hot gene-pairs/edges (or hot genes); e.g., the SNPs with significantly more interactive gene-pairs relative to other SNPs or the gene-pairs with significantly more interactive SNPs compared with other gene-pairs are kept in the following analysis (thus, their association matrix would be more dense than the original one);

(iv) Obtaining the network signatures (e.g., gene/gene-pair modules) based on their association profiles with SNPs,

e.g., the so-called nQTL module from hierarchical clustering on the nQTL association matrix, where a network signature is a cluster of edges corresponding to their edge traits (an edge signature is an edge corresponding to its edge trait) and a gene signature is a gene corresponding to its expression/gene trait;

(v) Estimating the functional enrichment of each network signature, e.g., GO or KEGG enrichment;

(vi) Inferring the link/association between network signatures (e.g., gene-pair co-expression measurements) and physiological phenotypes (e.g., clinical indices or factors) by partial least squares discriminant analysis (PLSDA) or canonical correlation analysis (CCA); for instance, in the following formula, when we expect to infer the association between a cluster of gene-pairs and a set of clinical indices (e.g., several markers), the co-expression data matrix of gene-pairs (e.g., X) and the phenotypic data matrix of indices (e.g., Y) on a group of same samples can be applied to calculate the canonical correlation (e.g., $cca(X, Y)$).

$$X' = a^T X$$
$$Y' = b^T Y$$
$$\mathrm{cca}(X,Y) = \max_{a,b} \frac{cov(X', Y')}{\sqrt{D(X')}\sqrt{D(Y')}}$$

(vii) Ranking and selecting the association cascades on genotype → network → phenotype based on significance of network trait, functional enrichment, and phenotypic relevance of network signatures together.

Besides, the "guilt-by-association" analysis can be carried out with nQTL, such as the comparison and combination with results from conventional eQTL, WGCNA, PCA, etc.

Note that an edge (gene-pair) is represented by its correlation strength (e_{ijk} for genes i and j on sample k), while a gene (node) is represented by its expression (x_{ik} for gene i on sample k) in this paper.

Synthetic Data Production and Method Assessment

For conventional eQTL analysis, the required data usually consist of two data matrices. One is the sequence variant data matrix (e.g., SNP data), where each row represents a SNP, each column represents a sample, and each element represents the genotype of one SNP in one sample. The other one is the expression data matrix (e.g., gene expression data), where one row represents a gene, one column represents the matched sample as same as that in the sequence variant data matrix (i.e., matched samples), and each element represents the expression level of one gene in one sample.

The general eQTL model of the association between SNPs and genes (i.e., expression trait) can be calculated as a simple linear regression or ANOVA. On the one hand, for each gene–SNP pair, with SNP encoded by the variant frequency, the linear association between gene expression g and genotype s is modeled as

$$g = \alpha + \beta \cdot s + \varepsilon,\ where\ \varepsilon \in \text{i.i.d. } N(0, \sigma^2)$$

On the other hand, for each gene–SNP pair, with SNP encoded by the categorical dominant effect, the ANOVA model for a linear regression between gene expression g and genotype s (e.g., s_1 and s_2 are dummy variables) is described as

$$g = \alpha + \beta_1 \cdot s_1 + \beta_2 \cdot s_2 + \varepsilon,\ where\ \varepsilon \in \text{i.i.d. } N(0, \sigma^2)$$

To effectively solve these models and handle covariate issues, the MatrixEQTL has been implemented to analyze different types of eQTL problems (Shabalin, 2012). Of note, in the above two models, the ε is measured for noise.

We generated eight sets of simulated datasets for identifying eQTL and nQTL, as supplied in MatrixEQTL (Shabalin, 2012).

For the first four sets of simulated data, the relation between gene and SNP (i.e., expression trait) is implemented by following formula:

$$\mathrm{pop} = 0.2{*}\mathrm{rnorm}(n)$$
$$\mathrm{snps} = \mathrm{rnorm}(n{*}nss) + \mathrm{pop}$$
$$\mathrm{ind} = \mathrm{snps}[, \mathrm{sample}(1{:}\ nss, ngs)]{*}(1{:}\ ngs)/ngs^{9/2}$$
$$\mathrm{gene} = \mathrm{rnorm}(n{*}ngs) + \mathrm{pop} + \mathrm{ind}$$

where n is the number of samples, ngs is the number of genes, nss is the number of SNPs, pop is the common signal in all variables, and ind is the individual genetic signal to gene expression. We set the sample number as 100, the gene number as 2,000 and 5,000, and the SNP number as 2,000 and 200,000. Obviously, there are four kinds of data combinations from gene and SNP. Of note, in order to make comparable eQTL evaluation with MatrixEQTL, we adopted the above data simulation approach as available through MatrixEQTL.

For the other four sets of synthetic data, SNP and gene data are generated randomly, and we associate gene-pair data with SNP (i.e., edge trait) according to the following formula:

$$\mathrm{gene} = \mathrm{rnorm}(n{*}ngs);$$
$$\mathrm{cordata} = \mathrm{cor} + \mathrm{pop} + \mathrm{ind};$$

where the additional variable cor means the value of gene-pairs. Then, we set the sample size as 100, the gene and gene-pair number as (2,000, 1,999*1,000) and (5,000, 4,999*2,500), and the SNP number as 2,000 and 200,000; again, there are four kinds of data combinations from gene/gene-pair and SNP.

Data Collection and Organization of Lung Cancer Dataset

As reported in the original study of Data GSE28571 (Micke et al., 2011), the surgically treated patients with primary non-small-cell lung cancer (NSCLC) were selected with fresh frozen NSCLC tissues, and patients with a history of other cancers or neoadjuvant treatment were excluded. Array experiments were performed according to the standard protocols for Affymetrix GeneChip Mapping 250K Nsp I arrays. In our reanalysis on these data, we have chosen 97 samples (as listed in **Supplementary Table S1**) because they have tested both the matched expression and SNP data with 261,549 SNPs and 23,322 annotated genes.

These samples also include short-term (<20 months; $n = 52$) and long-term survivors (>58 months; $n = 45$), which can be used as a clinical phenotype index. And the protein–protein interaction (PPI) network from the STRING database was applied to select gene-pairs for this, and following nQTL analysis on human data, where we have chosen the gene-pairs whose PPI score is >900. Thus, three preprocessed data matrices were used for conventional eQTL and new nQTL analyses. The SNP data *S* as a matrix consists of 261,549 rows and 97 columns, and each element indicates the SNP mutation type or score of one SNP in one sample. The expression data *E* consists of 23,322 rows and 97 columns, and each element indicates one gene's expression level in one sample. The co-expression data *C* consists of 147,645 rows and 97 columns, and each element indicates one gene-pair's co-expression level in one sample. Thus, *S* and *E* were the inputs of conventional eQTL analysis; *S* and *C* were the inputs of the new nQTL analysis.

Data Collection and Organization of CCLE Hematopoietic Data

We collected the CCLE: a compilation of gene expression, chromosomal copy number, and massively parallel sequencing data from 947 human cancer cell lines (Barretina et al., 2012). Coupled with pharmacological profiles for 24 anticancer drugs across 479 of these cell lines, this collection allowed identification of genetic, lineage, and gene-expression-based predictors of drug sensitivity. Genotyping/copy number analysis was performed by Affymetrix Genome-Wide Human SNP Array 6.0, and expression analysis used the GeneChip Human Genome U133 Plus 2.0 Array. Eight-point dose–response curves were generated for 24 anticancer drugs by an automated compound-screening platform.

We chose the hematopoietic and lymphoid tissue as a case study of nonsolid tumor. In this CCLE hematopoietic data, we selected 21,544 gene expression data and 906,600 SNP data for 151 matched samples (as listed in **Supplementary Table S1**). Of note, we focused on the nQTL compared to conventional eQTL, where the expression level of mRNA/protein is proportional to the quantitative trait and the prior-known gene/protein network are abundant. Thus, the data of annotated long noncoding RNA (ncRNA) or micro-RNA in the original CCLE hematopoietic data were not used in this study. In addition to these eQTLs and nQTLs in protein-coding genes, the QTLs in ncRNAs would also be attractive in future work, which could reveal the association between genotype and regulation of diverse ncRNAs (Ye et al., 2020).

Data Collection and Organization of the Human Peripheral Blood Mononuclear Cell (PBMC) Dataset

The original study focused on the cell-type-specific effects of genetic variation on genome-wide gene expression, which generated scRNA-seq data of ~25,000 PBMCs from 45 donors of the population-based cohort study Lifelines Deep (van der Wijst et al., 2018). To assure sufficient analysis power, cell types were merged to a more general classification: $CD4^+$ T cells, $CD8^+$ T cells, NK cells ($CD56^{dim}$ $CD16^+$ and $CD56^{bright}$ $CD16^{+/-}$), monocytes ($CD14^{bright}$ $CD16^-$ cMonocyte and $CD14^{dim}$ $CD16^+$ ncMonocyte), B cells and DCs ($CD1C^+$ myeloid, mDC, and plasmacytoid, pDC). In detail, the original study had proposed the concepts of co-expression QTLs by associating the SNP type and gene-pair co-expression based on single-cell data (van der Wijst et al., 2018), which can be thought of as a supervised way to infer the SNP and gene-pair association. It prepared the gene expression data of one individual/sample by averaging his/her single cells' gene expression, which could be thought of as a mimic bulk expression of individuals; and it also generated the gene co-expression data of one individual/sample by calculating the gene expression correlation of two genes on single cells from this individual; and then it carried out the eQTL analysis by combining the SNP *vs.* individual matrix and mimic bulk expression *vs.* individual matrix; it also detected co-expression QTL through the eQTL model on SNP *vs.* individual matrix and co-expression *vs.* individual matrix. Similar to such a study, in this work, we carried out eQTL and nQTL analyses, based on the expression data of $CD4^+$ T cells. Thus, in the 45 individuals, there are 7,478 genes in the expression matrix and 4,017,251 SNPs in the genotype matrix.

RESULTS

Assessment of nQTL Framework on Synthetic Datasets for Gene and Network Traits

Similar to the simulation experiment used by a conventional eQTL study (Shabalin, 2012), we generated eight sets of simulated datasets. The parameter setting includes the following: (1) for the first four datasets with preset expression traits, the sample size is 100, the gene number is from 2,000 to 5,000, and the SNP number is from 2,000 to 200,000; (2) for the latter four datasets with preset edge/network traits, the sample size is 100, the gene and gene-pair numbers are from (2,000, 1,999*1,000) to (5,000, 4,999*2,500), and the SNP number is from 2,000 to 200,000.

As known in GWAS, there will be a large deviation at one SNP site in the QQ-plot, suggesting that the deviation of the observed value of this SNP site is caused by the genetic effects of this SNP mutation. With the results shown in the QQ-plot, it is obvious that the eQTL analysis can find significant SNP–gene associations in the first four datasets with preset expression/node traits rather than other datasets with preset edge traits (**Figure 2A** and **Supplementary Figure S1**); the eQTL analysis can find the SNPs which have a bigger deviation; meanwhile, the nQTL analysis is comparable in spite of less deviation (**Figure 2B** and **Supplementary Figure S1**). By contrast, the nQTL analysis can detect significant SNP–(gene-pair) associations in the latter four datasets with preset edge traits (**Figure 2D** and **Supplementary Figure S1**), but the eQTL analysis only can detect a limited number of associations (**Figure 2C** and **Supplementary Figure S1**). In further quantitative comparison, the area under the curve (AUC) measurement was applied to compare the

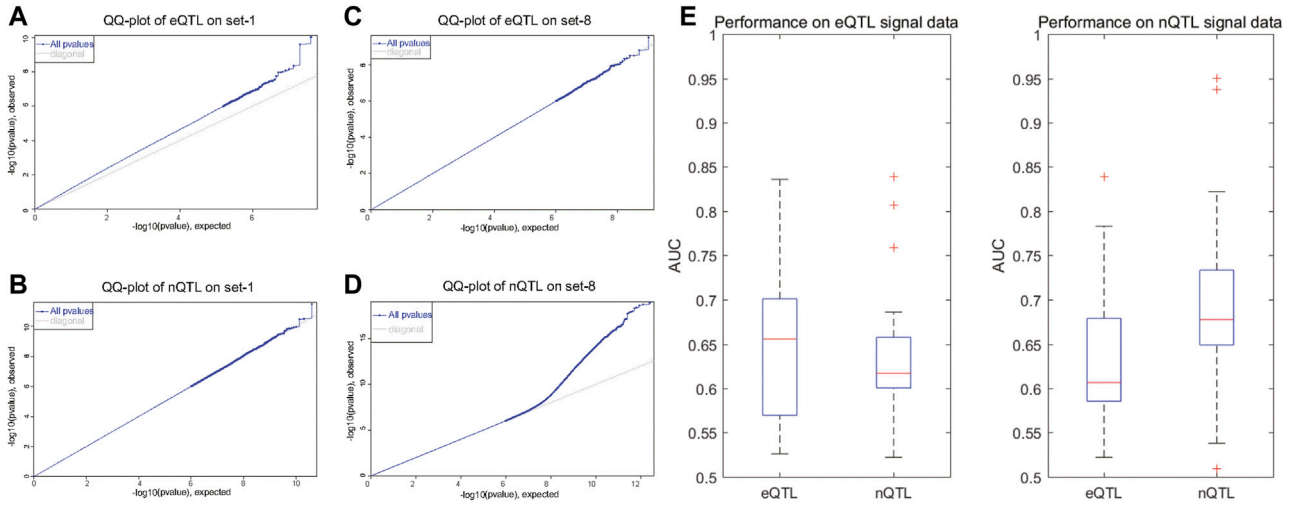


FIGURE 2 | Comparison of eQTL and nQTL identification on synthetic datasets based on expression traits and network traits. **(A,B)** QQ-plot of eQTL and nQTL, respectively, on a simulated data with preset expression traits. **(C,D)** QQ-plot of eQTL and nQTL, respectively, on a simulated data with preset edge traits. **(E)** AUC measurement comparison between nQTL and eQTL identifications on multiple datasets with preset expression traits. **(F)** AUC measurement comparison between nQTL and eQTL identifications on multiple datasets with preset edge traits.

method performance, which is an accuracy measurement considering the identification specificity and sensitivity under different threshold adoptions. As shown in **Figure 2E**, on multiple datasets with preset expression/node traits, the nQTL analysis has slightly small but comparable AUC values to those of eQTL (i.e., nonsignificant difference observed, $p = 0.3463$); meanwhile, as shown in **Figure 2F**, on multiple datasets with preset edge traits, nQTL analysis indeed has larger AUC values than those of eQTL analysis (i.e., significant difference observed, $p = 0.031$). These results on synthetic datasets support the ability and efficiency of the nQTL framework to detect novel network traits, in which each edge trait means an association between a genotype and a gene-pair co-expression (or an edge in network) and an expression trait (or a gene trait) means an association between a genotype and a gene expression (or a node in network).

nQTL Framework Identified Network Traits and Network Signatures Effective for Phenotype Discrimination and Disease Prognosis on Lung Cancer

We carried out the nQTL analysis on the public data GSE28571 (Micke et al., 2011) to detect new genotype–phenotype associations of lung cancer on a network level, with a comparison of conventional eQTL analysis.

According to the detected eQTLs (**Supplementary Table S2**) and nQTLs (**Supplementary Table S3**) with their QQ-plot (**Figure 3A** and **Supplementary Figure S2**), we found that the nQTL framework obviously tends to have more significant discoveries on edge traits. Several extremely significant SNPs, especially, can be filtered and shown in the Manhattan plot (**Figure 3B** and **Supplementary Figure S3**) by nQTL rather than eQTL analysis. Although these two methods can significantly detect similar SNPs and genes (**Supplementary Figure S4**), nQTL analysis can provide more alternative candidates than eQTL analysis. Of note, in this and following case studies, there is no more additional preprocessing or control (Micke et al., 2006; Nguyen et al., 2006; Botling et al., 2009; Platig et al., 2016) on the SNPs involved in nQTL or eQTL, which can reserve sufficient association information for downstream network and function analysis.

For the downstream analysis of the nQTL framework, the top-ranked 1,000 gene-pairs/edges (hot edges) with many SNP connections (i.e., each gene-pair can associate with more than 330 SNPs) are selected. Their association matrix can be extracted and clustered as in **Figure 3C** and **Supplementary Figure S5A**, and those edges can be obviously grouped into several network signatures (i.e., edge/gene-pair clusters in **Supplementary Figure S5B**). Many network properties of PPI subnetworks corresponding to such network signatures (e.g., **Supplementary Figures S6 and S7**) indicate that nQTL identification can effectively detect the SNPs associated with gene networks; by contrast, the gene cluster based on eQTL analysis would produce lots of small gene groups or even individual genes, which cannot reflect the biological network underlying QTL (**Supplementary Figure S7**). Here, a network signature is a gene-pair/edge cluster in terms of not their expressions but their correlations, whereas a gene signature is a gene in terms of its expression.

Then, according to the enriched biological functions of network signatures by enrichment analysis (**Supplementary Figure S8**), we found that nine network signatures have many enriched pathways or functions related to lung cancer, and especially one signature with our nQTL Module 1 (**Figure 3D**) is significantly associated with lung cancer. We also can find pathways in cancer (Takayama et al., 2006), small-cell lung cancer (Sanchez-Cespedes, 2003), and NSCLC (Breuer et al., 2005), which support that network signatures accurately detect the phenotype-associated common functions across samples.

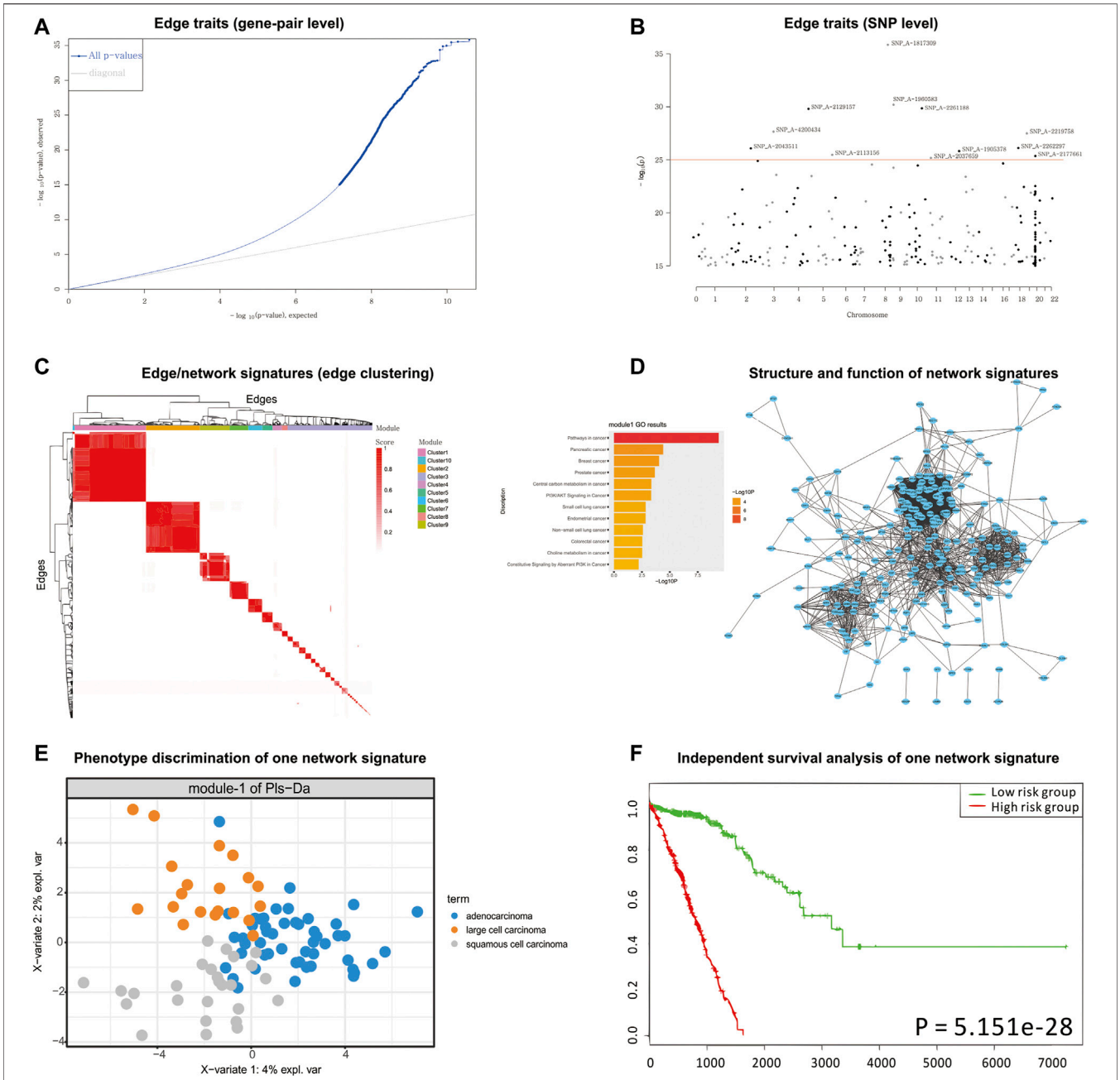


FIGURE 3 | nQTL analysis results from lung cancer dataset. **(A)** QQ-plot of edge trait on the gene level. **(B)** Manhattan plot of edge trait on the SNP level. **(C)** Network signature from edge/gene-pair clustering. **(D)** The network structure and function of one network signature. **(E)** PLSDA of one significant network signature for subtyping a disease. **(F)** Survival analysis and validation of one significant network signature.

Furthermore, some detailed signaling-related functions are also found, such as the phosphatidylinositol 3′-kinase (PI3K)-Akt signaling pathway (Hers et al., 2011), which is a key pathway associated with distinct metabolic phenotype characterized by low oxidative phosphorylation in small-cell lung cancer (SCLC) (Tripathi et al., 2017) or tumor angiogenesis in NSCLC (Zhou et al., 2017); Ras signaling pathway, which has been reported to be related with the proliferation, migration, and invasion of NSCLC cells (Levallet et al., 2012; Abdel-Rahman, 2016; Zheng et al., 2017); and MAPK signaling pathway whose activation would regulate lung cancer growth (Tang et al., 2016) and promote cell invasion and metastasis in NSCLC (Levallet et al., 2012). Besides, we also observed two well-known functions like the p53 signaling pathway and cell cycle pathway. The most recent studies have revealed that the suppression of the p53 signaling pathway could accelerate tumor progression in lung cancer (Hao et al., 2018) and that the cell cycle pathway is involved in leptomeningeal metastasis of NSCLC (Fan et al., 2018).

Finally, based on the subtypes of lung cancer available in the original study (e.g., large cell carcinoma, adenocarcinoma, and squamous cell carcinoma) (Micke et al., 2011), PLSDA can be executed and visualized on each network signature (**Figure 3E** and **Supplementary Figure S9**), and most network signatures showed significant phenotype discriminations on different subtypes. Besides, based on the binary survival data available in the original study (e.g., two groups of patients with long survival time or short survival time) (Micke et al., 2011), network signatures display satisfactory discrimination power again by PLSDA (**Figure 3E**, **Supplementary Figures S9 and S10**, and **Supplementary Table S4**). In addition, we also carried out independent survival analysis by nQTL network signatures' genes on independent lung cancer datasets from SurvExpress (Aguirre-Gamboa et al., 2013) (**Figure 3F** and **Supplementary Figure S10**), where all network signatures have *p*-values less than 0.05. Of note, compared to traditional methods using gene expression alone (e.g., WGCNA co-expression modules), nQTL analysis would be more efficient (**Supplementary Table S5**), where the nQTL framework has found six modules with a *p*-value <1E−10 and three modules with a *p*-value <1E−20; meanwhile, WGCNA has found five modules with a *p*-value <1E−10 and two modules with a *p*-value <1E−20. Thus, nQTL analysis can build functional association cascade as genotype → network (module) → phenotype. Furthermore, the nQTL framework also displays competitive and robust performance on distinguishing different cancer subtypes (**Supplementary Figures S11 and S13**), because all WGCNA modules had worse effects on large cell carcinoma identification (**Supplementary Figures S12 and S14**), indicating potential discovery bias of conventional methods. For example, the fifth WGCNA module can only separate samples from adenocarcinoma and squamous cell carcinoma but can mix samples from large cell carcinoma (**Supplementary Figure S12**); meanwhile, the AUC of the corresponding PLSDA (e.g., large cell carcinoma *vs.* others) is lower at 0.6284 (**Supplementary Figure S14**). Besides, on another large independent survival analysis on TCGA data (Gao et al., 2013), we have found one nQTL module with a *p*-value of 2.03E−03, which is below the Bonferroni significance threshold when doing 10 tests; by contrast, no WGCNA modules show such significance (**Supplementary Table S5**). Of course, there are indeed many ways to factorize or cluster expression data, and the additional benchmark comparison in the future would recover more characteristics of gene clusters or gene subnetworks associated with genotype, e.g., nQTL modules.

nQTL Framework Identified Network Traits and Network Signatures Effective for Drug Response Phenotype Discrimination and Interpretation on Cancer Cell Line

In this second case, we carried out nQTL analysis on the public CCLE dataset (Barretina et al., 2012) to detect new genotype–phenotype associations in different cancer cell lines on a network level, and we chose the hematopoietic and lymphoid tissues as an example of nonsolid tumor to illustrate the results.

According to the detected eQTLs (**Supplementary Table S6**) and nQTLs (**Supplementary Table S7**) with their QQ-plot (**Figure 4A** and **Supplementary Figure S15**), we found that nQTL analysis obviously tends to have more significant discoveries on edge traits and concentrate on less genes than eQTL analysis, although these methods have significant overlap on the detected SNPs and genes (**Figure 4B** and **Supplementary Figures S16 and S17**).

By the downstream analysis of the nQTL framework on the hematopoietic and lymphoid data, the top-ranked 1,000 edges (hot edges) with many SNP connections (more than 2,000) are selected. Their association matrix can be extracted and clustered as in **Figure 4C** and **Supplementary Figure S18A**, and those edges can be obviously grouped into several network signatures (**Supplementary Figure S18B**). Again, many network properties of PPI subnetworks corresponding to these network signatures (e.g., **Supplementary Figure S19**) indicate that nQTL identification can effectively detect the SNPs associated with interacting genes; meanwhile, the gene cluster based on eQTL analysis would only detect separate or incomplete structures of biological networks underlying QTL (**Supplementary Figure S19**).

Next, considering the enriched biological functions of network signatures (**Supplementary Figure S20**), we found they have similar enriched pathways or functions relevant to lymphoid cancer and leukemia, and especially the signature Module-1 (**Figure 4D**) is significantly associated with lymphoid cancer. We also found pathways in cancer, lymphocyte differentiation, and lymphocyte migration, which support that network signatures are able to detect the phenotype-associated shared functions across samples efficiently. Some detailed signaling-related functions were also found. The PI3K-Akt signaling pathway is activated by many types of cellular stimuli or toxic insults and regulates fundamental cellular functions such as transcription, translation, proliferation, growth, and survival (Engelman et al., 2006). The mammalian (mechanistic) target of rapamycin (mTOR) is a highly conserved serine/threonine protein kinase, and NPM-ALK induces GC resistance in lymphoid cells through the activation of the mTOR signaling pathway (Gu et al., 2008). The tumor necrosis factor (TNF), as a critical cytokine, can induce a wide range of intracellular signaling pathways including apoptosis and cell survival as well as inflammation and immunity (MacEwan, 2002).

Then, based on the binary category of cell lines derived from disease samples available in the original study (e.g., two groups of patients with acute leukemia and chronic leukemia (i.e., A *vs.* C) or two groups of patients with myelogenous leukemia and lymphocytic leukemia (i.e., M *vs.* L)), PLSDA can be executed, evaluated, and visualized for each network signatures' association with phenotypes (**Figure 4E** and **Supplementary Figure S21**). Many network signatures (i.e., correlations between genes) actually showed significant phenotype discriminations on two groups of samples (**Supplementary Figure S8**). Furthermore, we carried out independent survival analysis for these network signatures' genes by using independent data of chronic lymphocytic leukemia (Ramsay et al., 2013) (**Figure 4F** and **Supplementary Figure S22**), where seven of eight network

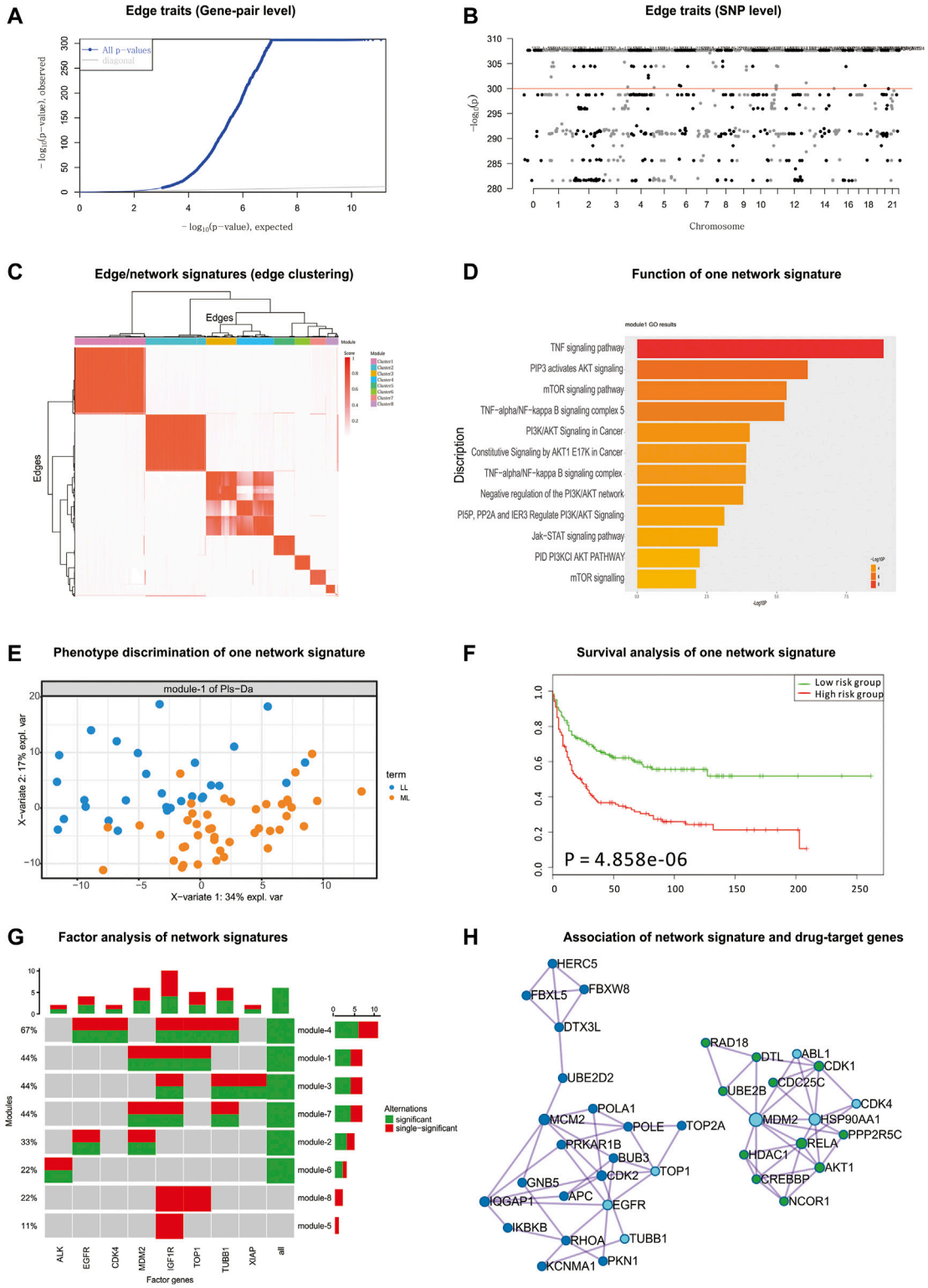


FIGURE 4 | nQTL analysis results from CCLE hematopoietic data. **(A)** QQ-plot of edge trait on the gene level. **(B)** Manhattan plot of edge trait on the SNP level. **(C)** Edge/network signature from edge/gene-pair clustering. **(D)** The biological function of one significant network signature. **(E)** PLSDA of one significant network signature for two groups of patients with myelogenous leukemia and lymphocytic leukemia (i.e., M *vs.* L). **(F)** Survival analysis and validation of one significant network signature. **(G)** CCA of network signatures and drug-target genes. **(H)** Association among network signature genes and drug-target genes on the STRING network, where nodes in dark blue and green are two groups of signature genes and nodes in light blue are drug-target genes.

signatures have *p*-values less than 0.05 and half of eight modules are still significant after multiple testing corrections, which indicates the efficiency of the nQTL framework and its network signature discovery for disease subtyping and prognosis potential.

Considering mainly two different hematopoietic and lymphoid cell lines, we have summarized the differentially expressed genes between different types of cell lines. We found that on average only 7% of genes in the nQTL module were differentially expressed. Thus, cell type would be a partial driver of the conventional co-expression module; however, the nQTL approach could discover many new modules relevant to genotype independent of cell type.

Finally, according to the drug and drug-target information from the CCLE database, we collected a drug-target gene list: IGF1R, TUBB1, MDM2, TOP1, EGFR, XIAP, CKD4, ALK, etc. By CCA, we detected the significant expression associations among our network signature genes and these drug-target genes. Interestingly, there are few associations between sole gene and network signatures; by contrast, the network signatures tend to associate with a combination of drug-target genes (**Figure 4G**). This is also supported by the STRING subnetwork among network signature genes and drug-target genes (**Figure 4H**), where TUBB1, TOP1, and EGFR can interact with a group of network signature genes enriched in the functional pathway "Pathways in cancer" and MDM2 and CKD4 can cooperate with another group of network-neighbor genes enriched in the functional pathway "Chronic myeloid leukemia".

nQTL Framework Identified Network Traits and Network Signatures Effective for Healthy Immunity Phenotype Discrimination on Both Human Bulk and Single-Cell Transcriptomes

In this third case, we conducted nQTL analysis on scRNA-seq-induced bulk expression data corresponding to $CD4^+$ T cells (van der Wijst et al., 2018), and we especially validated our discovered edge/network traits on the original scRNA-seq data in a unsupervised manner, which should be more efficient than the original method "co-expression QTL" (van der Wijst et al., 2018) when single-cell data are not available. Here, the supervision is different from that in the conventional classification model. Although co-expression QTL analysis did not require *a priori* selection of specific genes to test, it did require *a priori* selection of specific edges by a full scan of significant modifiers of edges. In other words, the co-expressions of edges are directly calculated from additional single-cell data, which can be considered as certain supervised information from external sources. By contrast, nQTL analysis is able to infer a similar association between SNPs and the correlated edges on the basis of bulk expression data, which would be validated by single-cell data independently. Thus, as summarized and shown in **Supplementary Figure S39**, we consider that nQTL analysis can make association analysis in an unsupervised manner, different from co-expression QTL analysis.

Of note, different from the disease study with prior-targeted phenotype, the healthy human data might have many potential phenotypes of interest. And we could focus on the nQTL analysis relevant to a post-targeted phenotype (i.e., immune factors or inflammatory factors), although some significant results might be relevant to other phenotypes.

Similarly, the obtained eQTLs (**Supplementary Table S9**), nQTLs (**Supplementary Table S10**), and their QQ-plots (**Figure 5A** and **Supplementary Figure S23**) suggest that nQTL analysis is able to identify more significant discoveries on edge traits and that nQTL analysis indeed can detect more phenotype-associated SNPs than eQTL analysis according to the extremely significant SNPs filtered (**Figure 5B** and **Supplementary Figure S24**). Meanwhile, nQTL analysis can provide more alternative candidates than eQTL analysis although these two methods can find significantly overlapping SNPs and genes (**Supplementary Figure S25**).

Next, by selecting the top-ranked gene-pairs (hot edges) with significantly more associations to all SNPs than other edges, the association matrix can be extracted and clustered (**Figure 5C** and **Supplementary Figure S26A**), based on which those edges/gene-pairs can be grouped into a few network signatures (**Supplementary Figure S26B**). The network properties of PPI subnetworks induced from nQTL modules corresponding to these network signatures reveal larger network centralities than those induced from eQTL analysis (**Supplementary Figures S27 and S28**). Thus, the nQTL framework can effectively detect the SNP-associated gene networks, and the network signatures (**Figure 5D**) would be more general on annotating biological functions rather than simply considering network structure. In accordance with the biological background of the data, we observed several pathways relevant to immunity, viruses, and antigens in the biological context of PBMC. Enriched in antigen processing and presentation pathway, the genes encoding MHC class II molecules are transcribed according to a strict cell-type-specific and quantitatively modulated pattern, which is pivotal for the adaptive immune system by guiding the development and activation of $CD4^+$ T-helper cells MHC class II molecules (Reith et al., 2005). And enriched in the autoimmune thyroid disease pathway, both autoantibodies and thyroid-specific cytotoxic T lymphocytes (CTLs) have been proposed to be responsible for autoimmune thyrocyte depletion (Stassi and De Maria, 2002), and self-reactive $CD4^+$ T lymphocytes (Th) recruit B cells and $CD8^+$ T cells (CTL) into the thyroid (Tsatsoulis, 2006).

Considering that the enriched biological functions for many network signatures (**Figure 5E** and **Supplementary Figure S29**) actually show significant relations with viruses, interferons, antigens, and especially immunity, we collected and used literature-reported immune factors or inflammatory factors as a post-targeted phenotype for the following association inference. These functional factors include IFIT5, IFITM1, IFIT2, IFI6, ISG20, ISG15, IFI35, IFIT3, and HERC6. By calculating the CCA between a group of gene-pairs from one network signature and a group of functional factors (**Figure 5F** and **Supplementary Figure S30**), several network signatures showed significant relevance with many functional factors, underlying the onset of the immune system of healthy

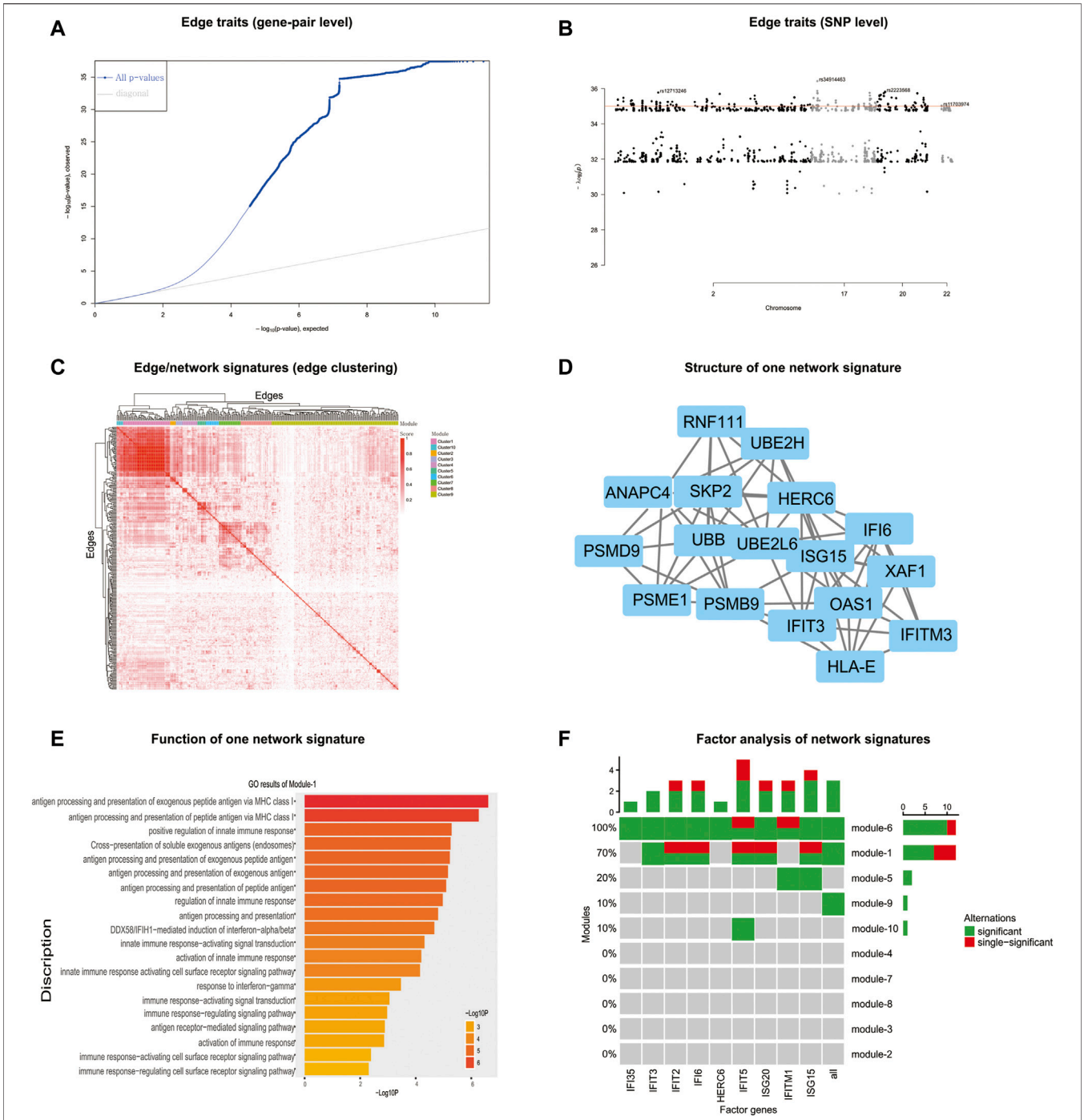


FIGURE 5 | nQTL analysis results from the human PBMC dataset. **(A)** QQ-plot of edge trait on the gene level. **(B)** Manhattan plot of edge trait on the SNP level. **(C)** Network signature from edge/gene-pair clustering. **(D)** The network structure of one network signature. **(E)** GO analysis of one significant network signature. **(F)** CCA of network signatures and immune factors or inflammatory factors.

humans. On the other hand, we also carried out survival analysis by these network signature genes on independent pan-cancer datasets, assuming the immune-associated survival risks detectable in our identified network signatures. On six cancer datasets with survival data from more than 500 samples (**Figure 6**, **Supplementary Figures S31–S36**, and **Supplementary Table S11**), all network signatures have *p*-values less than 0.05; and even after multiple testing corrections, there are still about 83% network signatures showing significant survival analysis (**Supplementary Table S11**), again indicating the efficiency of nQTL discovery.

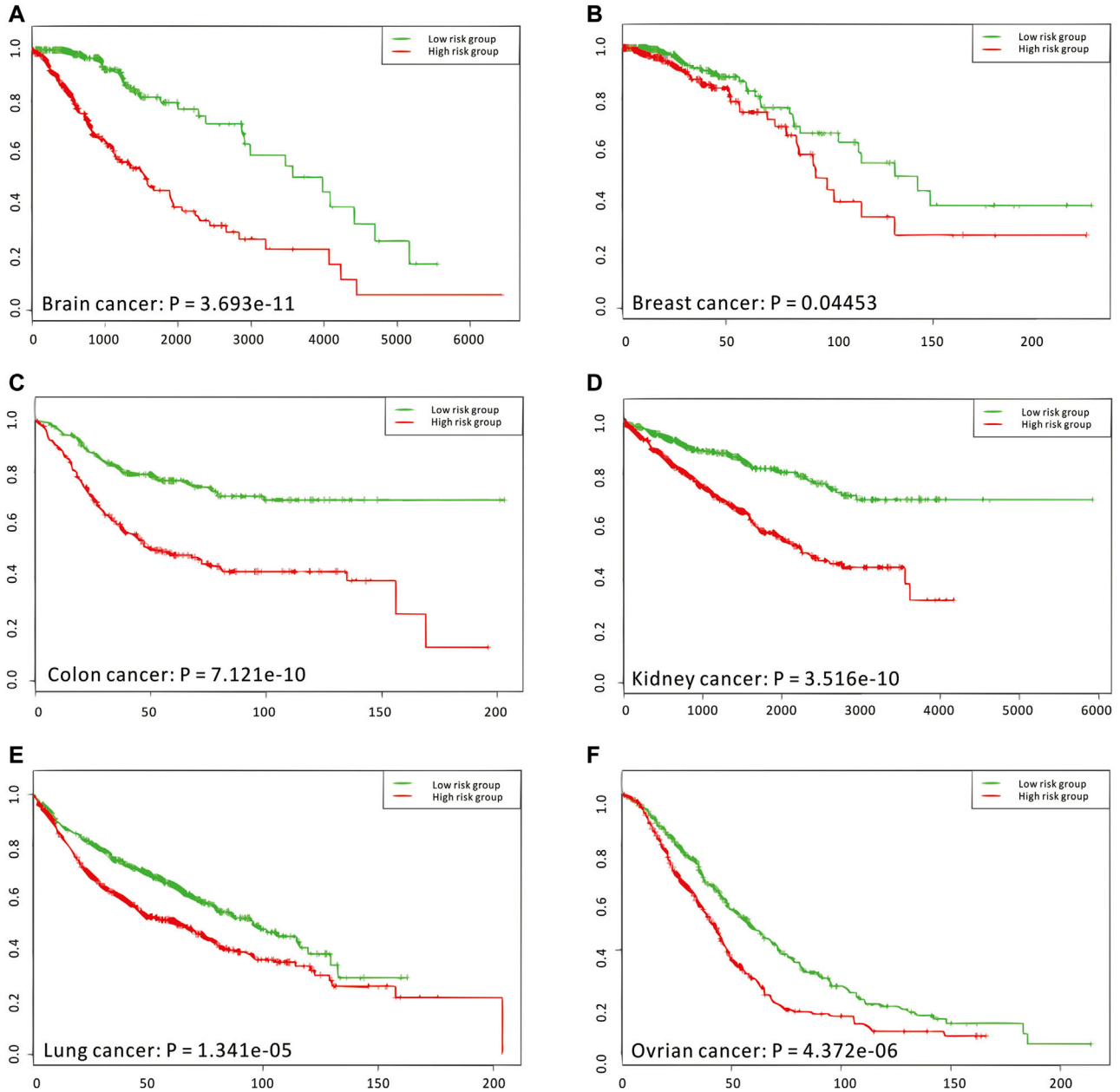


FIGURE 6 | Independent survival analyses of nQTL modules from the human PBMC dataset. **(A–F)** The validation on brain, breast, colon, kidney, lung, and ovarian cancer, respectively. And these analysis results are obtained from the SurvExpress web server, where *p*-values are produced from the Cox model.

As our design, the nQTL framework is able to detect the potential association between SNP and edge in an unsupervised manner. In fact, the strategy of one work to find "co-expression QTL" is directly using single-cell data to calculate the co-expression of a gene-pair (van der Wijst et al., 2018); however, it would not always be applicable in many applications due to there being no individual-specific single-cell data available. Actually, the original work reported only one detailed case of co-expression QTL, involving a gene-pair RPS26 and RPL21. In **Figures 7A,B** and **Supplementary Figure S37A**, they have differential co-expression corresponding to a particular SNP type. But it is obvious that these cells have distinguishing expression differences for different SNP types on gene RPS26; thus, this originally reported case would be one of the expected co-expression QTLs; however, it also dominated in conventional differential expression. By contrast, we selected two genes with the most significant association to one SNP according to our nQTL analysis on the given mimic bulk expression data, and we could also observe or validate their potential co-expression difference on unseen single-cell data (**Figures 7C–F** and **Supplementary Figure S37B**). Motivated by this idea, we rescreened the gene-pairs corresponding to edge traits from

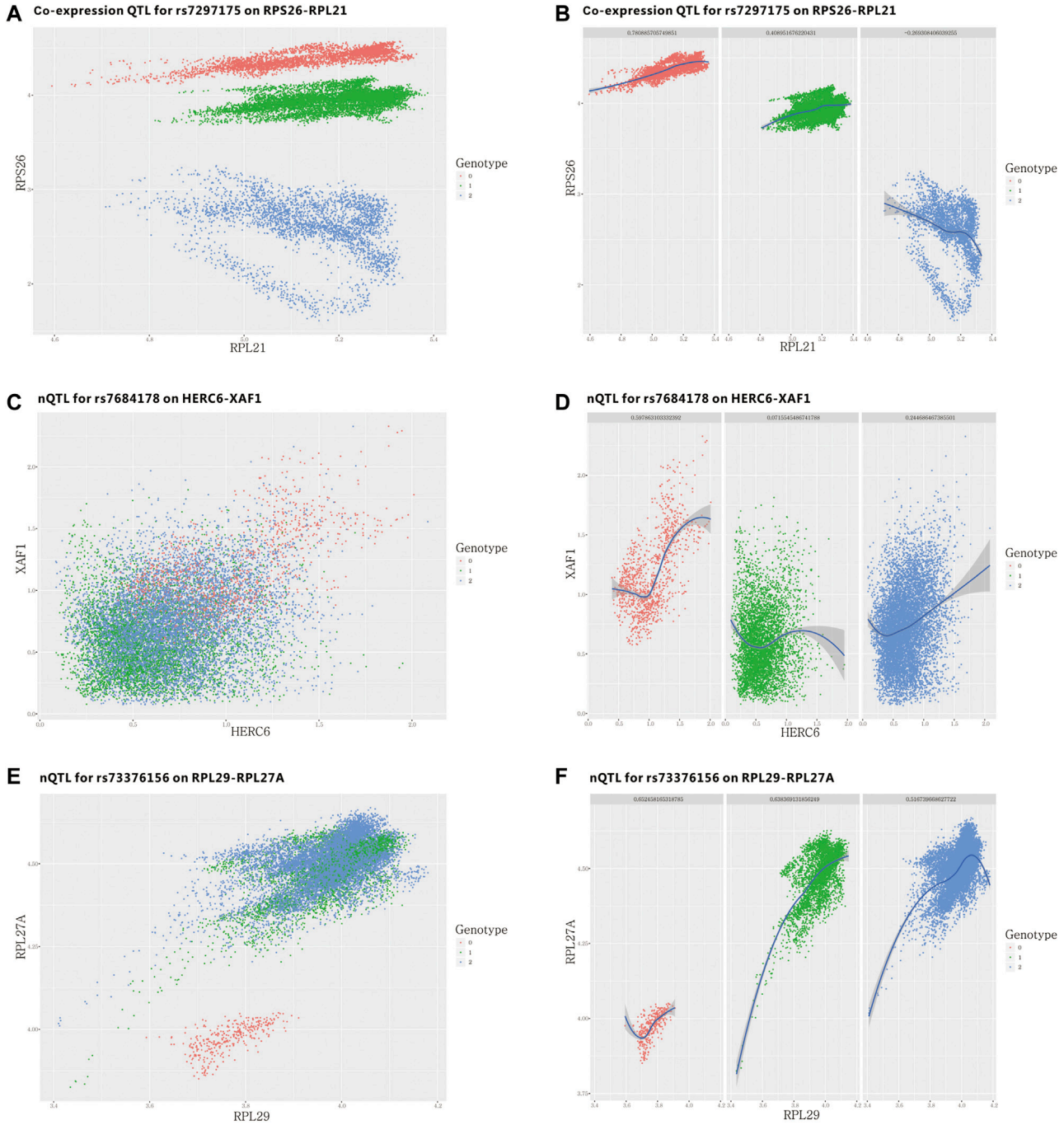


FIGURE 7 | SNP–edge associations (edge traits) detected unsupervised by the nQTL framework and their assessments on single-cell data from the human PBMC dataset. **(A,B)** Single-cell plot of previously reported SNP and gene-pair association. **(C–F)** Single-cell plot of our newly identified SNP and gene-pair associations.

our nQTL analysis on the mimic bulk expression data and further assessed their conditional co-expressions on the single-cell data from different groups of samples. As summarized in **Supplementary Figure S38**, nQTL analysis indeed can detect many similar "co-expression QTLs" without the supervised information from single-cell data, benefiting from the SSN approach.

DISCUSSION

The main merit of nQTL analysis is to capture the association of genotype and (molecular) phenotype on a sample-specific network level rather than a population level, which is different from previous methods including "co-expression QTLs" (van der Wijst et al., 2018) and "context-specific eQTL" (Zhernakova et al.,

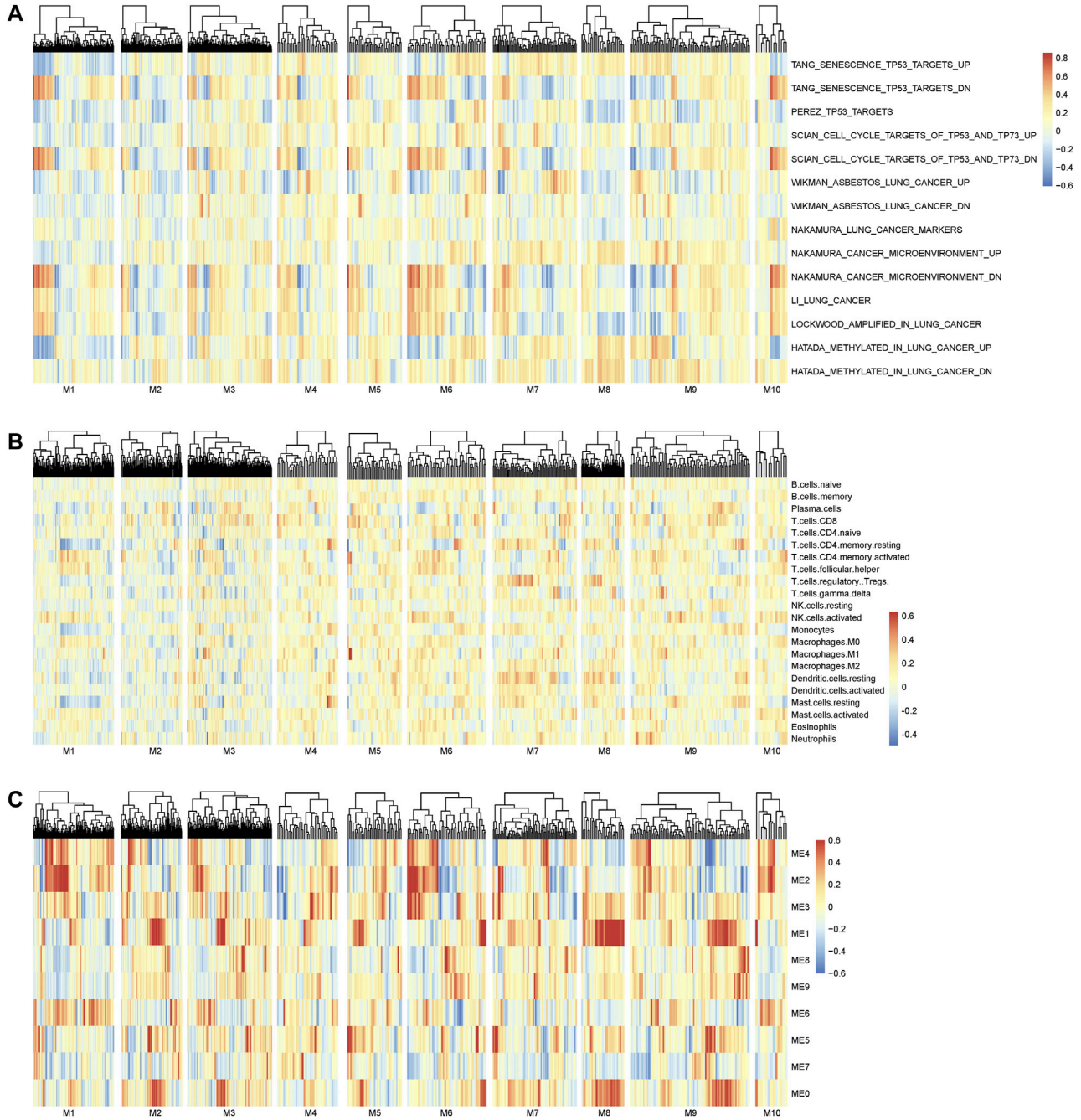


FIGURE 8 | Association analysis on the nQTL module and different factors. **(A)** The association between the nQTL module and (GSVA) pathway enrichment/activity. **(B)** The association between the nQTL module and (CIBERSORT) cell composition. **(C)** The association between the nQTL module and (WGCNA) co-expression module. A few nQTL modules are associated with or driven by potential biological mechanisms, e.g., pathway function, cell composition, or gene co-expression. And many other nQTL modules should be new ones that are determined by particular genotypes through biological network rewiring, according to their GO enrichment and network association.

2017). Of course, a feature of nQTL would reflect the association between genotype and network activity. As already known, the inference of network activity from its member genes' expressions would be an option when the experimental measurements of network activity are not directly available at present, which could be similarly used to build the association between genotype and network activity in the nQTL framework. Meanwhile, how to accurately infer the network activity is still another open question, which would be further investigated in our future work. With such an assumption, we aim to provide a proof-of-principle case

study on lung cancer data, by associating the genes' expressions from the nQTL module to sample-specific pathway enrichments from GSVA (Hanzelmann et al., 2013). As shown in **Figure 8A**, nQTL modules indeed can identify several networks altered by the genetic variants that are associated with a disease (e.g., subtype discriminations) and also be able to translate into upregulation or downregulation of a network/pathway (e.g., lung-cancer-related pathways).

And a previous work (Zhernakova et al., 2017) found that the different gene modules are not driven by individual biological networks but result in differences in cell type composition between the different studied samples. Thus, we have carried out additional investigation on the relationship between nQTL module and cell type composition. Mainly on the lung cancer data as a case study, we used CIBERSORT (Newman et al., 2015) to obtain the potential cell type composition of each sample and analyzed the associations between nQTL module gene expressions and cell type compositions. Indeed, there are a few module genes which are associated with a few cell types (**Figure 8B**), like naive CD8 or CD4 T cells. Thus, the nQTL framework can detect some modules similar to previous findings, which are partially driven/explained by cell type compositions. However, the nQTL framework is also able to find many new modules that cannot be simply explained by cell type compositions. Thus, cell type composition differences might act as a biomarker, worthy of future study.

Different from conventional co-expression-based modules, the nQTL module captures the genotype-associated co-expressions. It is possible that a transcriptional factor (TF) is an important mediator or regulatory driver between the genotype and molecular phenotype; thus, we have investigated the potential TF and its motifs for nQTL modules on the lung cancer data, by g: Profiler (Raudvere et al., 2019). We found that module1, module5, and module10 would be enriched by the E2F family, AHR, and NFR-1, respectively. Thus, nQTL modules are actually different from conventional co-expression modules like the WGCNA module (**Figure 8C**), and its relation with TF can be further detected directly by combining more experimental data including Hi-C and ATAC data. Of note, in these co-expression-based methods, genes within the same pathway or genes with a PPI are assumed to be co-expressed; however, their associations would be not necessarily related to particular genotypes. To avoid potential bias in pathway enrichment from co-expression, we could carry out multiple testing corrections for enriched pathways from multiple nQTL modules, and our findings are still consistent in this study (e.g., almost all enrichment p-values are less than the corrected threshold of 0.05/10, as shown in **Supplementary Figures S8, S20, and S29**).

It is well known that GWAS has provided many candidate loci for human phenotypes (e.g., complex diseases). However, more and more reports show that the efficiency of GWAS is limited due to the unknown functions of most recommended loci. Thus, the eQTL analysis is further adopted to investigate the reasonability of the locus by linking the locus' mutation type and gene expression type so as to introduce the molecular phenotype to explain the functional relevance of particular GWAS-suggested locus. Obviously, this is an integration strategy, which can combine the GWAS summary results and eQTL quantitative results together by the same candidate locus, e.g., SNPs. However, it is still limited by the loss of sample match information, and this means the risk to overlook individual specificity. By contrast, nQTL analysis can avoid this issue by integrating the multiple omics information from the same group of samples and introducing the sample-specific network to link the individual specific genotypes to networks and further to phenotypes, thus improving the genotype detection and functional explanation associated with the target phenotype. For instance, the summary-data-based Mendelian randomization analysis (Zhu et al., 2016) or Bayesian statistical framework (He et al., 2013) can be similarly used to integrate the GWAS summary data and nQTL analysis outcomes. Besides, the joint matrix decomposition model (Hu et al., 2020) is also an alternative approach for integrating GWAS and nQTL associations on the basis of their shared SNPs. Indeed, the identification of nQTL and GWAS combination with new knowledge of a common disease is worth studying in the future.

From a technical viewpoint, the key of nQTL analysis is to obtain the individual network for one sample from gene expression data. As well known to us, reconstructing a network generally requires multiple samples to estimate the associations between variables of the network, where the variables are represented as nodes/genes in terms of expression in a network and the variable associations are represented as edges/gene-pairs in terms of correlation/strength in the same network. Not only nQTL analysis but also other approaches in biological and biomedical studies are facing the problem of how to estimate the network (or its feature) when only one sample is available for a target phenotype. To address this issue, a few sample-specific network construction methods have been developed and applied. Differential edge-like transformation (DET) based on single-sample PCC (sPCC) has been developed by using the additivity of PCC on multiple samples and assigning the individual additive factors to each sample (Zeng et al., 2014), which can estimate the significant co-expression of two variables compared to the population average. Particularly, the mean of DET between two variables from a group of samples equals the PCC of this variable-pair on these samples. Meanwhile, the SSN method constructs the sample-specific differential network of one sample against a set of reference samples (Liu et al., 2016; Yu et al., 2017), which develops a statistic on the co-expression change when one sample is added into the reference samples. And the CSN focuses on the inference of network representation for each single cell from scRNA-seq data by transforming the data from an "unstable" gene expression form to a "stable" gene association form based on statistical independence (Dai et al., 2019). Besides, the Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS) reconstructs the individual specific network in a population of samples for each detailed sample, by calculating the correlation statistical significance between all samples and the samples without a given single sample (Kuijjer et al., 2019). In fact, the model of nQTL is comprehensive, and these sample-specific network construction methods are alternatives for it. Considering that the conventional eQTL model and its biological question always focus on the linear association

among population samples, the DET approach is adopted in this nQTL study. The integration of other methods such as SSN, CSN, and LIONESS is worthy of future evaluation, dependent on biological background and questions in research context. In addition to enriched genes and their associations within prior-known networks/contexts in the current nQTL analysis, using network reconstruction or prediction can further improve the power of nQTL analysis.

Different from GWAS/eQTL applications, nQTL analysis has great potential in studying many complex biological problems and provides new insights from a system/network viewpoint, e.g., individual specific association discovery on diverse cell types based on new single-cell technologies (van der Wijst et al., 2018; Camp et al., 2019). As known, using nQTL or eQTL analyses for the integration of omics data at multilevels requires multilevel data from the same experimental subject in concept. For example, in the conventional study, we required the SNP and expression data for the same person in eQTL analysis although the test sample might be from different locations or times for this person, and this kind of data is well known as matched-data organization. However, it is hard to obtain such multilevel data for the same cell with current single-cell technologies. Thus, we have to approximately apply nQTL or eQTL analysis or a similar integrative model in single-cell studies currently. It is necessary to develop integrative omics and analysis methods in focal research of wide single-cell fields. Of note, the potential extension of the nQTL concept and model on the integration of multi-omics data would still be burdened by correlations among and within the different omics layers. Although this method could help identify causal relations within and between omics layers, there are still many issues that need to be further studied, such as causality inference (Leng et al., 2020).

As a unified framework, our nQTL approach provides a systematical tool to support the whole analysis pipeline to detect the edge traits, edge signatures, and their relationship with phenotypes, which builds an association cascade from genotype to network and further to phenotype. Compared to conventional eQTL or GWAS analysis, nQTL analysis not only supplies the potential phenotype drivers corresponding to genotypes but also reveals functional insights of these relevant genotypes in the form of global network and local modules. Therefore, the development and adoption of advanced network inference methods and module decomposition approaches would benefit nQTL research by promoting the efficiency and robustness of analysis.

Of note, the quality control and significance correction should be further promoted for nQTL analysis. In these experiments, the used public datasets have no minor allele frequency (MAF) available, and thus, the MAF filter was not used further. In future studies with a focus on particular variants like GWAS outcomes, an nQTL study could use a MAF filter to narrow down the genotype analysis. Besides, in addition to the FDR used by MatrixEQTL, the empirical FDR adopted by "co-expression QTLs" (van der Wijst et al., 2018) and "context-specific eQTL" (Zhernakova et al., 2017) analyses was also calculated, where there are no result changes on the simulated data; and the empirical FDR is also low under the current experimental threshold setting in real datasets (see **Supplementary Table S12**). Indeed, the combination of different permutation strategies for significance correction should provide reliable information on the confidence of identified eQTLs and nQTLs.

CONCLUSION

Collectively, the nQTL framework is proposed to detect the association cascade: genotype → network → phenotype, which not only interprets the biological or biomedical significance of the discovered nQTL traits or features but also reveals the detailed function roles of those nQTLs as network determinates associated with phenotypes. Our multiple simulation studies and case studies on human healthy and disease phenotypes all support the efficiency of the nQTL framework, which can identify not only network traits for analyzing complex biological processes at a network level but also network signatures for phenotype discrimination.

DATA AVAILABILITY STATEMENT

The used data are from public databases like GEO (https://www.ncbi.nlm.nih.gov/geo/) and CCLE (https://portals.broadinstitute.org/ccle/). The nQTL codes are available in the GitHub repository, https://github.com/ztpub/nQTL, and other analysis scripts and data can be required from the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization: TZ, LC. Data curation: KY. Funding acquisition: TZ, LC. Methodology: KY, TZ. Software: KY. Supervision: TZ, LC. Visualization: KY, TZ. Writing—original draft: KY. Writing—review and editing: KY, TZ, LC.

FUNDING

This paper was supported by the National Key R&D Program (Nos. 2017YFA0505500 and 2017YFC1201200), the Strategic Priority Research Program of Chinese Academy of Sciences (Grant Nos. XDB38050200 and XDB38060100), the National Natural Science Foundation of China (Nos. 12131020, 11871456, 31930022, and 12026608), the Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01), and the Japan Science and Technology Agency Moonshot R&D (JPMJMS 2021). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.720321/full#supplementary-material

REFERENCES

Abdel-Rahman, O. (2016). Targeting the MEK Signaling Pathway in Non-small Cell Lung Cancer (NSCLC) Patients with RAS Aberrations. *Ther. Adv. Respir. Dis.* 10 (3), 265–274. doi:10.1177/1753465816632111

Aguirre-Gamboa, R., Gomez-Rueda, H., Martínez-Ledesma, E., Martínez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., et al. (2013). SurvExpress: an Online Biomarker Validation Tool and Database for Cancer Gene Expression Data Using Survival Analysis. *PloS one* 8 (9), e74250. doi:10.1371/journal.pone.0074250

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* 483 (7391), 603–607. doi:10.1038/nature11003

Botling, J., Edlund, K., Segersten, U., Tahmasebpoor, S., Engström, M., Sundström, M., et al. (2009). Impact of Thawing on RNA Integrity and Gene Expression Analysis in Fresh Frozen Tissue. *Diagn. Mol. Pathol.* 18 (1), 44–52. doi:10.1097/PDM.0b013e3181857e92

Breuer, R. H. J., Postmus, P. E., and Smit, E. F. (2005). Molecular Pathology of Non-small-cell Lung Cancer. *Respiration* 72 (3), 313–330. doi:10.1159/000085376

Budach, S., Heinig, M., and Marsico, A. (2016). Principles of microRNA Regulation Revealed through Modeling microRNA Expression Quantitative Trait Loci. *Genetics* 203 (4), 1629–1640. doi:10.1534/genetics.116.187153

Camp, J. G., Platt, R., and Treutlein, B. (2019). Mapping Human Cell Phenotypes to Genotypes with Single-Cell Genomics. *Science* 365 (6460), 1401–1405. doi:10.1126/science.aax6648

Chen, L., Tong, T., and Zhao, H. (2008). Considering Dependence Among Genes and Markers for False Discovery Control in eQTL Mapping. *Bioinformatics* 24 (18), 2015–2022. doi:10.1093/bioinformatics/btn373

Chen, M. H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J. E., Moscati, A., et al. (2020). Trans-ethnic and Ancestry-specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* 182 (5), 1198–e14. e1114. doi:10.1016/j.cell.2020.06.045

Cheng, W., Shi, Y., Zhang, X., and Wang, W. (2015). Fast and Robust Group-wise eQTL Mapping Using Sparse Graphical Models. *BMC Bioinformatics* 16, 2. doi:10.1186/s12859-014-0421-z

Cheng, W., Shi, Y., Zhang, X., and Wang, W. (2016). Sparse Regression Models for Unraveling Group and Individual Associations in eQTL Mapping. *BMC Bioinformatics* 17, 136. doi:10.1186/s12859-016-0986-9

Cheng, W., Zhang, X., Guo, Z., Shi, Y., and Wang, W. (2014). Graph-regularized Dual Lasso for Robust eQTL Mapping. *Bioinformatics* 30 (12), i139–i148. doi:10.1093/bioinformatics/btu293

Chipman, K. C., and Singh, A. K. (2011). Using Stochastic Causal Trees to Augment Bayesian Networks for Modeling eQTL Datasets. *BMC Bioinformatics* 12, 7. doi:10.1186/1471-2105-12-7

Cui, Y., Zeng, X., Xiong, Q., Wei, D., Liao, J., Xu, Y., et al. (2021). Combining Quantitative Trait Locus and Co-expression Analysis Allowed Identification of New Candidates for Oil Accumulation in Rapeseed. *J. Exp. Bot.* 72 (5), 1649–1660. doi:10.1093/jxb/eraa563

Dai, H., Li, L., Zeng, T., and Chen, L. (2019). Cell-specific Network Constructed by Single-Cell RNA Sequencing Data. *Nucleic Acids Res.* 47, e62. doi:10.1093/nar/gkz172

De Maeyer, D., Weytjens, B., De Raedt, L., and Marchal, K. (2016). Network-Based Analysis of eQTL Data to Prioritize Driver Mutations. *Genome Biol. Evol.* 8 (3), 481–494. doi:10.1093/gbe/evw010

Engelman, J. A., Luo, J., and Cantley, L. C. (2006). The Evolution of Phosphatidylinositol 3-kinases as Regulators of Growth and Metabolism. *Nat. Rev. Genet.* 7 (8), 606–619. doi:10.1038/nrg1879

Fan, Y., Zhu, X., Xu, Y., Lu, X., Xu, Y., Wang, M., et al. (2018). Cell-Cycle and DNA-Damage Response Pathway Is Involved in Leptomeningeal Metastasis of Non-small Cell Lung Cancer. *Clin. Cancer Res.* 24 (1), 209–216. doi:10.1158/1078-0432.CCR-17-1582

Fraser, H. B., Moses, A. M., and Schadt, E. E. (2010). Evidence for Widespread Adaptive Evolution of Gene Expression in Budding Yeast. *Proc. Natl. Acad. Sci.* 107 (7), 2977–2982. doi:10.1073/pnas.0912245107

Gamazon, E. R., Huang, R. S., Cox, N. J., and Dolan, M. E. (2010). Chemotherapeutic Drug Susceptibility Associated SNPs Are Enriched in Expression Quantitative Trait Loci. *Proc. Natl. Acad. Sci.* 107 (20), 9287–9292. doi:10.1073/pnas.1001827107

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* 6 (269), pl1. doi:10.1126/scisignal.2004088

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S.-L., et al. (2010). Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. *Plos Genet.* 6 (5), e1000952. doi:10.1371/journal.pgen.1000952

Gu, L., Gao, J., Li, Q., Zhu, Y. P., Jia, C. S., Fu, R. Y., et al. (2008). Rapamycin Reverses NPM-ALK-Induced Glucocorticoid Resistance in Lymphoid Tumor Cells by Inhibiting mTOR Signaling Pathway, Enhancing G1 Cell Cycle Arrest and Apoptosis. *Leukemia* 22 (11), 2091–2096. doi:10.1038/leu.2008.204

Guo, X., Lin, W., Bao, J., Cai, Q., Pan, X., Bai, M., et al. (2018). A Comprehensive Cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *Am. J. Hum. Genet.* 102 (5), 890–903. doi:10.1016/j.ajhg.2018.03.016

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* 14, 7. doi:10.1186/1471-2105-14-7

Hao, X.-l., Han, F., Zhang, N., Chen, H.-q., Jiang, X., Yin, L., et al. (2018). TC2N, a Novel Oncogene, Accelerates Tumor Progression by Suppressing P53 Signaling Pathway in Lung Cancer. *Cell Death Differ* 26, 1235–1250. doi:10.1038/s41418-018-0202-8

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., et al. (2013). Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. *Am. J. Hum. Genet.* 92 (5), 667–680. doi:10.1016/j.ajhg.2013.03.022

Hers, I., Vincent, E. E., and Tavaré, J. M. (2011). Akt Signalling in Health and Disease. *Cell Signal.* 23 (10), 1515–1527. doi:10.1016/j.cellsig.2011.05.004

Ho, Y.-Y., Cope, L. M., and Parmigiani, G. (2014). Modular Network Construction Using eQTL Data: an Analysis of Computational Costs and Benefits. *Front. Genet.* 5, 40. doi:10.3389/fgene.2014.00040

Holloway, B., Luck, S., Beatty, M., Rafalski, J.-A., and Li, B. (2011). Genome-wide Expression Quantitative Trait Loci (eQTL) Analysis in maize. *BMC Genomics* 12, 336. doi:10.1186/1471-2164-12-336

Hu, J., Zeng, T., Xia, Q., Huang, L., Zhang, Y., Zhang, C., et al. (2020). Identification of Key Genes for the Ultrahigh Yield of Rice Using Dynamic Cross-Tissue Network Analysis. *Genomics, Proteomics & Bioinformatics* 18 (3), 256–270. doi:10.1016/j.gpb.2019.11.007

Hu, J., Zeng, T., Xia, Q., Qian, Q., Yang, C., Ding, Y., et al. (2018). Unravelling miRNA Regulation in Yield of rice (Oryza Sativa) Based on Differential Network Model. *Sci. Rep.* 8 (1), 8498. doi:10.1038/s41598-018-26438-w

Huang, Y., Wuchty, S., Ferdig, M. T., and Przytycka, T. M. (2009). Graph Theoretical Approach to Study eQTL: a Case Study of Plasmodium Falciparum. *Bioinformatics* 25 (12), i15–i20. doi:10.1093/bioinformatics/btp189

Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S.-x., Nguyen, Q. T., et al. (2017). Integrated Genome-wide Analysis of Expression Quantitative Trait Loci Aids Interpretation of Genomic Association Studies. *Genome Biol.* 18 (1), 16. doi:10.1186/s13059-016-1142-6

Ju, J. H., Shenoy, S. A., Crystal, R. G., and Mezey, J. G. (2017). An Independent Component Analysis Confounding Factor Correction Framework for Identifying Broad Impact Expression Quantitative Trait Loci. *Plos Comput. Biol.* 13 (5), e1005537. doi:10.1371/journal.pcbi.1005537

Kabakchiev, B., and Silverberg, M. S. (2013). Expression Quantitative Trait Loci Analysis Identifies Associations between Genotype and Gene Expression in Human Intestine. *Gastroenterology* 144 (7), 1488–1496. e1481-1483. doi:10.1053/j.gastro.2013.03.001

Kang, H. P., Yang, X., Chen, R., Zhang, B., Corona, E., Schadt, E. E., et al. (2012). Integration of Disease-specific Single Nucleotide Polymorphisms, Expression Quantitative Trait Loci and Coexpression Networks Reveal Novel Candidate Genes for Type 2 Diabetes. *Diabetologia* 55 (8), 2205–2213. doi:10.1007/s00125-012-2568-3

Kasela, S., Kisand, K., Tserel, L., Kaleviste, E., Remm, A., Fischer, K., et al. (2017). Pathogenic Implications for Autoimmune Mechanisms Derived by Comparative eQTL Analysis of CD4+ versus CD8+ T Cells. *Plos Genet.* 13 (3), e1006643. doi:10.1371/journal.pgen.1006643

Kolberg, L., Kerimov, N., Peterson, H., and Alasoo, K. (2020). Co-expression Analysis Reveals Interpretable Gene Modules Controlled by Trans-acting Genetic Variants. *Elife* 9, e58705. doi:10.7554/eLife.58705

Kugler, K. G., Siegwart, G., Nussbaumer, T., Ametz, C., Spannagl, M., Steiner, B., et al. (2013). Quantitative Trait Loci-dependent Analysis of a Gene Co-expression Network Associated with Fusarium Head Blight Resistance in Bread Wheat (Triticum aestivumL). *BMC Genomics* 14, 728. doi:10.1186/1471-2164-14-728

Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J., and Glass, K. (2019). Estimating Sample-specific Regulatory Networks. *iScience* 14, 226–240. doi:10.1016/j.isci.2019.03.021

Lee, S.-I., Dudley, A. M., Drubin, D., Silver, P. A., Krogan, N. J., Pe'er, D., et al. (2009). Learning a Prior on Regulatory Potential from eQTL Data. *Plos Genet.* 5 (1), e1000358. doi:10.1371/journal.pgen.1000358

Leng, S., Ma, H., Kurths, J., Lai, Y.-C., Lin, W., Aihara, K., et al. (2020). Partial Cross Mapping Eliminates Indirect Causal Influences. *Nat. Commun.* 11 (1), 2632. doi:10.1038/s41467-020-16238-0

Levallet, G., Bergot, E., Antoine, M., Creveuil, C., Santos, A. O., Beau-Faller, M., et al. (2012). High TUBB3 Expression, an Independent Prognostic Marker in Patients with Early Non-small Cell Lung Cancer Treated by Preoperative Chemotherapy, Is Regulated by K-Ras Signaling Pathway. *Mol. Cancer Ther.* 11 (5), 1203–1213. doi:10.1158/1535-7163.MCT-11-0899

Li, J., Wang, L., Jiang, T., Wang, J., Li, X., Liu, X., et al. (2016). eSNPO: An eQTL-Based SNP Ontology and SNP Functional Enrichment Analysis Platform. *Sci. Rep.* 6, 30595. doi:10.1038/srep30595

Liang, L., Morar, N., Dixon, A. L., Lathrop, G. M., Abecasis, G. R., Moffatt, M. F., et al. (2013). A Cross-Platform Analysis of 14,177 Expression Quantitative Trait Loci Derived from Lymphoblastoid Cell Lines. *Genome Res.* 23 (4), 716–726. doi:10.1101/gr.142521.112

Liu, G., Bao, X., and Wang, R. (2015). Expression Quantitative Trait Loci Regulate HNF4A and PTBP1 Expression in Human Brains. *Proc. Natl. Acad. Sci. USA.* 112 (30), E3975. doi:10.1073/pnas.1509048112

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016). Personalized Characterization of Diseases Using Sample-specific Networks. *Nucleic Acids Res.* 44 (22), e164. doi:10.1093/nar/gkw772

Lu, Y., Fang, Z., Li, M., Chen, Q., Zeng, T., Lu, L., et al. (2019). Dynamic Edge-Based Biomarker Non-invasively Predicts Hepatocellular Carcinoma with Hepatitis B Virus Infection for Individual Patients Based on Blood Testing. *J. Mol. Cel Biol.* 11, 665–677. doi:10.1093/jmcb/mjz025

Lynch, J. B., and Hsiao, E. Y. (2019). Microbiomes as Sources of Emergent Host Phenotypes. *Science* 365 (6460), 1405–1409. doi:10.1126/science.aay0240

MacEwan, D. J. (2002). TNF Receptor Subtype Signalling: Differences and Cellular Consequences. *Cell Signal.* 14 (6), 477–492. doi:10.1016/s0898-6568(01)00262-5

Mason, V. C., Schaefer, R. J., McCue, M. E., Leeb, T., and Gerber, V. (2018). eQTL Discovery and Their Association with Severe Equine Asthma in European Warmblood Horses. *BMC Genomics* 19 (1), 581. doi:10.1186/s12864-018-4938-9

Michaelson, J. J., Loguercio, S., and Beyer, A. (2009). Detection and Interpretation of Expression Quantitative Trait Loci (eQTL). *Methods* 48 (3), 265–276. doi:10.1016/j.ymeth.2009.03.004

Micke, P., Edlund, K., Holmberg, L., Kultima, H. G., Mansouri, L., Ekman, S., et al. (2011). Gene Copy Number Aberrations Are Associated with Survival in Histologic Subgroups of Non-small Cell Lung Cancer. *J. Thorac. Oncol.* 6 (11), 1833–1840. doi:10.1097/JTO.0b013e3182295917

Micke, P., Ohshima, M., Tahmasebpoor, S., Ren, Z.-P., Östman, A., Pontén, F., et al. (2006). Biobanking of Fresh Frozen Tissue: RNA Is Stable in Nonfixed Surgical Specimens. *Lab. Invest.* 86 (2), 202–211. doi:10.1038/labinvest.3700372

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Nguyen, D.-Q., Webber, C., and Ponting, C. P. (2006). Bias of Selection on Human Copy-Number Variants. *Plos Genet.* 2 (2), e20. doi:10.1371/journal.pgen.0020020

O'Brien, H. E., Hannon, E., Hill, M. J., Toste, C. C., Robertson, M. J., Morgan, J. E., et al. (2018). Expression Quantitative Trait Loci in the Developing Human Brain and Their Enrichment in Neuropsychiatric Disorders. *Genome Biol.* 19 (1), 194. doi:10.1186/s13059-018-1567-1

Peters, J. E., Lyons, P. A., Lee, J. C., Richard, A. C., Fortune, M. D., Newcombe, P. J., et al. (2016). Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *Plos Genet.* 12 (3), e1005908. doi:10.1371/journal.pgen.1005908

Peterson, C. B., Service, S. K., Jasinska, A. J., Gao, F., Zelaya, I., Teshiba, T. M., et al. (2016). Characterization of Expression Quantitative Trait Loci in Pedigrees from Colombia and Costa Rica Ascertained for Bipolar Disorder. *Plos Genet.* 12 (5), e1006046. doi:10.1371/journal.pgen.1006046

Platig, J., Castaldi, P. J., DeMeo, D., and Quackenbush, J. (2016). Bipartite Community Structure of eQTLs. *Plos Comput. Biol.* 12 (9), e1005033. doi:10.1371/journal.pcbi.1005033

Ramsay, A. J., Martínez-Trillos, A., Jares, P., Rodríguez, D., Kwarciak, A., and Quesada, V. (2013). Next-generation Sequencing Reveals the Secrets of the Chronic Lymphocytic Leukemia Genome. *Clin. Transl Oncol.* 15 (1), 3–8. doi:10.1007/s12094-012-0922-z

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: a Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update). *Nucleic Acids Res.* 47 (W1), W191–W198. doi:10.1093/nar/gkz369

Reith, W., LeibundGut-Landmann, S., and Waldburger, J.-M. (2005). Regulation of MHC Class II Gene Expression by the Class II Transactivator. *Nat. Rev. Immunol.* 5 (10), 793–806. doi:10.1038/nri1708

Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., McDowell, I. C., et al. (2017). Co-expression Networks Reveal the Tissue-specific Regulation of Transcription and Splicing. *Genome Res.* 27 (11), 1843–1858. doi:10.1101/gr.216721.116

Sanchez-Cespedes, M. (2003). Dissecting the Genetic Alterations Involved in Lung Carcinogenesis. *Lung Cancer* 40 (2), 111–121. doi:10.1016/s0169-5002(03)00033-3

Shabalin, A. A. (2012). Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations. *Bioinformatics* 28 (10), 1353–1358. doi:10.1093/bioinformatics/bts163

Son, H.-Y., Hwangbo, Y., Yoo, S.-K., Im, S.-W., Yang, S. D., Kwak, S.-J., et al. (2017). Genome-wide Association and Expression Quantitative Trait Loci Studies Identify Multiple Susceptibility Loci for Thyroid Cancer. *Nat. Commun.* 8, 15966. doi:10.1038/ncomms15966

Stassi, G., and De Maria, R. (2002). Autoimmune Thyroid Disease: New Models of Cell Death in Autoimmunity. *Nat. Rev. Immunol.* 2 (3), 195–204. doi:10.1038/nri750

Strober, B. J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A., et al. (2019). Dynamic Genetic Regulation of Gene Expression during Cellular Differentiation. *Science* 364 (6447), 1287–1290. doi:10.1126/science.aaw0040

Strunz, T., Grassmann, F., Gayán, J., Nahkuri, S., Souza-Costa, D., Maugeais, C., et al. (2018). A Mega-Analysis of Expression Quantitative Trait Loci (eQTL) Provides Insight into the Regulatory Architecture of Gene Expression Variation in Liver. *Sci. Rep.* 8 (1), 5865. doi:10.1038/s41598-018-24219-z

Sun, W., Yu, T., and Li, K.-C. (2007). Detection of eQTL Modules Mediated by Activity Levels of Transcription Factors. *Bioinformatics* 23 (17), 2290–2297. doi:10.1093/bioinformatics/btm327

Suthram, S., Beyer, A., Karp, R. M., Eldar, Y., and Ideker, T. (2008). eQED: an Efficient Method for Interpreting eQTL Associations Using Protein Networks. *Mol. Syst. Biol.* 4, 162. doi:10.1038/msb.2008.4

Takayama, T., Miyanishi, K., Hayashi, T., Sato, Y., and Niitsu, Y. (2006). Colorectal Cancer: Genetics of Development and Metastasis. *J. Gastroenterol.* 41 (3), 185–192. doi:10.1007/s00535-006-1801-6

Tanaka, Y., Tamada, Y., Ikeguchi, M., Yamashita, F., and Okuno, Y. (2020). System-Based Differential Gene Network Analysis for Characterizing a Sample-specific Subnetwork. *Biomolecules* 10 (2), 306. doi:10.3390/biom10020306

Tang, Z., Yu, W., Zhang, C., Zhao, S., Yu, Z., Xiao, X., et al. (2016). CREB-binding Protein Regulates Lung Cancer Growth by Targeting MAPK and CPSF4 Signaling Pathway. *Mol. Oncol.* 10 (2), 317–329. doi:10.1016/j.molonc.2015.10.015

Thibodeau, S. N., French, A. J., McDonnell, S. K., Cheville, J., Middha, S., Tillmans, L., et al. (2015). Identification of Candidate Genes for Prostate Cancer-Risk SNPs Utilizing a normal Prostate Tissue eQTL Data Set. *Nat. Commun.* 6, 8653. doi:10.1038/ncomms9653

Tripathi, S. C., Fahrmann, J. F., Celiktas, M., Aguilar, M., Marini, K. D., Jolly, M. K., et al. (2017). MCAM Mediates Chemoresistance in Small-Cell Lung Cancer via

the PI3K/AKT/SOX2 Signaling Pathway. *Cancer Res.* 77 (16), 4414–4425. doi:10.1158/0008-5472.CAN-16-2874

Tsatsoulis, A. (2006). The Role of Stress in the Clinical Expression of Thyroid Autoimmunity. *Ann. N.Y Acad. Sci.* 1088, 382–395. doi:10.1196/annals.1366.015

van der Wijst, M. G. P., Brugge, H., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., et al. (2018). Single-cell RNA Sequencing Identifies Celltype-specific Cis-eQTLs and Co-expression QTLs. *Nat. Genet.* 50 (4), 493–497. doi:10.1038/s41588-018-0089-9

Verbeke, L. P. C., Cloots, L., Demeester, P., Fostier, J., and Marchal, K. (2013). EPSILON: an eQTL Prioritization Framework Using Similarity Measures Derived from Local Networks. *Bioinformatics* 29 (10), 1308–1316. doi:10.1093/bioinformatics/btt142

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., et al. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* 182 (5), 1214–e11. e1211. doi:10.1016/j.cell.2020.08.008

Wang, W., and Zhang, X. (2011). Network-based Group Variable Selection for Detecting Expression Quantitative Trait Loci (eQTL). *BMC Bioinformatics* 12, 269. doi:10.1186/1471-2105-12-269

Westra, H.-J., Arends, D., Esko, T., Peters, M. J., Schurmann, C., Schramm, K., et al. (2015). Cell Specific eQTL Analysis without Sorting Cells. *Plos Genet.* 11 (5), e1005223. doi:10.1371/journal.pgen.1005223

Wu, C., Delano, D. L., Mitro, N., Su, S. V., Janes, J., McClurg, P., et al. (2008). Gene Set Enrichment in eQTL Data Identifies Novel Annotations and Pathway Regulators. *Plos Genet.* 4 (5), e1000070. doi:10.1371/journal.pgen.1000070

Ye, Y., Zhang, Z., Liu, Y., Diao, L., and Han, L. (2020). A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine. *Trends Genet.* 36 (5), 318–336. doi:10.1016/j.tig.2020.01.009

Young, A. I., Benonisdottir, S., Przeworski, M., and Kong, A. (2019). Deconstructing the Sources of Genotype-Phenotype Associations in Humans. *Science* 365 (6460), 1396–1400. doi:10.1126/science.aax3710

Yu, X., Zeng, T., Wang, X., Li, G., and Chen, L. (2015). Unravelling Personalized Dysfunctional Gene Network of Complex Diseases Based on Differential Network Model. *J. Transl Med.* 13, 189. doi:10.1186/s12967-015-0546-5

Yu, X., Zhang, J., Sun, S., Zhou, X., Zeng, T., and Chen, L. (2017). Individual-specific Edge-Network Analysis for Disease Prediction. *Nucleic Acids Res.* 45 (20), e170. doi:10.1093/nar/gkx787

Yuan, L., Zhu, L., Guo, W.-L., Zhou, X., Zhang, Y., Huang, Z., et al. (2017). Nonconvex Penalty Based Low-Rank Representation and Sparse Regression for eQTL Mapping. *Ieee/acm Trans. Comput. Biol. Bioinf.* 14 (5), 1154–1164. doi:10.1109/TCBB.2016.2609420

Zeggini, E., Gloyn, A. L., Barton, A. C., and Wain, L. V. (2019). Translational Genomics and Precision Medicine: Moving from the Lab to the Clinic. *Science* 365 (6460), 1409–1413. doi:10.1126/science.aax4588

Zeng, T., Zhang, W., Yu, X., Liu, X., Li, M., Liu, R., et al. (2014). Edge Biomarkers for Classification and Prediction of Phenotypes. *Sci. China Life Sci.* 57 (11), 1103–1114. doi:10.1007/s11427-014-4757-4

Zhang, L., and Kim, S. (2014). Learning Gene Networks under SNP Perturbations Using eQTL Datasets. *Plos Comput. Biol.* 10 (2), e1003420. doi:10.1371/journal.pcbi.1003420

Zhang, T., Choi, J., Kovacs, M. A., Shi, J., Xu, M., Goldstein, A. M., et al. (2018). Cell-type-specific eQTL of Primary Melanocytes Facilitates Identification of Melanoma Susceptibility Genes. *Genome Res.* 28 (11), 1621–1635. doi:10.1101/gr.233304.117

Zhang, W., Zeng, T., and Chen, L. (2014). EdgeMarker: Identifying Differentially Correlated Molecule Pairs as Edge-Biomarkers. *J. Theor. Biol.* 362, 35–43. doi:10.1016/j.jtbi.2014.05.041

Zhang, W., Zeng, T., Liu, X., and Chen, L. (2015). Diagnosing Phenotypes of Single-Sample Individuals by Edge Biomarkers. *J. Mol. Cel Biol.* 7 (3), 231–241. doi:10.1093/jmcb/mjv025

Zheng, G., Shen, Z., Chen, H., Liu, J., Jiang, K., Fan, L., et al. (2017). Metapristone Suppresses Non-small Cell Lung Cancer Proliferation and Metastasis via Modulating RAS/RAF/MEK/MAPK Signaling Pathway. *Biomed. Pharmacother.* 90, 437–445. doi:10.1016/j.biopha.2017.03.091

Zhernakova, D. V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., et al. (2017). Identification of Context-dependent Expression Quantitative Trait Loci in Whole Blood. *Nat. Genet.* 49 (1), 139–145. doi:10.1038/ng.3737

Zhou, Y., Li, S., Li, J., Wang, D., and Li, Q. (2017). Effect of microRNA-135a on Cell Proliferation, Migration, Invasion, Apoptosis and Tumor Angiogenesis through the IGF-1/PI3K/Akt Signaling Pathway in Non-small Cell Lung Cancer. *Cell Physiol Biochem.* 42 (4), 1431–1446. doi:10.1159/000479207

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets. *Nat. Genet.* 48 (5), 481–487. doi:10.1038/ng.3538

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Yuan, Zeng and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 31 January 2022
doi: 10.3389/fcell.2022.817643

Identification of Ferroptosis-Related Biomarkers for Prognosis and Immunotherapy in Patients With Glioma

Junfeng Shi[1,2†], *Donglin Lai*[3,4†], *Xiaojia Zuo*[3†], *Dingsheng Liu*[3], *Bing Chen*[5], *Yanjun Zheng*[3*], *Changlian Lu*[3*] and *Xuefeng Gu*[3,4,6*]

[1]Department of Prosthodontics, Shanghai Ninth People's Hospital, College of Stomatology, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [2]Shanghai Engineering Research Center of Advanced Dental Technology and Materials, National Clinical Research Center for Oral Diseases, Shanghai, China, [3]Shanghai Key Laboratory of Molecular Imaging, Zhoupu Hospital, Shanghai University of Medicine & Health Sciences, Shanghai, China, [4]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China, [5]Department of Neurosurgery, Affiliated Hospital of Guangdong Medical University, Guangzhou, China, [6]School of Pharmacy, Shanghai University of Medicine & Health Sciences, Shanghai, China

Ferroptosis is a novel type of iron- and ROS-dependent cell death and is involved in various diseases. LncRNAs are involved and play important roles in the occurrence and development of several cancers. However, researches about the role of ferroptosis-related lncRNAs in glioma are relatively rare. Here, we identified nine ferroptosis-related lncRNAs and then constructed a prognostic model by the LASSO and Cox analysis. The model could predict overall survival with high sensitivity and specificity according to ROC curves. In addition, the cell cycle, p53 signaling, apoptosis, and oxidative phosphorylation pathways were obviously enriched in the pathogenesis of glioma by gene set enrichment analysis. A nomogram was constructed by integrating several independent prognostic clinicopathological features, and it could provide a valuable predictive tool for overall survival. Furthermore, a strong correlation between these nine lncRNAs and immunotherapy was found. Glioma patients in the high-risk group had higher TMB using somatic mutation data, different immune infiltration, and higher expression of immune checkpoints, indicating these patients might benefit from immune checkpoint inhibitor therapy. In summary, these nine ferroptosis-related lncRNAs were promising biomarkers for predicting overall survival and guiding immunotherapy or future immune checkpoint inhibitor development for glioma patients.

Keywords: ferroptosis, lncRNA, biomarker, prognosis, immune checkpoint, immunotherapy, glioma

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Lei Chen,
Shanghai Maritime University, China
Qi Sun,
Shenzhen University, China

****Correspondence:***
Xuefeng Gu
guxf@sumhs.edu.cn
Changlian Lu
lvcl@sumhs.edu.cn
Yanjun Zheng
zhengyj@sumhs.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *18 November 2021*
Accepted: *03 January 2022*
Published: *31 January 2022*

Citation:
Shi J, Lai D, Zuo X, Liu D, Chen B, Zheng Y, Lu C and Gu X (2022) Identification of Ferroptosis-Related Biomarkers for Prognosis and Immunotherapy in Patients With Glioma. Front. Cell Dev. Biol. 10:817643. doi: 10.3389/fcell.2022.817643

Abbreviations: BP, biological process; CC, cellular component; ceRNAs, competitive endogenous RNAs; CGGA, the Chinese Glioma Genome Atlas; C-index, the concordance index; DEmRNAs, differentially expressed mRNAs; GBM, glioblastoma; GO, gene ontology; GSEA, gene set enrichment analysis; HGG, high-grade glioma; ICIs, immune checkpoint inhibitors; KEGG, Kyoto Encyclopedia of Genes and Genomes; LASSO, least absolute shrinkage and selection operator; LGG, low-grade glioma; lncRNAs, long noncoding RNAs; MF, molecular function; OS, overall survival; ROS, reactive oxygen species; ssGSEA, single-sample GSEA; TCGA, The Cancer Genome Atlas; TIMER, Tumor immune estimation resource; TMB, tumor mutational burden; WHO, the World Health Organization.

INTRODUCTION

Glioma is the most common primary malignant tumor of the central nervous system in adults, accounting for approximately 80% of cases (Stupp et al., 2015; Ostrom et al., 2020). According to the 2016 classification standards of the World Health Organization (WHO), the pathological types of glioma can be divided into low-grade glioma (LGG, grade 1–2) and high-grade glioma (HGG, grade 3–4) (Louis et al., 2016). LGG has high differentiation and a good prognosis, and the median overall survival (OS) is 8–10 years. However, HGG has low differentiation, high malignancy, and poor prognosis. Among the types of HGG, the median OS of anaplastic glioma (WHO grade 3) is 3–4 years, and the prognosis of glioblastoma (GBM, WHO grade 4) is the worst, with a median OS of only 14.6–17 months (Smoll et al., 2013; Stupp et al., 2015; Gu et al., 2019; Litak et al., 2019). Glioma is prone to immune infiltration and recurrence after surgical resection, and there are limited therapeutic options to date. Patients with glioma suffer from a high mortality rate and poor quality of life. In recent years, it has been reported that immune checkpoint inhibitors (ICIs, i.e., PD-L1 inhibitors) have a certain effect in the treatment of glioma (Ampudia-Mesias et al., 2021).

Ferroptosis is a novel form of programmed cell death that is different from apoptosis and autophagy; ferroptosis mainly occurs through Fe^{2+} or lipoxygenases, which catalyse the lipid peroxidation of unsaturated fatty acids that are highly expressed on the cell membrane, thereby inducing cell death (Dixon 2017; Woo et al., 2020; Tang et al., 2021). The level of reactive oxygen species (ROS) in tumor cells is higher than that in normal cells, and excessive accumulation of ROS activates the apoptotic pathway and promotes the death of tumor cells (Wang et al., 2016; Zhao et al., 2019). Iron overload can result in ferroptosis, which can be activated in cancer cells to fight cancer (Brown et al., 2020). And inhibiting ferroptosis results in decreased chemosensitivity (Zhang et al., 2020). The application of ferroptosis inducers can enhance tumor sensitivity to chemotherapy and radiotherapy, providing a more promising therapeutic strategy for killing drug-resistant cancer cells (Hassannia et al., 2019; Zhao et al., 2020). The most interesting finding was that the ferroptosis inducer erastin can enhance the sensitivity of GBM cells to temozolomide (Chen et al., 2015).

With the improvement of gene sequencing technology, long noncoding RNAs (lncRNAs) have attracted increasing attention because they can regulate gene expression at multiple levels, such as the chromatin, transcription, and posttranscriptional levels, and they can participate in various biological processes, such as cell differentiation, cell cycle regulation, and stem cell pluripotency maintenance (Lu Q. et al., 2020; Zhuo et al., 2020; Jiang, et al., 2021). In recent years, multiple studies have confirmed that regulation of the expression of lncRNAs (i.e., *LINC00618, LINC00336, PVT1*, and *ZFAS1*) is closely related to ferroptosis (Wang et al., 2019; Lu J. et al., 2020; Yang et al., 2020; Wang et al., 2021; Yao, et al., 2021; Zhang et al., 2021). Some lncRNAs (i.e., *LINC00336* and *ZFAS1*) also act as competitive endogenous RNAs (ceRNAs) to prevent

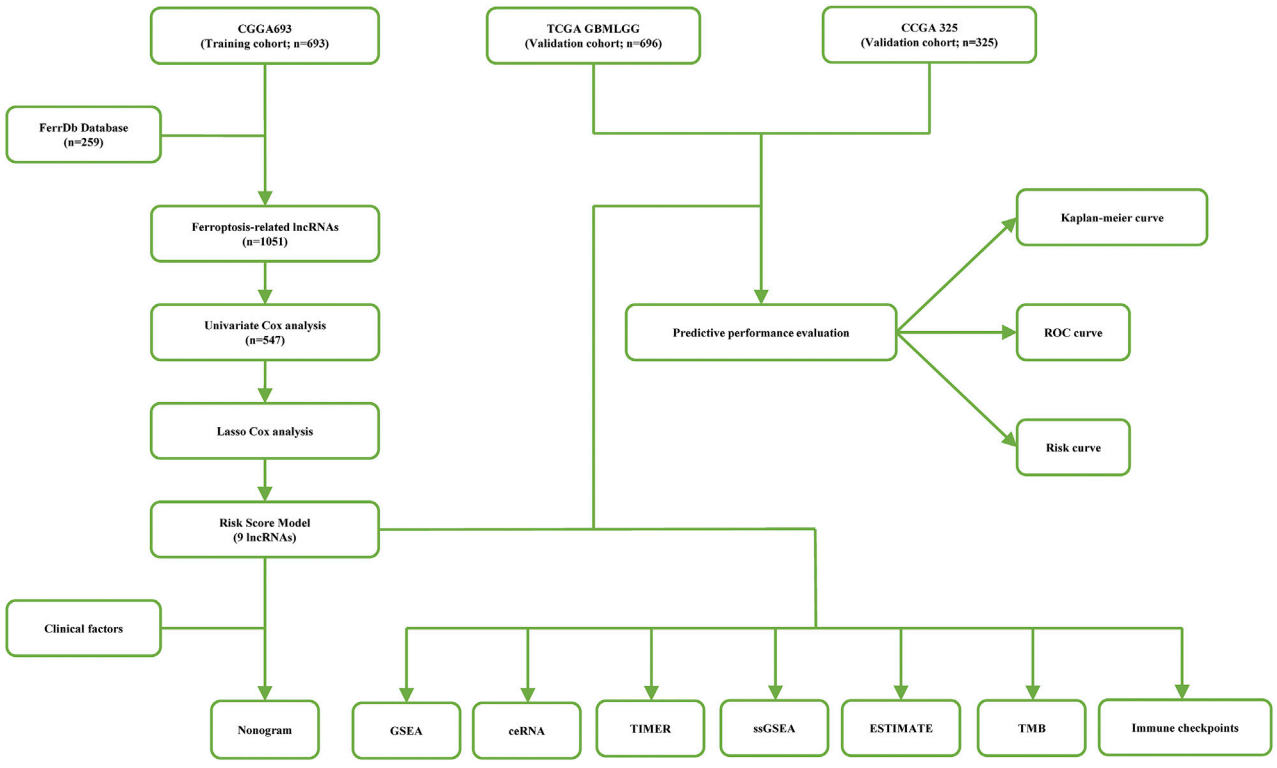


FIGURE 1 | A flow chart of data analysis methods and processing in this study.

TABLE 1 | The clinical characteristics of glioma patients in CGGA.

Characteristics	Training	Validation 1
	CGGA693 (*n* = 693)	CGGA325 (*n* = 325)
PRS type		
Primary	422	229
Recurrent	271	62
Secondary	0	30
Grade		
WHO II	188	103
WHO III	255	79
WHO IV	249	139
Gender		
Female	295	122
Male	398	203
Age		
≥ 60	83	33
<60	609	292
Follow-up state		
Alive	266	96
Dead	397	220
Radiotherapy		
Untreated	136	66
Treated	510	244
Chemotherapy		
Untreated	161	111
Treated	486	193
IDH mutation status		
Mutant	356	175
Wildtype	286	149
1p19q codeletion status		
Non-codel	478	250
Codel	145	67
MGMTp methylation status		
Methylated	315	157
Un-methylated	227	149

TABLE 2 | The clinical characteristics of glioma patients in TCGA.

Characteristics	Validation 2
	TCGA (*n* = 696)
Type	
GBM	168
LGG	528
Follow-up state	
Alive	420
Dead	273
Gender	
Female	298
Male	398
Age	
≥ 60	158
<60	538

peroxidation, thereby inhibiting ferroptosis (Wang et al., 2019; Yang et al., 2020).

Nevertheless, researches about the potential mechanism of ferroptosis-related lncRNAs in glioma are relatively rare. Herein, nine ferroptosis-related lncRNAs (*AC062021.1, FAM66C, MIR497HG, TMEM72-AS1, AC010729.2, FAM225B, HOXA-AS2, LINC00662, and LINC00665*) were identified to construct a risk model in patients with glioma. This risk model aimed to reveal the potential roles of these biomarkers in the prognosis and treatment prospects of glioma patients and to further explore the relationships among tumor mutational burden (TMB), immune checkpoints, and ferroptosis-related lncRNAs. In particular, the extra application of these biomarkers in glioma patients during treatment with immune checkpoint inhibitors (ICIs) was also explored.

MATERIALS AND METHODS

Data Acquisition

A flow chart of the data analysis and processing methods used in this study is shown in **Figure 1**. Patient data were collected from the Chinese Glioma Genome Atlas (CGGA) (Zhao, et al., 2021) and The Cancer Genome Atlas (TCGA). After excluding samples without complete clinical data, a total of 1714 samples were obtained for final analysis, including the training CGGA693 cohort (n = 693), the validation CGGA325 cohort (n = 325), and the validation TCGA cohort (LGG, n = 528; GBM, n = 168). Two hundred fifty-nine ferroptosis-related genes were obtained from the FerrDb database (**Supplementary Table S1**) (Zhou and Bao 2020).

Construction of a Prognostic Model

The 259 ferroptosis-related genes and lncRNAs were analyzed in the training CGGA693 cohort by Spearman analysis, and thus, 1051 ferroptosis-related lncRNAs were obtained under the "absolute value of correlation coefficient≥0.3, $p < 0.05$" screening condition. Next, 547 prognosis-related lncRNAs were acquired by univariate Cox regression analysis with the "$p < 0.05$"; screening condition. Then, the least absolute shrinkage and selection operator (LASSO) regression analysis was performed for variable screening and dimension reduction to build a simplified and accurate model. The best lambda and corresponding variables were usually obtained with the least mean square. Lambda. min was 0.09938703, and 24 ferroptosis-related lncRNAs were identified. Multivariate Cox regression analysis was further performed, and 9 ferroptosis-related lncRNAs were ultimately identified to establish a more stable prognostic risk model. The risk score was calculated with the regression coefficients and expression of genes. A Sankey diagram was constructed to visualize the relationship between the 259 ferroptosis-related genes and lncRNAs. According to the expression levels of ferroptosis-related lncRNAs, glioma patients were divided into high- and low-expression groups in the CGGA693 cohort, and then, Kaplan-Meier curves of the two groups were constructed to further explore the value of these 9 ferroptosis-related lncRNA prognostic biomarkers.

Assessment and Validation

According to the median risk score, glioma patients in the training CGGA693 cohort were divided into high- and low-risk groups. Kaplan-Meier analysis and the long-rank test were applied to identify whether the OS rates of the two risk groups

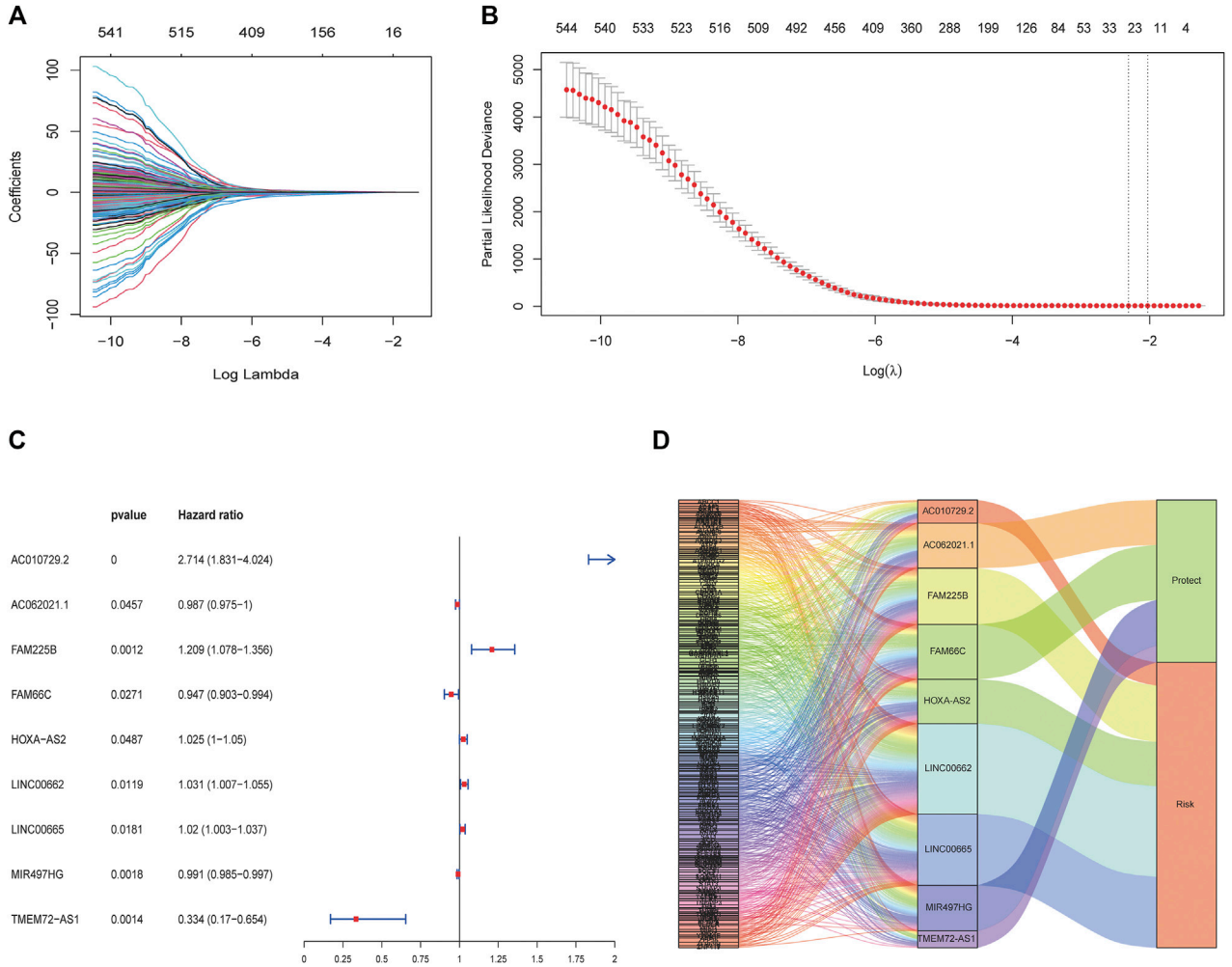


FIGURE 2 | Construction of the ferroptosis-related lncRNA prognostic biomarkers. **(A, B)** LASSO regression model. The red dots indicate the partial likelihood of deviance values, the grey lines indicate the standard error (SE), and the two vertical dashed lines on the left and right indicate optimal values by minimum criteria and 1-SE criteria, respectively. **(C)** Multivariate Cox regression analysis revealed that 9 ferroptosis-related lncRNAs were independent prognostic factors for glioma patients. **(D)** Sankey diagram shows the relationship among ferroptosis-related genes, ferroptosis-related lncRNAs, and risk types.

were significantly different. ROC curves (AUCs for 1-, 3-, and 5-years OS) were used to assess the specificity and sensitivity of the predictive model, which could be analyzed by using the "survivalROC" R package. A risk score map, survival status distribution map, and expression heatmap were plotted to display the distribution of patients and the expression of biomarkers between the low- and high-risk groups. To verify whether the risk model was reliable for predicting OS, the same methods were performed in two external cohorts (CGGA325 and TCGA).

Association of Prognostic Biomarkers and Clinicopathologic Features

To explore the relationship between ferroptosis-related prognostic lncRNAs and clinicopathologic factors, the expression of prognostic biomarkers among subgroups with different statuses was analyzed, including risk stratification, WHO grade, primary or recurrent type, IDH mutation status, and 1p19q codeletion status. These statistical analyses were applied to both the training and validation cohorts.

Functional Enrichment Analysis

To elucidate the underlying biological mechanism of the present results, gene set enrichment analysis (GSEA) was performed in GSEA V4.1.0 based on the Molecular Signatures Database v7.4 (Subramanian et al., 2005). To explore the regulatory mechanism of the screened candidate lncRNAs, the relationships among miRNAs, mRNAs, and lncRNAs were evaluated, and a potential ceRNA network in the regulation of glioma was constructed. These 9 biomarkers were input into the DIANA tool LncBase v.2 to predict the corresponding miRNAs (Paraskevopoulou et al., 2016). Then, mRNAs targeted by

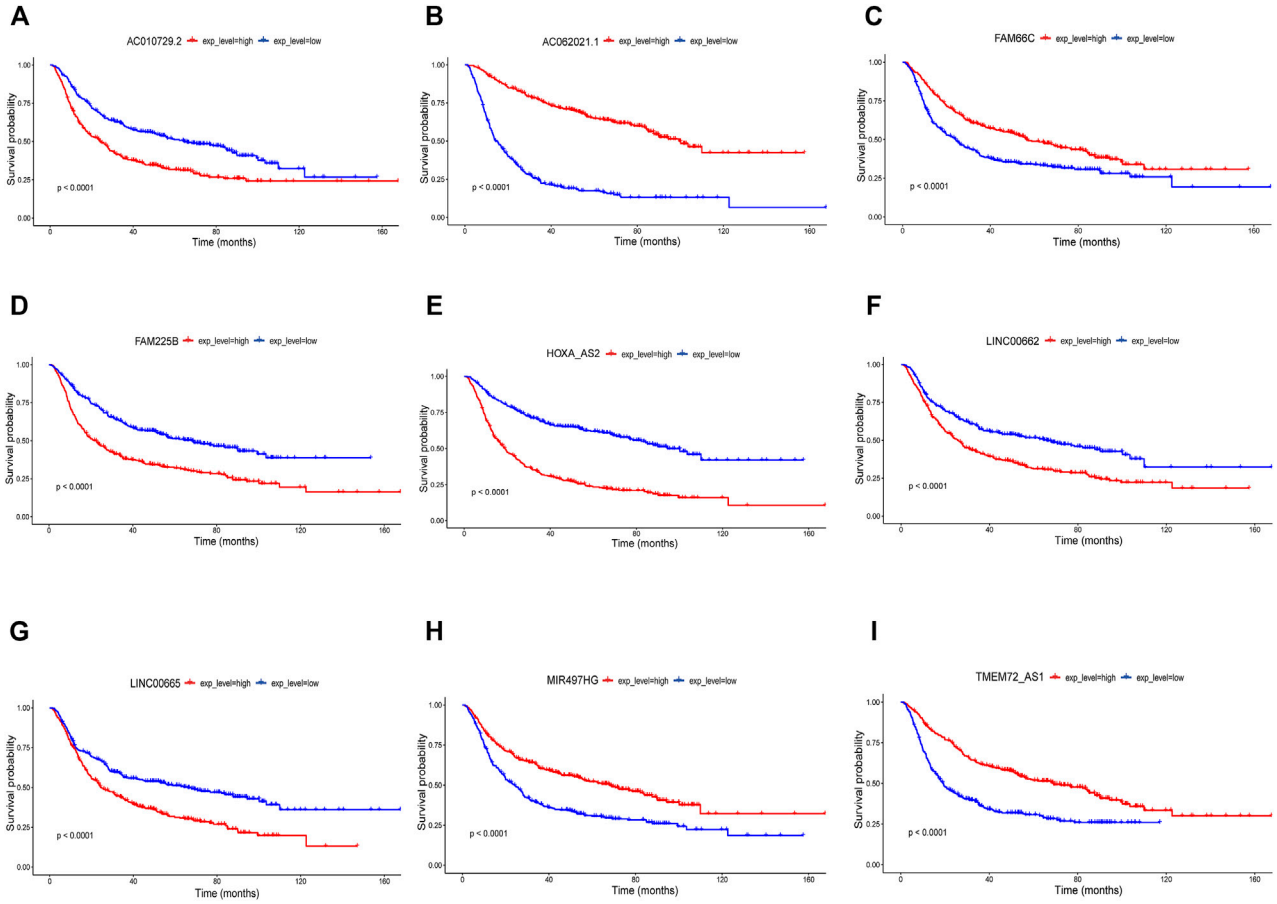


FIGURE 3 | Nine biomarkers in Kaplan-Meier analysis in the training CGGA693 cohort **(A)** AC010729.2, **(B)** AC062021.1, **(C)** FAM66C, **(D)** FAM225B, **(E)** HOXA-AS2, **(F)** LINC00662, **(G)** LINC00665, **(H)** MIR497HG and **(I)** TMEM72-AS1.

miRNAs were predicted with the condition of no less than three target-predicting programs in the Encyclopedia of RNA Interactomes (Li, et al., 2014). The differentially expressed mRNAs (DEmRNAs) between the high- and low-risk groups were identified by "limma" R. The final mRNAs were identified by integrating predictive mRNAs and DEmRNAs. The ceRNA network was constructed and visualized by Cytoscape (version: v3.7.1) (Shannon et al., 2003). Afterward, the functional enrichment results of these mRNAs in the ceRNA network were obtained by gene ontology (GO), including the biological process (BP), cellular component (CC), molecular function (MF), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis via the "clusterProfiler" R package. The p cut-off-value and q cut-off-value were both 0.05.

Nomogram Construction and Assessment

Univariate and multivariate Cox regression analyses of clinicopathological features were used to identify independent prognostic factors. Then, a nomogram was constructed based on these independent prognostic factors. This nomogram could calculate the OS rates of glioma patients at 1, 3, and 5 years, which provides valuable suggestions for clinicians to judge the OS of glioma patients. In addition, the concordance index (C-index) was calculated to evaluate the predictive ability of a nomogram, and calibration curves were used to evaluate the accuracy of the nomogram. The higher the C-index is, the more accurate the prediction model. Calibration curves, including 1-, 3-, and 5-years curves, were drawn in the "rms" R package to assess whether there was a good match between the actual probability of OS and the predicted probability.

Immune Infiltration

The association between these ferroptosis biomarkers and six types of immune cells (B cells, $CD8^+$ T cells, $CD4^+$ T cells, neutrophils, macrophages, and dendritic cells) was identified with the TIMER database (Li T. et al., 2020). Pearson's correlation coefficients and estimated *p* values were calculated. Subsequently, single-sample GSEA (ssGSEA) was performed to assess the immune response in the two risk groups using the "gsva" R package. Additionally, the ssGSEA scores of 13 immune cell infiltrate and 16 immune-related functions were calculated in the two risk groups and

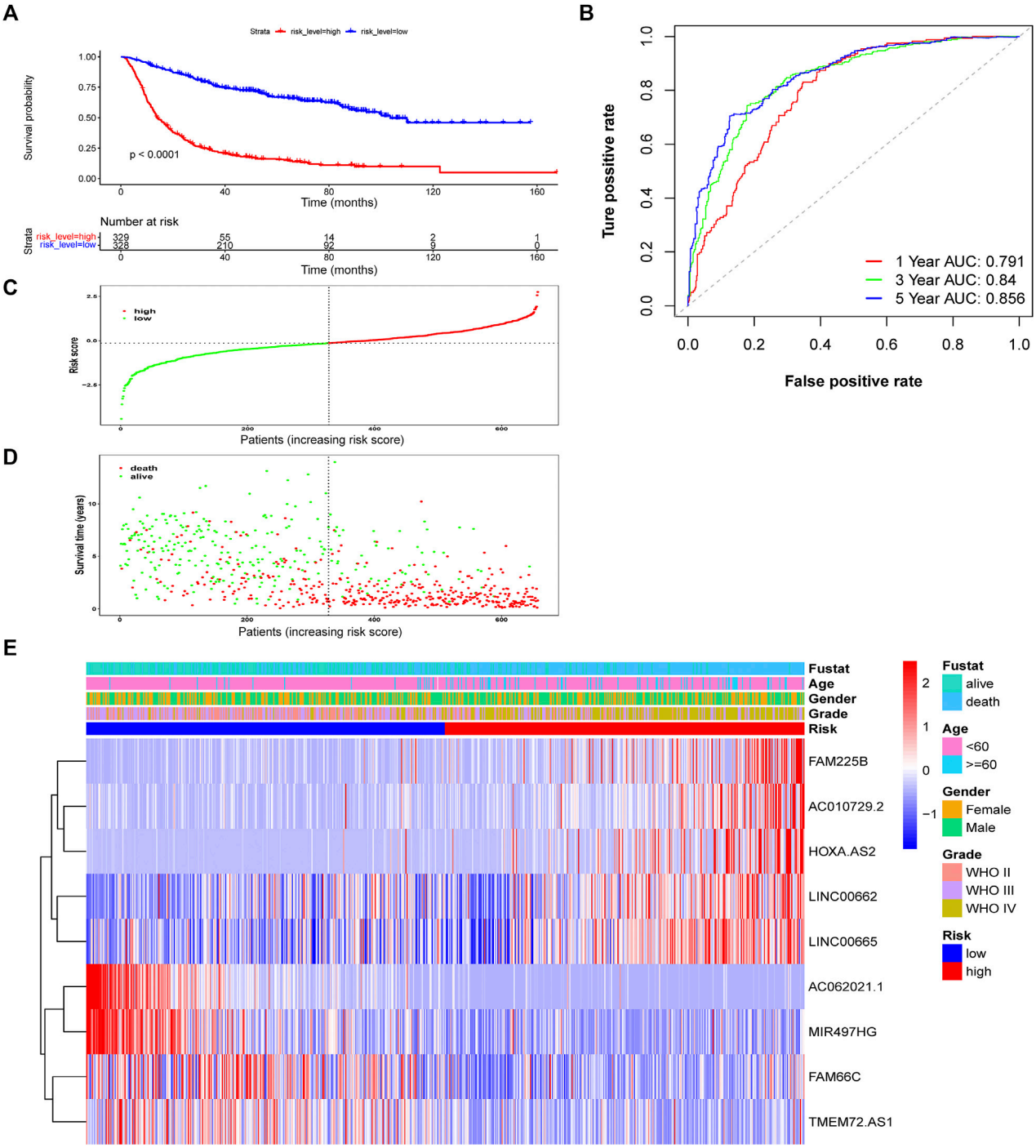


FIGURE 4 | Assessment of the risk score model based on ferroptosis-related prognostic biomarkers in the training CGGA693 cohort. **(A)** Kaplan-Meier survival based on risk level. **(B)** ROC curves for predicting OS of glioma patients based on the risk score. **(C–E)** The risk score, survival distribution of patients with increased risk scores, and expression heatmap of ferroptosis-related lncRNAs based on risk level.

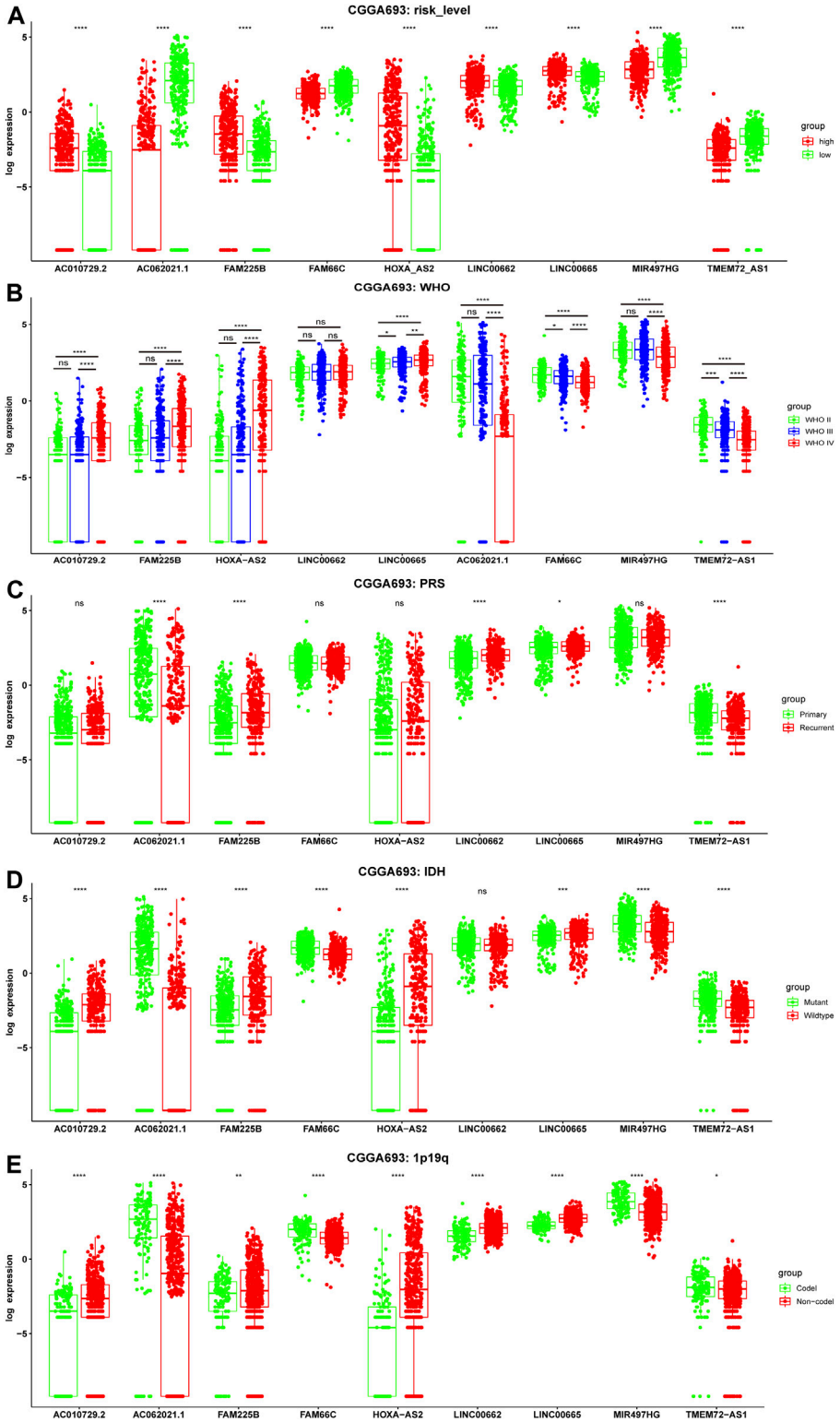


FIGURE 5 | Correlation between the expression level of nine ferroptosis-related lncRNAs and clinicopathological features in the training CGGA693 cohort. **(A–E)** Risk level, WHO grade, PRS type, IDH mutation status, and 1p19q codeletion status. WHO: World Health Organization; P: Primary, R: Recurrent, S: Secondary Recurrent; ns: Not significant, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

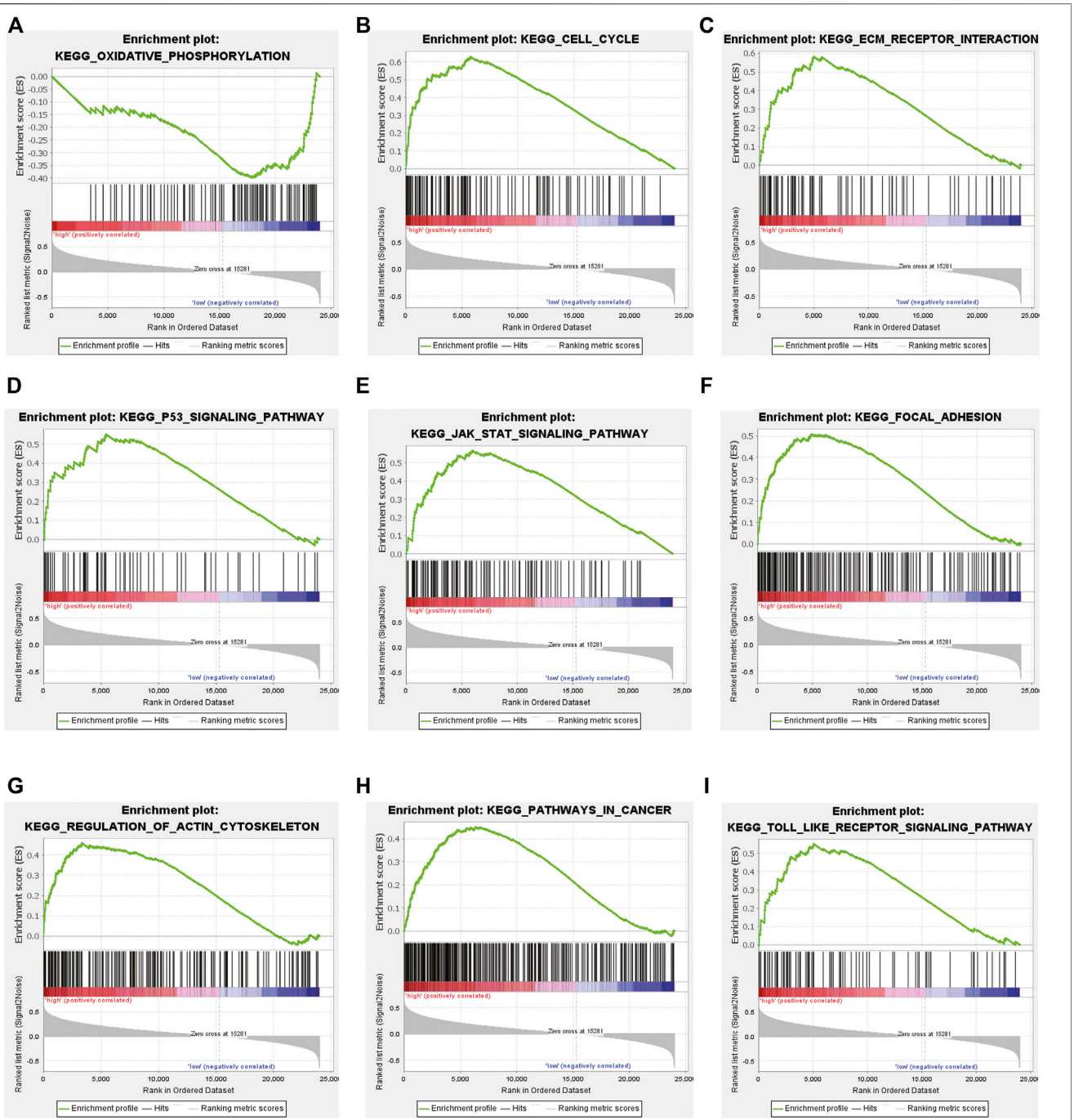


FIGURE 6 | Gene set enrichment analysis (GSEA) in the training CGGA693 cohort. **(A)** GSEA suggested notable enrichment of pathways in the low-risk group and **(B–I)** GSEA suggested significant enrichment in the high-risk group.

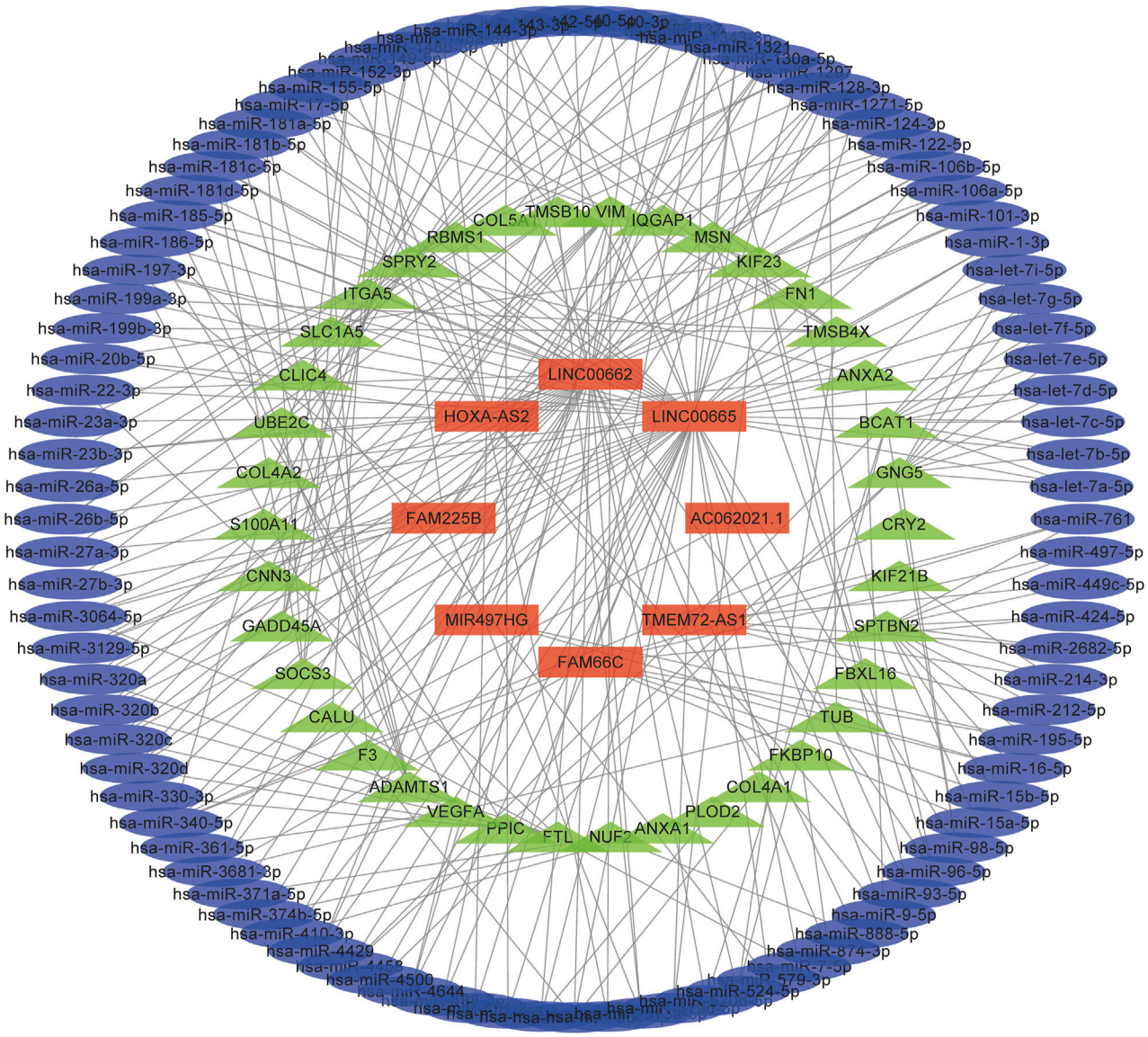


FIGURE 7 | A ceRNA network in the regulation of glioma. Light blue, light green, and red colours represent miRNAs, mRNAs, and lncRNAs, respectively.

displayed as a heatmap and boxplots. The immune difference between the two risk groups was depicted in the heatmap and violin plots. In addition, stromal scores and immune scores for two risk groups were calculated and compared according to the ESTIMATE algorithm (Yoshihara, et al., 2013).

Tumor Mutational Burden and Immune Checkpoints

Tumor cells with many mutations were far different from normal cells and thus more easily found by the immune system. Thus, patients with higher TMB theoretically benefit more from immunotherapy. Mutation data acquired by Mutect software were downloaded from the TCGA database, and TMB was calculated using R software. TMB values were analyzed between high- and low-risk groups in the TCGA cohort. Based on the median TMB value, glioma patients were divided into low- and high-TMB groups. In the two TMB groups, survival analysis and correlation analysis between TMB and ferroptosis-related biomarkers were conducted. Moreover, the expression levels of immune checkpoints were compared between the two risk groups.

Statistical Analysis

The data were statistically processed by the R package. The Wilcox test was used to compare two groups in boxplots. Survival differences were determined by the Kaplan-Meier

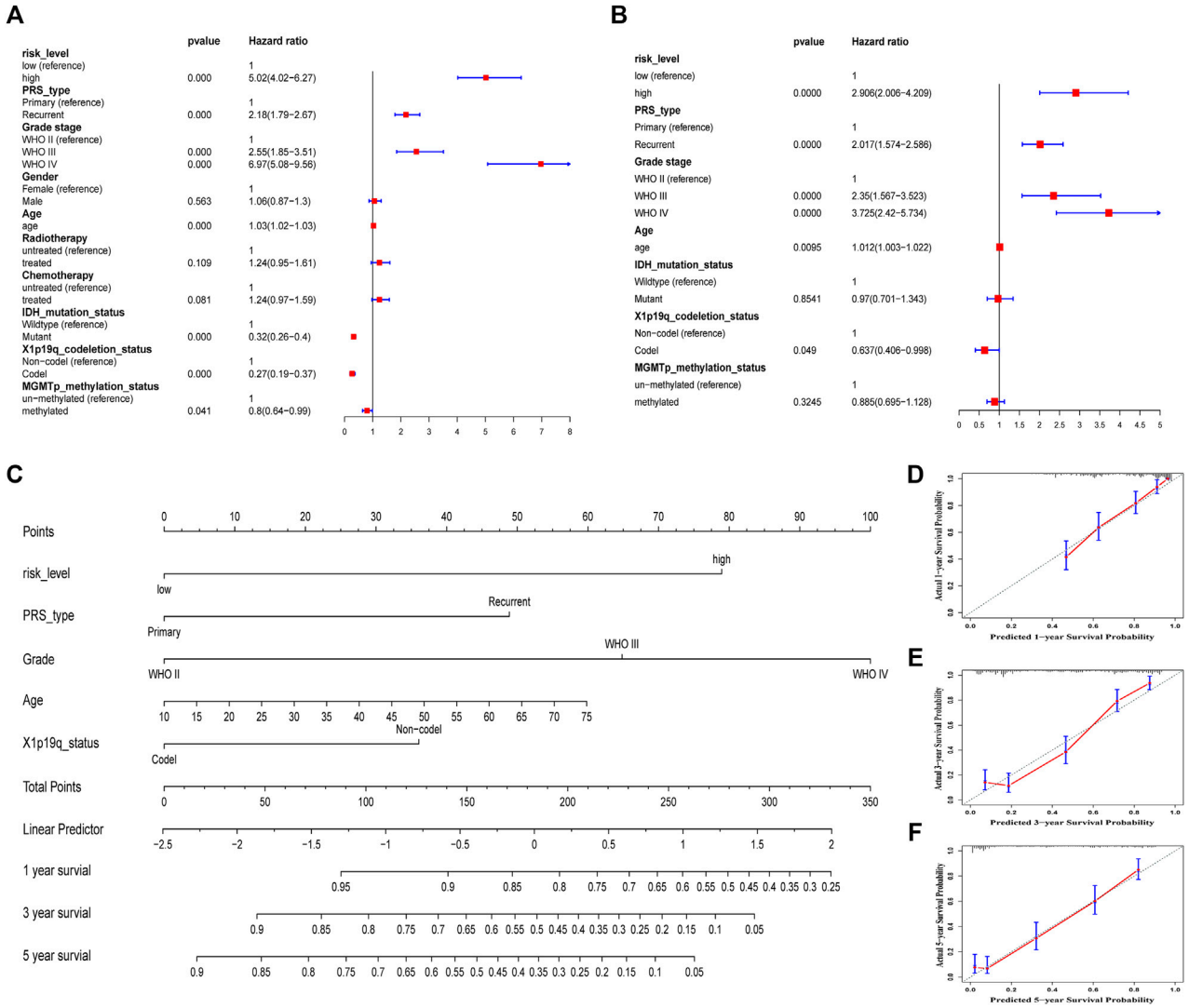


FIGURE 8 | Construction and assessment of a nomogram in the training CGGA693 cohort. **(A)** Univariate Cox regression. **(B)** Multivariate cox regression. **(C)** A prognostic nomogram was used to predict the 1-, 3-, and 5-years survival probability of patients with glioma. **(D–F)** Calibration curves of the nomogram at 1, 3, and 5 years. The *Y*-axis and *X*-axis represent the actual and nomogram-predicted survival probability, respectively.

curve and long-rank test. $p < 0.05$ was considered to indicate statistically significant differences.

RESULTS

Construction of a Prognostic Model

The clinical information of the glioma patients obtained from the CGGA and TCGA cohorts is summarized in **Table 1** and **Table 2**. Nine ferroptosis-related prognostic lncRNAs were identified in the training CGGA693 cohort after serial analysis, including Spearman analysis, univariate Cox regression analysis, and LASSO-Cox regression analysis (**Figures 2A–C**). The risk score was thus calculated by the expression of 9 ferroptosis-related lncRNAs and their corresponding regression coefficients: Risk score = 0.9849102 * $Exp_{AC010729.2}$−0.01261145 * $Exp_{AC062021.1}$ + 0.18957587 * $Exp_{FAM225B}$—0.05426328 * Exp_{FAM66C} + 0.02465331 * $Exp_{HOXA-AS2}$ + 0.03022672 * $Exp_{LINC00662}$ + 0.01995719 * $Exp_{LINC00665}$−0.00924465 * $Exp_{MIR497HG}$—1.09779996 * $Exp_{TME72-AS1}$ (**Supplementary Table S2**). These lncRNAs included 4 protective biomarkers (*AC062021.1, FAM66C, MIR497HG, and TMEM72-AS1*) and 5 risk biomarkers (*AC010729.2, FAM225B, HOXA-AS2, LINC00662, and LINC00665*). A Sankey diagram was drawn to visualize the association among these 259 ferroptosis-related genes, 9 ferroptosis-related lncRNAs, and OS in patients with glioma (**Figure 2D**). Furthermore, survival analysis of these biomarkers in the training CGGA693 cohort was conducted (**Figure 3**).

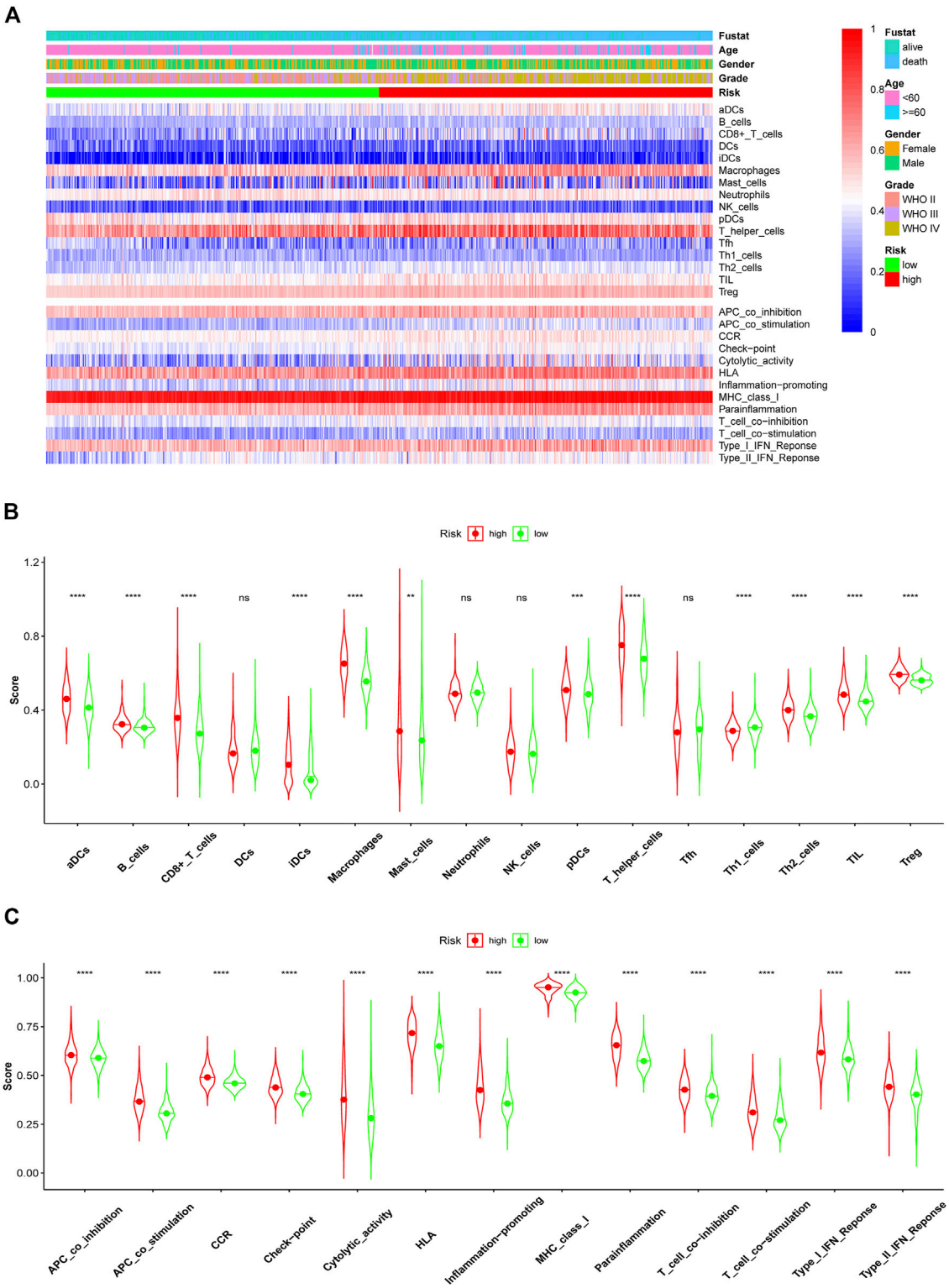


FIGURE 9 | Comparison of immune cell infiltration and functions elevated by ssGSEA between the high- and low-risk groups in the training CGGA693 cohort. **(A)** Heatmap of ssGSEA scores of immune cells and functions. The ssGSEA scores of 16 immune cells **(B)** and 13 immune functions **(C)** between the high- and low-risk groups of glioma patients in violin plots.

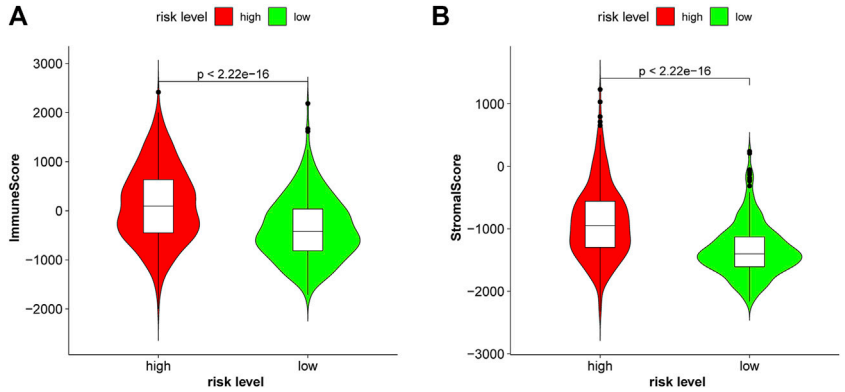


FIGURE 10 | Patients with glioma in the high-risk group had higher immune scores **(A)** and higher stromal scores **(B)** in the CGGA693.

Assessment and Validation

According to the median risk score, patients in the training CGGA693 cohort were divided into two groups. The survival analysis showed that the OS (median OS: 94.5 months) of the low-risk group was significantly prolonged compared with that (median OS: 22.2 months) of the high-risk group (**Figure 4A**, $p < 0.0001$). To assess the specificity and sensitivity of this prognostic risk model in glioma, time-independent ROC curves were plotted, and the AUCs used to predict the 1-, 3-, and 5-years OS of glioma patients were 0.791, 0.84, and 0.856, respectively (**Figure 4B**). The risk score curve, the survival status curve, and the expression heatmap of these prognostic biomarkers in glioma patients are shown in **Figures 4C–E**. The survival status curve showed a larger proportion of dead patients in the high-risk group, which was consistent with the survival analysis. The results were validated in the CGGA325 and TCGA cohorts (**Supplementary Figures S1, S2**).

Association of Prognostic Biomarkers and Clinicopathologic Features

To determine whether these ferroptosis-related prognostic lncRNAs were associated with clinical features, expression in different clinical groups was analyzed. There were significant differences in the expression of 9 ferroptosis-related prognostic lncRNAs between the high- and low-risk groups in the CGGA693 cohort (**Figure 5A**), which were validated in the CGGA325 cohort (**Supplementary Figure S3A**) and the TCGA cohort (**Supplementary Figure S4A**). Compared with the WHO grade II group, the expression of *FAM66C* and *TMEM72-AS1* was significantly different in the WHO grade III group, and the expression of eight prognostic biomarkers was significantly different in the WHO grade IV groups (CGGA693 and CGGA325 cohorts) (**Figure 5B**, **Supplementary Figure S3B**). Compared with the primary glioma group, the expression of *AC062021.1, FAM225B, LINC00662, LINC00665*, and *TMEM72-AS1* was significantly different in the recurrent glioma group of the training CGGA693 cohort (**Figure 5C**). However, only *AC062021.1, HOXA-AS2*, and *TMEM72-AS1* in the recurrent glioma group were significantly different, and only *AC062021.1, FAM66C,* and *LINC00662* were significantly different in the secondary recurrent glioma group of the validation CGGA325 cohort, which may be caused by the limited sample size (**Supplementary Figure S3C**). There were different expression levels of eight ferroptosis-related prognostic lncRNAs between the groups with and without IDH gene mutations in the CGGA693 cohort (**Figure 5D**), which was consistent with the CGGA325 cohort (**Supplementary Figure S3D**). Similar findings were demonstrated between the two types of groups with and without 1p19q codeletion (**Figure 5E**), which was validated in the CGGA325 cohort (**Supplementary Figure S3E**). In addition, the expression of all 9 lncRNAs was significantly different between the GBM and LGG subgroups in TCGA cohort (**Supplementary Figure S4B**).

Functional Enrichment Analyses

To elucidate the underlying biological mechanism of the differences between the two risk groups, GSEA was performed in the CGGA693 cohort. The results showed that oxidative phosphorylation was significantly enriched in the low-risk group (**Figure 6A**). Several key pathways (cell cycle, ECM-receptor interaction, p53 signaling, JAK-STAT signaling, focal adhesion, regulation of actin cytoskeleton, cancer and Toll-like receptor (TLR) signaling pathways) were significantly enriched in the high-risk group (**Figures 6B–I**).

The expression correlations of lncRNA-miRNA and miRNA-mRNA in potential ceRNA pairs in glioma were predicted to explore the possible regulatory mechanism of lncRNAs. Only eight of 9 lncRNAs were found to have regulatory mechanisms. (**Figure 7**, **Supplementary Table S3**). They may play important roles in the transcriptional regulation of glioma. GO analysis showed that the regulation of angiogenesis was obviously enriched in the BP term of glioma (**Supplementary Figure S5**). Chen et al. reported that elevated ATF4 expression can enhance proliferation, migration, and angiogenesis in glioma, but the ferroptosis inducer erastin can attenuate this effect. This result implied that ferroptosis could inhibit tumor angiogenesis (Chen, et al., 2017). Moreover, the PI3K-Akt signaling pathway was enriched in the KEGG term of glioma (**Supplementary Figure S5**). Yi et al. found that activating the mutation of PI3K can result in ferroptosis resistance in tumor cells, while the expression of *SREBP1* or *SCD1* can inhibit the PI3K/Akt/mTOR axis by

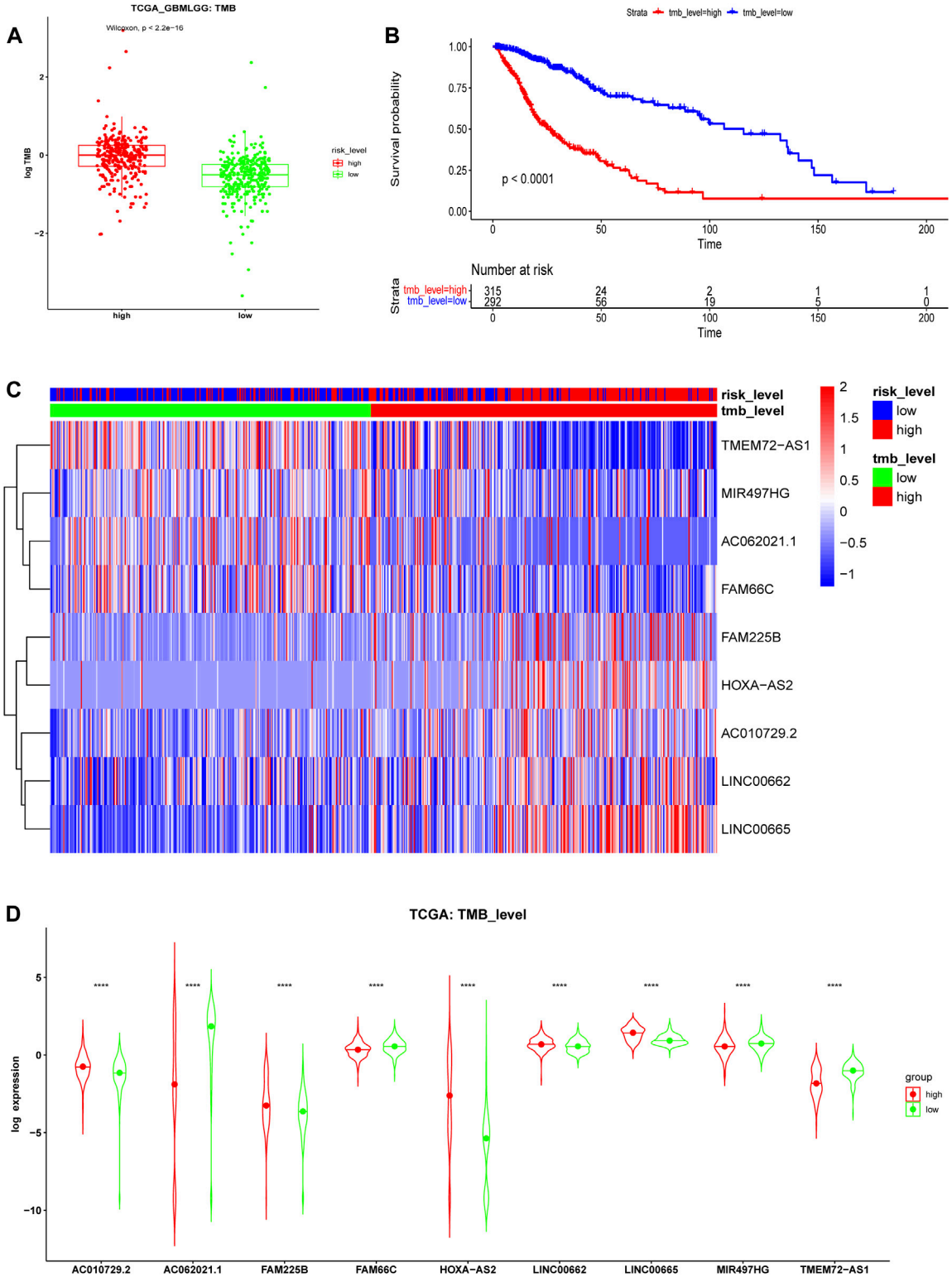


FIGURE 11 | Tumor mutational burden (TMB) analysis in the TCGA cohort. **(A)** TMB levels of high- and low-risk groups. **(B)** Kaplan-Meier survival analysis between high- and low-TMB groups divided by the median TMB. **(C)** Expression heatmap and **(D)** violin plots of 9 ferroptosis-related lncRNAs between the high- and low-level TMB groups in the TCGA cohort.

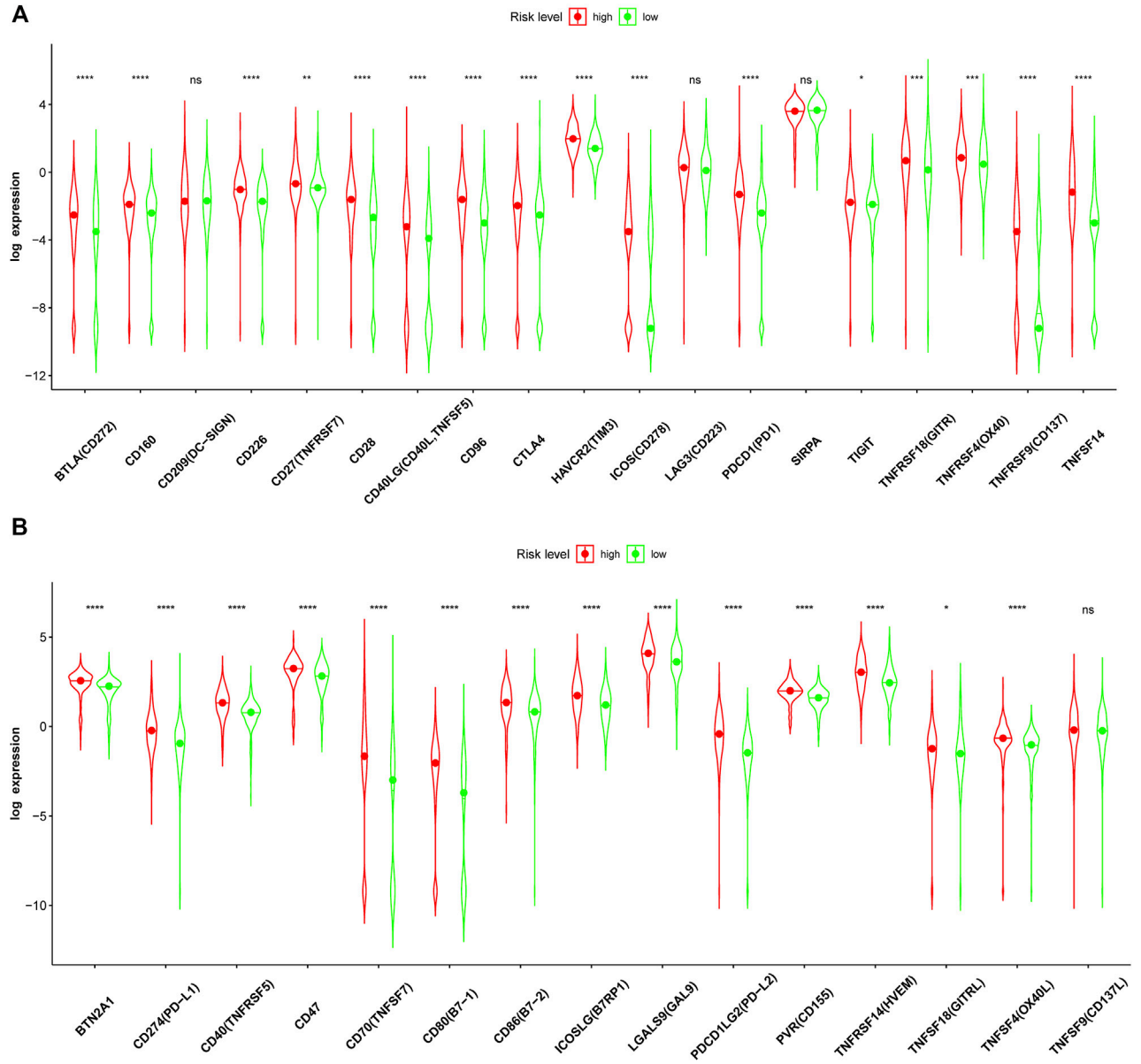


FIGURE 12 | Immune checkpoint analysis in the CGGA693 cohort. **(A, B)** Expression of immune checkpoints between the high- and low-risk groups in the training CGGA693 cohort.

TABLE 3 | Currently ongoing clinical trials for immune checkpoint inhibitors in cancer.

Immune checkpoint	Monoclonal antibody	NCT number
CTLA4	Ipilimumab, tremelimumab	NCT04084951
HAVCR2 (TIM3)	Cobolimab (TSR-022), MBG-453, INCAGN02390	NCT04139902, NCT02608268, NCT04370704
PDCD1 (PD1)	Pembrolizumab, nivolumab, tislelizumab, toripalimab, TSR-042	NCT02563002, NCT03113266, NCT03307785
TIGIT	BMS-986207, AMR101, Icosapent Ethyl Oral Capsule	NCT04570839, NCT03661047, NCT04682665
TNFRSF4 (OX40)	PF-04518600	NCT03092856, NCT03971409
TNFRSF9 (CD137)	GVAX	NCT03767582
CD274 (PD-L1)	Atezolizumab, durvalumab, avelumab, IMC-001	NCT04084951, NCT04230759, NCT04268368, NCT04196465
CD40 (TNFRSF5)	APX005M (sotigalimab)	NCT03597282, NCT03719430
CD47	PF-07257876, TTI-621, HX009	NCT04881045, NCT02890368, NCT02663518, NCT04886271
CD80 (B7-1)	IMC-001	NCT04196465

regulating lipid metabolism, which can sensitize the ferroptosis of cancer cells and play an antitumor role (Yi, et al., 2020).

Nomogram Construction and Assessment

Some clinical characteristics related to the prognosis of glioma were identified after univariate and multivariate Cox regression analyses in the CGGA693 cohort. To improve predictive ability and provide a quantitative tool to predict the survival outcomes of glioma patients in clinical practice, a nomogram was constructed based on these independent prognostic factors, including risk level, primary or recurrent type, grade, age, and 1p19q codeletion status (**Figures 8A–C**). The C-index was 0.786, which indicated that this nomogram model had a good predictive value. Calibration curves confirmed again that there was a good match between the actual probability of 1-year, 3-years, and 5-years OS and the predicted probability of the nomogram (**Figures 8D–F**).

Immune Infiltration

Glioma, especially GBM, with high malignancy and a high recurrence rate, did not respond well to traditional treatment. Here, we studied whether these lncRNAs were associated with immune infiltration and thus could guide their potential application in immunotherapy in glioma. The correlation of *FAM66C* and distinct immune cells was elucidated by the TIMER database (**Supplementary Figure S6**). In addition, different scores were discovered for several immune cells and immune-related functions in the two risk groups by ssGSEA. The heatmap of the ssGSEA score displayed distinct immune infiltration statuses in the two risk groups of this ferroptosis-related model (**Figure 9A**). In the Violin diagram, most of the 16 immune cells had different scores between the high- and low-risk groups (**Figure 9B**). Similar results were also observed for 13 immune-related functions (**Figure 9C**), especially for immune checkpoints. These results implied that comprehensive analysis of immunological cells and functions may be essential before immunotherapy, especially before ICI therapy. In addition, patients in the high-risk group had higher immune scores and stromal scores (**Figure 10**).

Tumor Mutational Burden and Immune Checkpoints

TMB is a novel biomarker of the response of patients with malignant tumors to ICIs (Ready, et al., 2019). The FDA approved high TMB levels as one of the treatment standards for solid tumor patients to receive medical prescriptions with ICIs. The analysis results of the TCGA cohort showed that the risk stratification calculated from our risk model was consistent with the level of TMB. Patients at high risk had a higher TMB level than those at low risk (**Figure 11A**, $p < 0.001$), which indicated that glioma patients at high risk may benefit from immunotherapy. Kaplan-Meier curves showed that the OS in the high-TMB group was significantly lower than that in the low-TMB group (**Figure 11B**, $p < 0.001$). The expression heatmap of these 9 ferroptosis-related prognosis lncRNAs and TMB was shown in **Figure 11C**. The expression of these nine lncRNAs was significantly different between the high- and low-level TMB groups (**Figure 11D**), which affirmed that TMB was associated with the lncRNAs in our model.

Given the higher ssGSEA score of immune checkpoints and higher TMB in the high-risk group, the expression of immune checkpoints based on the risk level was further analyzed in detail. The results showed that the expression levels of many immune checkpoints in the high-risk group were significantly higher than those in the low-risk group in the CGGA693 cohort (**Figure 12**). Therefore, glioma patients with high risk in our model might receive more benefits from ICI therapy.

DISCUSSION

The role of ferroptosis, a new form of cell death, in malignant tumors has been gradually elucidated, and more ferroptosis-related genes have been identified (Liang et al., 2020; Liu et al., 2020; Zhuo et al., 2020). In recent years, ferroptosis-related lncRNAs, which play a regulatory role in protein-coding, have attracted increasing attention (Jiang et al., 2021; Wu and Liu 2021; Zhang et al., 2021). After data mining and analysis, we constructed a ferroptosis-related lncRNA risk model with prognostic value in the CGGA693 cohort and further verified the results in the CGGA325 and TCGA cohorts. In addition, a nomogram was constructed to provide valuable suggestions for clinicians to judge the OS of glioma patients. Based on the TMB level, the immune status and expression of immune checkpoints displayed a strong positive correlation between the two groups in our risk model. Thus, our risk model might be used to guide immunotherapy in patients with glioma.

The dysregulation of 9 ferroptosis-related lncRNAs was related to the OS of glioma patients, especially those with wild-type IDH, 1p19q codeletion, disease recurrence, and tumor malignancy. Some of these lncRNAs have been previously reported to play a role in glioma and other malignant tumors. For example, *FAM225B* plays an important role in cell migration and focal adhesion and is related to the OS of glioma patients (Li, et al., 2020; Ma and Liu 2021). *HOXA-AS2* is involved in the pathogenesis of glioma and regulates glioma cell viability, cell migration, and invasion, participating in the occurrence of vasculogenic mimicry and apoptosis (Gao et al., 2018; Wu et al., 2019). Wu et al. revealed that *LINC00662* in the ceRNA network (the LINC00662/miR-107/HMGB1 axis) regulated cell proliferation and the invasion of glioma and could serve as a therapeutic target for patients with glioma (Wu et al., 2020). More interestingly, *LINC00665* can encode a special micropeptide (CIP2A-BP) to inhibit the progression of three negative breast cancers or act as a ceRNA that is involved in regulating the LINC00665/AGTR1/miR-34a-5p axis and other biological behaviours in glioma (Guo et al., 2020; Ruan et al., 2020; Dai et al., 2021). Additionally, *MIR497HG* was reported to be associated with proliferation, migration, invasion, and lymph node metastasis in bladder cancer and breast cancer (Zhang et al., 2019; Zhuang et al., 2020). However, until now, there have been no reports on the function of these lncRNAs in ferroptosis; therefore, future in-depth work is required.

GSEA showed that oxidative phosphorylation-related genes were significantly enriched in the low-risk group. Oxidative phosphorylation (OXPHOS) and glycolysis can maintain tumor propagation by isogenic glioma stem cells (GSCs) but the former can switch to the latter under metabolic stress (Shibao et al., 2018). In addition, both of them can sustain the emergence of glioma independently by detecting the metabolic requirements of GSCs (Saga et al., 2014; Yoshida and Saya 2021). Gboxin, an oxidative phosphorylation inhibitor, exerts specific toxicity in glioblastoma (Shi et al., 2019). Another p53 signaling pathway enriched in the high-risk group may play an essential role in the ferroptosis-related regulation of glioma. EX527, a Sirt-1 inhibitor, can inhibit the growth of glioma by activating the p53 signaling pathway, which was reported to be related to ferroptosis in lung cancer, and HOXA-AS2 may be associated with the p53 gene in hepatocellular carcinoma (Mao et al., 2018; Lu J. et al., 2020; Lu Q. et al., 2020; Wang et al., 2020; Zhao et al., 2020). Therefore, ferroptosis-related lncRNAs may play an important role in the potential ferroptosis-related regulatory mechanism of glioma.

Several studies have shown that TMB is a very effective prognostic marker and is associated with immune checkpoints (PD-1, PD-L1, etc.) in many cancers, including GBM (Yarchoan et al., 2017; Litak et al., 2019; Ready et al., 2019; Wang and Li 2019; Marabelle et al., 2020; Yin et al., 2020). The FDA approved TMB-H as one of the treatment standards for solid tumor patients to receive ICIs (i.e., a PD-1 inhibitor, pembrolizumab) (Strickler et al., 2021). Our study confirmed that the TMB level predicts the OS of glioma patients (**Figure 11B**) and is associated with ferroptosis-related biomarkers (**Figure 11D**). Given the important role of ICIs in diverse cancers (André et al., 2020; Cacciotti et al., 2020; Ott et al., 2020), we further analyzed all immune checkpoints in glioma and found that most of those immune checkpoints, such as PD-1, PD-L1, PD-L2, CTLA4, and TIM3, were significantly different between the two risk groups (**Figure 12**). Gliomas with a high TMB may benefit from PD-1 inhibitors (Touat et al., 2020). However, gliomas usually have a low TMB and are associated with a highly immunosuppressive microenvironment, which may be a potential mechanism of immunotherapy resistance (Touat et al., 2020). Moreover, ICIs activate the body's antitumor immune response by blocking immune checkpoints, and the side effects (immune-related adverse events, irAEs) induced by activating the immune system are a major challenge in clinical practice (Postow et al., 2018). Notably, irAEs of immunotherapy have a very high incidence (54–76%), and different ICIs cause different toxic effects (Xu et al., 2018). For instance, nivolumab (anti-PD-1) usually causes endocrine toxicities; atezolizumab (anti-PD-L1) mainly causes hypothyroidism, nausea, and vomiting; pembrolizumab (anti-PD-1) mainly causes arthralgia, pneumonitis, and hepatic toxicities; and iplimumab (anti-CTLA4) mainly causes skin, gastrointestinal, and renal toxicities. Therefore, more ICIs need further exploration to reveal their efficacy and safety. Some ICIs (MEDI6469, tislelizumab, sotigalimab, avelumab, and tremelimumab) of currently ongoing clinical trials in cancers are summarized in **Table 3** and have been used in animal studies, clinical trials, or patients with cancer (André et al., 2020; Cacciotti et al., 2020; Herbst et al., 2020; Powles et al., 2020; Armand et al., 2021; Baas et al., 2021; Duhen et al., 2021; Goldman et al., 2021; O'Hara et al., 2021; Ye et al., 2021; Zhai et al., 2021). Some of the immune checkpoints in our study do not currently have corresponding ICIs, suggesting that ICIs could be further developed and may play a crucial role in glioma immunotherapy in the near future.

CONCLUSION

In summary, we first identified 9 ferroptosis-related lncRNAs that could be independent prognostic biomarkers of glioma patients and created a prognostic risk model of glioma. This risk model based on these nine biomarkers can predict the outcome of patients with glioma in some clinical conditions, such as IDH wildtype, 1p19q codeletion, and disease recurrence; in particular, the nomogram can predict the OS rate of clinical patients, which may provide a valuable suggestion for clinicians to judge the OS of glioma patients. Additionally, this risk model may have potential application value for guiding immunotherapy or future ICI development for glioma patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

XG, CL, and YZ conceived and designed the study. JS, DL, and XZ were responsible for the collection and assembly of data, data analysis, and interpretation. XZ, DL, and XG were involved in the writing of the manuscript. BC and DSL provided help in revising the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (32000812, 81772829 and 81830052), the Special Program for Collaborative Innovation, the Construction Project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400), "Top-100 Talent Cultivation Plan" of Shanghai University of Medicine and Health Sciences, and Discipline Construction Program of the Pudong New District Health Bureau of Shanghai (Grant No. PWZzk2017-31).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.817643/full#supplementary-material

REFERENCES

Ampudia-Mesias, E., Puerta-Martinez, F., Bridges, M., Zellmer, D., Janeiro, A., Strokes, M., et al. (2021). CD200 Immune-Checkpoint Peptide Elicits an Anti-glioma Response through the DAP10 Signaling Pathway. *Neurotherapeutics.* 18 (3), 1980–1994. doi:10.1007/s13311-021-01038-1

André, T., Shiu, K. K., Kim, T. W., Jensen, B. V., Jensen, L. H., Punt, C., et al. (2020). Pembrolizumab in Microsatellite-Instability-High Advanced Colorectal Cancer. *New Engl. J. Med.* 383, 2207–2218. doi:10.1056/NEJMoa2017699

Armand, P., Lesokhin, A., Borrello, I., Timmerman, J., Gutierrez, M., Zhu, L., et al. (2021). A Phase 1b Study of Dual PD-1 and CTLA-4 or KIR Blockade in Patients with Relapsed/refractory Lymphoid Malignancies. *Leukemia* 35, 777–786. doi:10.1038/s41375-020-0939-1

Baas, P., Scherpereel, A., Nowak, A. K., Fujimoto, N., Peters, S., Tsao, A. S., et al. (2021). First-line Nivolumab Plus Ipilimumab in Unresectable Malignant Pleural Mesothelioma (CheckMate 743): a Multicentre, Randomised, Open-Label, Phase 3 Trial. *Lancet.* 397, 375–386. doi:10.1016/s0140-6736(20)32714-8

Brown, R. A. M., Richardson, K. L., Kabir, T. D., Trinder, D., Ganss, R., and Leedman, P. J. (2020). Altered Iron Metabolism and Impact in Cancer Biology, Metastasis, and Immunology. *Front. Oncol.* 10, 476. doi:10.3389/fonc.2020.00476

Cacciotti, C., Choi, J., Alexandrescu, S., Zimmerman, M. A., Cooney, T. M., Chordas, C., et al. (2020). Immune Checkpoint Inhibition for Pediatric Patients with Recurrent/refractory CNS Tumors: a Single Institution Experience. *J. Neurooncol.* 149, 113–122. doi:10.1007/s11060-020-03578-6

Chen, D., Fan, Z., Rauh, M., Buchfelder, M., Eyupoglu, I. Y., and Savaskan, N. (2017). ATF4 Promotes Angiogenesis and Neuronal Cell Death and Confers Ferroptosis in a xCT-dependent Manner. *Oncogene* 36, 5593–5608. doi:10.1038/onc.2017.146

Chen, L., Li, X., Liu, L., Yu, B., Xue, Y., and Liu, Y. (2015). Erastin Sensitizes Glioblastoma Cells to Temozolomide by Restraining xCT and Cystathionine-γ-Lyase Function. *Oncol. Rep.* 33, 1465–1474. doi:10.3892/or.2015.3712

Dai, Y., Zhang, Y., Hao, M., and Zhu, R. (2021). LINC00665 Functions as a Competitive Endogenous RNA to Regulate AGTR1 Expression by Sponging miR-34a-5p in Glioma. *Oncol. Rep.* 45, 1202–1212. doi:10.3892/or.2021.7949

Dixon, S. J. (2017). Ferroptosis: Bug or Feature? *Immunological Rev.* 277, 150–157. doi:10.1111/imr.12533

Duhen, R., Ballesteros-Merino, C., Frye, A. K., Tran, E., Rajamanickam, V., Chang, S. C., et al. (2021). Neoadjuvant Anti-OX40 (MEDI6469) Therapy in Patients with Head and Neck Squamous Cell Carcinoma Activates and Expands Antigen-specific Tumor-Infiltrating T Cells. *Nat. Commun.* 12, 1047. doi:10.1038/s41467-021-21383-1

Gao, Y., Yu, H., Liu, Y., Liu, X., Zheng, J., Ma, J., et al. (2018). Long Non-coding RNA HOXA-AS2 Regulates Malignant Glioma Behaviors and Vasculogenic Mimicry Formation via the MiR-373/EGFR Axis. *Cell Physiol. Biochem.* 45, 131–147. doi:10.1159/000486253

Goldman, J. W., Dvorkin, M., Chen, Y., Reinmuth, N., Hotta, K., Trukhin, D., et al. (2021). Durvalumab, with or without Tremelimumab, Plus Platinum-Etoposide versus Platinum-Etoposide Alone in First-Line Treatment of Extensive-Stage Small-Cell Lung Cancer (CASPIAN): Updated Results from a Randomised, Controlled, Open-Label, Phase 3 Trial. *Lancet Oncol.* 22, 51–65. doi:10.1016/s1470-2045(20)30539-8

Gu, X., Wan, G., Chen, N., Li, J., Chen, B., Tang, Y., et al. (2019). DGKζ Plays Crucial Roles in the Proliferation and Tumorigenicity of Human Glioblastoma. *Int. J. Biol. Sci.* 15, 1872–1881. doi:10.7150/ijbs.35193

Guo, B., Wu, S., Zhu, X., Zhang, L., Deng, J., Li, F., et al. (2020). Micropeptide CIP2A-BP Encoded by LINC00665 Inhibits Triple-Negative Breast Cancer Progression. *EMBO J.* 39, e102190. doi:10.15252/embj.2019102190

Hassannia, B., Vandenabeele, P., and Vanden Berghe, T. (2019). Targeting Ferroptosis to Iron Out Cancer. *Cancer cell* 35, 830–849. doi:10.1016/j.ccell.2019.04.002

Herbst, R. S., Giaccone, G., de Marinis, F., Reinmuth, N., Vergnenegre, A., Barrios, C. H., et al. (2020). Atezolizumab for First-Line Treatment of PD-L1-Selected Patients with NSCLC. *New Engl. J. Med.* 383, 1328–1339. doi:10.1056/NEJMoa1917346

Jiang, Z., Shi, Y., Zhao, W., Zhou, L., Zhang, B., Xie, Y., et al. (2021). Association between Chronic Periodontitis and the Risk of Alzheimer's Disease: Combination of Text Mining and GEO Dataset. *BMC oral health* 21, 466. doi:10.1186/s12903-021-01827-2

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucleic Acids Res.* 42, D92–D97. doi:10.1093/nar/gkt1248

Li, J., Zhang, Q., Ge, P., Zeng, C., Lin, F., Wang, W., et al. (2020). FAM225B Is a Prognostic lncRNA for Patients with Recurrent Glioblastoma. *Dis. markers* 2020, 8888085. doi:10.1155/2020/8888085

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells. *Nucleic Acids Res.* 48, W509–W514. doi:10.1093/nar/gkaa407

Liang, J. Y., Wang, D. S., Lin, H. C., Chen, X. X., Yang, H., Zheng, Y., et al. (2020). A Novel Ferroptosis-Related Gene Signature for Overall Survival Prediction in Patients with Hepatocellular Carcinoma. *Int. J. Biol. Sci.* 16, 2430–2441. doi:10.7150/ijbs.45050

Litak, J., Mazurek, M., Grochowski, C., Kamieniak, P., and Roliński, J. (2019). PD-L1/PD-1 Axis in Glioblastoma Multiforme. *Int. J. Mol. Sci.* 20, 5347. doi:10.3390/ijms20215347

Liu, H. J., Hu, H. M., Li, G. Z., Zhang, Y., Wu, F., Liu, X., et al. (2020). Ferroptosis-Related Gene Signature Predicts Glioma Cell Death and Glioma Patient Progression. *Front. Cel. Dev. Biol.* 8, 538. doi:10.3389/fcell.2020.00538

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a Summary. *Acta neuropathol.* 131, 803–820. doi:10.1007/s00401-016-1545-1

Lu, J., Xu, F., and Lu, H. (2020). LncRNA PVT1 Regulates Ferroptosis through miR-214-Mediated TFR1 and P53. *Life Sci.* 260, 118305. doi:10.1016/j.lfs.2020.118305

Lu, Q., Gao, J., Tang, S., Li, Z., Wang, X., Deng, C., et al. (2020). Integrated RNA Sequencing and Single-Cell Mass Cytometry Reveal a Novel Role of LncRNA HOXA-AS2 in Tumorigenesis and Stemness of Hepatocellular Carcinoma. *OncoTargets Ther.* 13, 10901–10916. doi:10.2147/ott.S272717

Ma, X., and Liu, L. (2021). Knockdown of FAM225B Inhibits the Progression of the Hypertrophic Scar Following Glaucoma Surgery by Inhibiting Autophagy. *Mol. Med. Rep.* 23, 204. doi:10.3892/mmr.2021.11843

Mao, C., Wang, X., Liu, Y., Wang, M., Yan, B., Jiang, Y., et al. (2018). A G3BP1-Interacting lncRNA Promotes Ferroptosis and Apoptosis in Cancer via Nuclear Sequestration of P53. *Cancer Res.* 78, 3484–3496. doi:10.1158/0008-5472.Can-17-3454

Marabelle, A., Fakih, M., Lopez, J., Shah, M., Shapira-Frommer, R., Nakagawa, K., et al. (2020). Association of Tumour Mutational burden with Outcomes in Patients with Advanced Solid Tumours Treated with Pembrolizumab: Prospective Biomarker Analysis of the Multicohort, Open-Label, Phase 2 KEYNOTE-158 Study. *Lancet Oncol.* 21, 1353–1365. doi:10.1016/s1470-2045(20)30445-9

O'Hara, M. H., O'Reilly, E. M., Varadhachary, G., Wolff, R. A., Wainberg, Z. A., Ko, A. H., et al. (2021). CD40 Agonistic Monoclonal Antibody APX005M (Sotigalimab) and Chemotherapy, with or without Nivolumab, for the Treatment of Metastatic Pancreatic Adenocarcinoma: an Open-Label, Multicentre, Phase 1b Study. *Lancet Oncol.* 22, 118–131. doi:10.1016/s1470-2045(20)30532-5

Ostrom, Q. T., Patil, N., Cioffi, G., Waite, K., Kruchko, C., and Barnholtz-Sloan, J. S. (2020). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017. *Neuro oncol* 22, iv1–iv96. doi:10.1093/neuonc/noaa200

Ott, P. A., Hu-Lieskovan, S., Chmielowski, B., Govindan, R., Naing, A., Bhardwaj, N., et al. (2020). A Phase Ib Trial of Personalized Neoantigen Therapy Plus Anti-PD-1 in Patients with Advanced Melanoma, Non-small Cell Lung Cancer, or Bladder Cancer. *Cell* 183, 347–362.e24. doi:10.1016/j.cell.2020.08.053

Paraskevopoulou, M. D., Vlachos, I. S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., et al. (2016). DIANA-LncBase V2: Indexing microRNA Targets on Non-coding Transcripts. *Nucleic Acids Res.* 44, D231–D238. doi:10.1093/nar/gkv1270

Postow, M. A., Sidlow, R., and Hellmann, M. D. (2018). Immune-Related Adverse Events Associated with Immune Checkpoint Blockade. *New Engl. J. Med.* 378, 158–168. doi:10.1056/NEJMra1703481

Powles, T., Park, S. H., Voog, E., Caserta, C., Valderrama, B. P., Gurney, H., et al. (2020). Avelumab Maintenance Therapy for Advanced or Metastatic Urothelial Carcinoma. *New Engl. J. Med.* 383, 1218–1230. doi:10.1056/NEJMoa2002788

Ready, N., Hellmann, M. D., Awad, M. M., Otterson, G. A., Gutierrez, M., Gainor, J. F., et al. (2019). First-Line Nivolumab Plus Ipilimumab in Advanced Non-small-cell Lung Cancer (CheckMate 568): Outcomes by Programmed Death Ligand 1 and Tumor Mutational Burden as Biomarkers. *J. Clin. Oncol.* 37, 992–1000. doi:10.1200/jco.18.01042

Ruan, X., Zheng, J., Liu, X., Liu, Y., Liu, L., Ma, J., et al. (2020). lncRNA LINC00665 Stabilized by TAF15 Impeded the Malignant Biological Behaviors of Glioma Cells via STAU1-Mediated mRNA Degradation. *Mol. ther. Nucleic Acids* 20, 823–840. doi:10.1016/j.omtn.2020.05.003

Saga, I., Shibao, S., Okubo, J., Osuka, S., Kobayashi, Y., Yamada, S., et al. (2014). Integrated Analysis Identifies Different Metabolic Signatures for Tumor-Initiating Cells in a Murine Glioblastoma Model. *Neuro oncol* 16, 1048–1056. doi:10.1093/neuonc/nou096

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Shi, Y., Lim, S. K., Liang, Q., Iyer, S. V., Wang, H. Y., Wang, Z., et al. (2019). Gboxin Is an Oxidative Phosphorylation Inhibitor that Targets Glioblastoma. *Nature* 567, 341–346. doi:10.1038/s41586-019-0993-x

Shibao, S., Minami, N., Koike, N., Fukui, N., Yoshida, K., Saya, H., et al. (2018). Metabolic Heterogeneity and Plasticity of Glioma Stem Cells in a Mouse Glioblastoma Model. *Neuro oncol.* 20, 343–354. doi:10.1093/neuonc/nox170

Smoll, N. R., Schaller, K., and Gautschi, O. P. (2013). Long-term Survival of Patients with Glioblastoma Multiforme (GBM). *J. Clin. Neurosci.* 20, 670–675. doi:10.1016/j.jocn.2012.05.040

Strickler, J. H., Hanks, B. A., and Khasraw, M. (2021). Tumor Mutational Burden as a Predictor of Immunotherapy Response: Is More Always Better? *Clin. Cancer Res.* 27, 1236–1241. doi:10.1158/1078-0432.Ccr-20-3054

Stupp, R., Taillibert, S., Kanner, A. A., Kesari, S., Steinberg, D. M., Toms, S. A., et al. (2015). Maintenance Therapy with Tumor-Treating Fields Plus Temozolomide vs Temozolomide Alone for Glioblastoma: A Randomized Clinical Trial. *Jama* 314, 2535–2543. doi:10.1001/jama.2015.16669

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. USA.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Tang, D., Chen, X., Kang, R., and Kroemer, G. (2021). Ferroptosis: Molecular Mechanisms and Health Implications. *Cel Res.* 31, 107–125. doi:10.1038/s41422-020-00441-1

Touat, M., Li, Y. Y., Boynton, A. N., Spurr, L. F., Iorgulescu, J. B., Bohrson, C. L., et al. (2020). Mechanisms and Therapeutic Implications of Hypermutation in Gliomas. *Nature* 580, 517–523. doi:10.1038/s41586-020-2209-9

Wang, M., Mao, C., Ouyang, L., Liu, Y., Lai, W., Liu, N., et al. (2019). Long Noncoding RNA LINC00336 Inhibits Ferroptosis in Lung Cancer by Functioning as a Competing Endogenous RNA. *Cel Death Differ.* 26, 2329–2343. doi:10.1038/s41418-019-0304-y

Wang, R., Ma, L., Weng, D., Yao, J., Liu, X., and Jin, F. (2016). Gallic Acid Induces Apoptosis and Enhances the Anticancer Effects of Cisplatin in Human Small Cell Lung Cancer H446 Cell Line via the ROS-dependent Mitochondrial Apoptotic Pathway. *Oncol. Rep.* 35, 3075–3083. doi:10.3892/or.2016.4690

Wang, T., Li, X., and Sun, S. L. (2020). EX527, a Sirt-1 Inhibitor, Induces Apoptosis in Glioma via Activating the P53 Signaling Pathway. *Anticancer Drugs* 31, 19–26. doi:10.1097/cad.0000000000000824

Wang, X., and Li, M. (2019). Correlate Tumor Mutation burden with Immune Signatures in Human Cancers. *BMC Immunol.* 20, 4. doi:10.1186/s12865-018-0285-5

Wang, Z., Chen, X., Liu, N., Shi, Y., Liu, Y., Ouyang, L., et al. (2021). A Nuclear Long Non-coding RNA LINC00618 Accelerates Ferroptosis in a Manner Dependent upon Apoptosis. *Mol. Ther.* 29, 263–274. doi:10.1016/j.ymthe.2020.09.024

Woo, Y., Lee, H. J., Jung, Y. M., and Jung, Y. J. (2020). Regulated Necrotic Cell Death in Alternative Tumor Therapeutic Strategies. *Cells* 9, 2709. doi:10.3390/cells9122709

Wu, H., and Liu, A. (2021). Long Non-coding RNA NEAT1 Regulates Ferroptosis Sensitivity in Non-small-cell Lung Cancer. *J. Int. Med. Res.* 49, 300060521996183. doi:10.1177/0300060521996183

Wu, J., Guo, X., Xu, D., and Zhang, H. (2020). LINC00662 Sponges miR-107 Accelerating the Invasiveness and Proliferation of Glioma Cells. *J. Cancer* 11, 5700–5712. doi:10.7150/jca.46381

Wu, L., Zhu, X., Song, Z., Chen, D., Guo, M., Liang, J., et al. (2019). Long Non-coding RNA HOXA-AS2 Enhances the Malignant Biological Behaviors in Glioma by Epigenetically Regulating RND3 Expression. *Onco Targets Ther.* 12, 9407–9419. doi:10.2147/ott.S225678

Xu, C., Chen, Y. P., Du, X. J., Liu, J. Q., Huang, C. L., Chen, L., et al. (2018). Comparative Safety of Immune Checkpoint Inhibitors in Cancer: Systematic Review and Network Meta-Analysis. *BMJ (Clinical research ed.)* 363, k4226. doi:10.1136/bmj.k4226

Yang, Y., Tai, W., Lu, N., Li, T., Liu, Y., Wu, W., et al. (2020). lncRNA ZFAS1 Promotes Lung Fibroblast-To-Myofibroblast Transition and Ferroptosis via Functioning as a ceRNA through miR-150-5p/SLC38A1 axis. *Aging* 12, 9085–9102. doi:10.18632/aging.103176

Yao, J., Chen, X., Liu, X., Li, R., Zhou, X., and Qu, Y. (2021). Characterization of a Ferroptosis and Iron-Metabolism Related lncRNA Signature in Lung Adenocarcinoma. *Cancer Cel. Int.* 21, 340. doi:10.1186/s12935-021-02027-2

Yarchoan, M., Hopkins, A., and Jaffee, E. M. (2017). Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *New Engl. J. Med.* 377, 2500–2501. doi:10.1056/NEJMc1713444

Ye, D., Liu, J., Zhou, A., Zou, Q., Li, H., Fu, C., et al. (2021). Tislelizumab in Asian Patients with Previously Treated Locally Advanced or Metastatic Urothelial Carcinoma. *Cancer Sci.* 112, 305–313. doi:10.1111/cas.14681

Yi, J., Zhu, J., Wu, J., Thompson, C. B., and Jiang, X. (2020). Oncogenic Activation of PI3K-AKT-mTOR Signaling Suppresses Ferroptosis via SREBP-Mediated Lipogenesis. *Proc. Natl. Acad. Sci. USA.* 117, 31189–31197. doi:10.1073/pnas.2017152117

Yin, W., Jiang, X., Tan, J., Xin, Z., Zhou, Q., Zhan, C., et al. (2020). Development and Validation of a Tumor Mutation Burden-Related Immune Prognostic Model for Lower-Grade Glioma. *Front. Oncol.* 10, 1409. doi:10.3389/fonc.2020.01409

Yoshida, G. J., and Saya, H. (2021). Molecular Pathology Underlying the Robustness of Cancer Stem Cells. *Regenerative Ther.* 17, 38–50. doi:10.1016/j.reth.2021.02.002

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Zhai, Y., Celis-Gutierrez, J., Voisinne, G., Mori, D., Girard, L., Burlet-Schiltz, O., et al. (2021). Opposing Regulatory Functions of the TIM3 (HAVCR2) Signalosome in Primary Effector T Cells as Revealed by Quantitative Interactomics. *Cell Mol. Immunol.* 18, 1581–1583. doi:10.1038/s41423-020-00575-7

Zhang, H., Deng, T., Liu, R., Ning, T., Yang, H., Liu, D., et al. (2020). CAF Secreted miR-522 Suppresses Ferroptosis and Promotes Acquired Chemo-Resistance in Gastric Cancer. *Mol. Cancer* 19, 43. doi:10.1186/s12943-020-01168-8

Zhang, K., Ping, L., Du, T., Liang, G., Huang, Y., Li, Z., et al. (2021). A Ferroptosis-Related lncRNAs Signature Predicts Prognosis and Immune Microenvironment for Breast Cancer. *Front. Mol. biosciences* 8, 678877. doi:10.3389/fmolb.2021.678877

Zhang, X., Zhuang, J., Liu, L., He, Z., Liu, C., Ma, X., et al. (2019). Integrative Transcriptome Data Mining for Identification of Core lncRNAs in Breast Cancer. *PeerJ* 7, e7821. doi:10.7717/peerj.7821

Zhao, Q., Liu, Y., Zhong, J., Bi, Y., Liu, Y., Ren, Z., et al. (2019). Pristimerin Induces Apoptosis and Autophagy via Activation of ROS/ASK1/JNK Pathway in Human Breast Cancer *In Vitro* and *In Vivo*. *Cel Death Discov.* 5, 125. doi:10.1038/s41420-019-0208-0

Zhao, Y., Li, Y., Zhang, R., Wang, F., Wang, T., and Jiao, Y. (2020). The Role of Erastin in Ferroptosis and its Prospects in Cancer Therapy. *OncoTargets Ther.* 13, 5429–5441. doi:10.2147/ott.S254995

Zhao, Z., Zhang, K. N., Wang, Q., Li, G., Zeng, F., Zhang, Y., et al. (2021). Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients. *Genomics, proteomics & bioinformatics* 19, 1–12. doi:10.1016/j.gpb.2020.10.005

Zhou, N., and Bao, J. (2020). FerrDb: a Manually Curated Resource for Regulators and Markers of Ferroptosis and Ferroptosis-Disease Associations. *Database (Oxford)* 2020, baaa021. doi:10.1093/database/baaa021

Zhuang, C., Liu, Y., Fu, S., Yuan, C., Luo, J., Huang, X., et al. (2020). Silencing of lncRNA MIR497HG via CRISPR/Cas13d Induces Bladder Cancer Progression through Promoting the Crosstalk between Hippo/Yap and TGF-β/Smad Signaling. *Front. Mol. biosciences* 7, 616768. doi:10.3389/fmolb.2020.616768

Zhuo, S., Chen, Z., Yang, Y., Zhang, J., Tang, J., and Yang, K. (2020). Clinical and Biological Significances of a Ferroptosis-Related Gene Signature in Glioma. *Front. Oncol.* 10, 590861. doi:10.3389/fonc.2020.590861

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Shi, Lai, Zuo, Liu, Chen, Zheng, Lu and Gu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 31 January 2022
doi: 10.3389/fcell.2021.817086

Identification of Prognostic Biomarkers for Bladder Cancer Based on DNA Methylation Profile

Shumei Zhang[1†], Jingyu Zhang[2†], Qichao Zhang[1], Yingjian Liang[3,4], Youwen Du[5] and Guohua Wang[1]*

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]Department of Neurology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China, [3]Department of General Surgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [4]Key Laboratory of Hepatosplenic Surgery, Ministry of Education, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [5]School of Life Sciences, Anhui Medical University, Hefei, China

Background: DNA methylation is an important epigenetic modification, which plays an important role in regulating gene expression at the transcriptional level. In tumor research, it has been found that the change of DNA methylation leads to the abnormality of gene structure and function, which can provide early warning for tumorigenesis. Our study aims to explore the relationship between the occurrence and development of tumor and the level of DNA methylation. Moreover, this study will provide a set of prognostic biomarkers, which can more accurately predict the survival and health of patients after treatment.

Methods: Datasets of bladder cancer patients and control samples were collected from TCGA database, differential analysis was employed to obtain genes with differential DNA methylation levels between tumor samples and normal samples. Then the protein-protein interaction network was constructed, and the potential tumor markers were further obtained by extracting Hub genes from subnet. Cox proportional hazard regression model and survival analysis were used to construct the prognostic model and screen out the prognostic markers of bladder cancer, so as to provide reference for tumor prognosis monitoring and improvement of treatment plan.

Results: In this study, we found that DNA methylation was indeed related with the occurrence of bladder cancer. Genes with differential DNA methylation could serve as potential biomarkers for bladder cancer. Through univariate and multivariate Cox proportional hazard regression analysis, we concluded that FASLG and PRKCZ can be used as prognostic biomarkers for bladder cancer. Patients can be classified into high or low risk group by using this two-gene prognostic model. By detecting the methylation status of these genes, we can evaluate the survival of patients.

Conclusion: The analysis in our study indicates that the methylation status of tumor-related genes can be used as prognostic biomarkers of bladder cancer.

Keywords: bladder cancer, DNA methylation, protein-protein interaction network, survival analysis, prognostic markers

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Chunquan Li,
Harbin Medical University, China
Junwei Han,
Harbin Medical University, China

****Correspondence:***
Guohua Wang
ghwang@nefu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *17 November 2021*
Accepted: *22 December 2021*
Published: *31 January 2022*

Citation:
Zhang S, Zhang J, Zhang Q, Liang Y, Du Y and Wang G (2022) Identification of Prognostic Biomarkers for Bladder Cancer Based on DNA Methylation Profile. Front. Cell Dev. Biol. 9:817086. doi: 10.3389/fcell.2021.817086

INTRODUCTION

In recent years, statistical research on global diseases showed that cancer has become an important killer of human health, and seriously threatens people's life and health. It is generally believed that cancer is caused by the accumulation of mutations in cancer susceptibility genes and resulting abnormal cell growth, but a large number of recent studies have shown that in addition to genetic variation, abnormal DNA methylation also plays an important role in the occurrence and development of cancer (Tang et al., 2018). There are many studies demonstrating the importance of DNA methylation (Mo et al., 2020; Tanaka et al., 2020; Zuo et al., 2020; Luo et al., 2021). Numerous studies have shown that global hypomethylation of DNA and hypermethylation of cytosine-phosphate-guanine (CpG)-enriched regions are common in cancers (Jaenisch and Bird, 2003; Karsli-Ceppioglu et al., 2014; Sahu and Pattanayak, 2020; Yalcin and Otu, 2021). Methylation of promoters inhibits gene transcription, and abnormal methylation is one of the main causes of genomic instability, oncogene activation and tumor suppressor gene suppression. For example, abnormal methylation in colorectal tumors is characterized by hypermethylation in promoters and transcriptional silencing of tumor suppressors or DNA repair genes (Bariol et al., 2003; Øster et al., 2011; Shen and Zou, 2020; Cheng et al., 2021), coexisting with global methylation loss that leads to chromosomal and microsatellite instability and oncogene activation (Beggs et al., 2013). Both promoter hypermethylation and global hypomethylation are markers of the early stage of colorectal cancer (Lao and Grady, 2011; Luo et al., 2014; Gao et al., 2020; Mohammed et al., 2021). Therefore, abnormal methylation may contribute greatly to the pathogenesis and progression of cancer.

Alterations in DNA methylation patterns on CpG islands were the first and most significant epigenetic abnormalities found in cancer cells. DNA methylation, an epigenetic regulator of gene expression, can lead to gene silencing. Elevated methylation of tumor suppressor genes is an early event in many tumors, suggesting that altered patterns of DNA methylation may be one of the first detectable tumor-specific changes associated with tumorigenesis. Cancer-specific methylation patterns have been found to be biomarkers for cancer. Yang et al. identified novel DNA methylation markers associated with epithelial ovarian cancer risk in a genome-wide methylation study of 63,000 women of European ancestry, and proposed that methylation of multiple CpG sites may influence epithelial ovarian cancer risk by regulating gene expression (Yang et al., 2018; Singh et al., 2021). In addition, DNA methylation has several advantages over somatic mutation analysis for cancer detection, such as higher clinical sensitivity and dynamic range. Abnormal DNA methylation can also interact with genomic and other epigenomic changes, such as histone modification and nucleosome localization during tumorigenesis and development (Feinberg et al., 2006; Esteller, 2007; Baylin and Jones, 2011). Numerous studies have demonstrated that cancer-specific methylation patterns can serve as biomarkers for cancer. Hypermethylation of the promoter region of ST18 (breast cancer suppressor gene 18) and expression loss of ST18 in tumor cells of breast cancer patients suggest that this epigenetic mechanism is a marker responsible for tumor-specific down-regulation (Jandrig et al., 2004). Currently, abnormal DNA methylation is widely studied as a disordered epigenetic mechanism in gastric cancer. For example, it has been shown that some tumor suppressor genes or tumor-related genes (such as p16, Runx3, MLH1 and CDH1) in gastric cancer can be silenced by promoter methylation (Calcagno et al., 2013). In addition, CpG island methylation also occurs in non-neoplastic gastric mucosa infected with *Helicobacter pylori* (Maekita et al., 2006; Tahara et al., 2009; Iqubal et al., 2020). Since *Helicobacter pylori* infection is closely related to the occurrence of gastric cancer (Uemura et al., 2001), the occurrence of DNA methylation in gastric mucosa after infection increases the risk of gastric cancer. All these studies confirmed the important value of DNA methylation as a tumor marker.

In addition, because abnormal DNA methylation is an early event in many tumors, the precancerous changes in these tumors were often recorded. Studies have observed a dynamic change in DNA methylation from normal gastric mucosa to precancerous lesions to gastric cancer (Lee et al., 2004). Genome-wide DNA methylation analysis for cervical intraepithelial neoplasia identified novel specific methylation markers, which can be used as a population-screening method for HPV-positive women (Boers et al., 2016). Significant epigenetic heterogeneity exists in the early precancerous liver tissues of individual patients, highlighting the potential contribution of DNA methylation to the progression of liver disease and revealing strategies for identifying epigenetic drivers of HCC (Hlady et al., 2017). An analysis of genome-wide gene expression microarray and DNA methylation microarray explored genes with abnormal DNA methylation in the early detection of breast cancer, among which HOXA4 and IGF1 were identified as increased DNA methylation in breast cancer, which may be used as biomarkers for early detection of breast cancer (Li et al., 2019; Zhuang et al., 2020; Kanathezath et al., 2021). In addition, an association between DNA methylation status and tumor survival has been demonstrated. Zhang et al. analyzed EGLNs DNA methylation data in tumor tissue samples from 1230 early-stage NSCLC patients from 5 international cohort groups and gene expression data from the Cancer Genome Atlas (TCGA) and found that EGLNs DNA methylation was associated with overall survival in early-stage NSCLC patients (Zhang et al., 2019). This suggests that DNA methylation of genes can be used as a biomarker for tumor prognosis. Therefore, it is an important research direction to interpret the occurrence and development of tumors from the perspective of epigenetic abnormalities.

Bladder cancer is a malignant tumor occurring on the bladder mucosa. It is the most common malignant tumor of the urinary system and one of the top ten common tumors. The etiology of bladder cancer is complex, including internal genetic factors and external environmental factors. A large number of studies have focused on abnormalities in DNA methylation and its important role in the occurrence and development of bladder cancer. Kawakami et al. reported for the first time that MSH3 epigenetic regulation by means of DNA methylation might

contribute to gene silencing, being implicated in bladder cancer carcinogenesis (Kawakami et al., 2004). Chen et al. demonstrated that Urine DNA methylation was closely associated with early detection and recurrence monitoring for bladder cancer. They discovered 26 significant methylation markers of bladder cancer in combined analyses. They built and validated a 2-marker diagnostic model that discriminated among patients with bladder cancer with high accuracy, sensitivity and specificity (Chen et al., 2020). In addition, there are many researchers aimed of the prognosis of bladder cancer at the level of DNA methylation. Recently, with BLCA sample transcriptome data and methylation data from The Cancer Genome Atlas (TCGA), 18 target genes were identified and the signature based on them was considered an effective and independent prognostic factor (Liu et al., 2021). A meta-analysis of nineteen studies showed that patients with methylated DNA had poorer OS, compared with those with unmethylated DNA, the combined HR was 2.766 (95% CI: 2.099-3.806). Simultaneously, methylated DNA was considerably associated with shorter PFS (HR = 2.872, 95% CI: 1.971-4.185) (Yu et al., 2018). This indicated that DNA methylation was negatively correlated with the prognosis of BC patients and might become a promising biomarker. These results indicated that it is feasible and significant to analyze the prognostic markers of bladder cancer from the perspective of DNA methylation.

In this study, the DNA methylation data of bladder cancer from TCGA database were used to preliminarily obtain the potential biomarkers of bladder cancer by identifying the genes with differential DNA methylation between bladder cancer and normal control samples. However, this may not be comprehensive enough, so we further screen for potential cancer biomarkers by constructing a protein-protein interaction network. Literature excavation was conducted to confirm whether the association of these biomarkers with bladder cancer has been studied, thus proving the reliability of our screening method. Finally, we screened out prognostic biomarkers for bladder cancer by applying univariate and multivariate Cox proportional risk regression models, and divided patients into two risk groups by a prognostic index model, and verified the reliability of the prognostic model in the test set. Our study can provide a reference for clinicians to monitor patients' survival status and timely change of treatment regimens, and provide several targets for biological experimentalists for experimental verification.

MATERIALS AND METHODS

Data Acquisition and Processing

The Illumina Infinium HumanMethylation450 Bead Chip DNA methylation profile data of bladder cancer and clinical information and survival data of the samples were obtained from TCGA database (Cancer Genome Atlas Research Network, 2013), including 408 tumor samples, 14 normal control samples. The methylation level of each probe was represented by β-value, which ranges from 0 to 1, corresponding to unmethylated and fully methylated, respectively. Probes with missing data in more than 70% of the samples were removed. The remaining probes with not available (NAs) were imputed using the k-nearest neighbors (knn) imputation procedure. Unstable genomic sites, including CpGs in sex chromosomes and single nucleotide polymorphisms were removed. Because DNA methylation in promoter regions strongly influences gene expression, we selected CpGs within promotor regions. Promoter regions were defined as 2 kb upstream to 0.5 kb downstream from transcription start sites.

Identification of Differentially Methylated Genes

DNA methylation is the most extensively documented epigenetic modification that can influence cell fate and gene expression (Jones and Baylin, 2007), which finally leads to the inhibition of gene expression through formation of heterochromatin in the gene regulatory region (Li and Zhang, 2014). In this study, identification of differentially methylated genes between cancer samples and adjacent control samples for bladder cancer is our first task.

We used *t*-test to recognize the differentially methylated CpGs between cancer patients and adjacent control tissues. Benjamini-Hochberg method was used in multiple tests to adjust the *p* values. The CpGs whose adjusted *p* value were less than 0.05 and the difference of the average of β were more than 10 percent were considered distinctly differentially methylated CpG between cancer patients and adjacent control tissues. All processes were programmed using R software.

Construction of Protein-Protein Interaction Network and Extraction of Subnet

The protein-protein interaction network is composed of individual protein through the interaction between each other to participate in all aspects of the life process such as biological signal transmission, gene expression regulation, energy and material metabolism, and cell cycle regulation. This network has many important applications in bioinformatics research (Cheng et al., 2018; Liu et al., 2019; Cheng et al., 2019; Zhai et al., 2020). In this study, we chose an integrated PPI network as background, which was integrated from the Biomolecular Interaction Network Database (BIND), the Biological General Repository for Interaction Data sets (BioGRID), the Database of Interacting Proteins (DIP), the Human Protein Reference Database (HPRD), IntAct, the Molecular INTeraction database (MINT), the mammalian PPI database of the Munich Information Center on Protein Sequences (MIPS), PDZBase (a PPI database for PDZ-domains) and Reactome. Then we used the genes mapped by the DMGs (differentially methylated genes) identified above as seed gene set. They were mapped into integrated PPI network and a sub-network was extracted. The sub-network was composed of the seed genes and their one-step neighbor genes which were connecting with the seed genes in integrated PPI network. The degree of nodes in sub-network was defined as the number of genes which were connected with the specific genes. It could be used to predict the importance of

identified cancer-related DMGs. Genes with larger degree are often regarded as the Hub genes in the network. The network was visualized through the Cytoscape (http://www.cytoscape.org/), an open-source software for constructing biological networks. Cytoscape is one of the most popular open-source software tools for visualizing biomedical networks composed of proteins, genes and other types of interactions. It provides a multi-functional interactive visual interface for researchers to explore complex biological interconnections supported by various annotations and experimental data, thus facilitating research tasks such as predicting gene function and constructing pathways (Su et al., 2014; Cheng et al., 2020).

Literature Mining Confirms the Genes Identified From the Network

To test whether the biomarkers screened in our study are indeed associated with bladder cancer, we used PubMed (www.pubmed.gov) to conduct a literature search, and analyzed whether the genes were indeed related to bladder cancer in previous reports, so as to prove that the screening of tumor-related genes through the network is effective. PubMed (White, 2020) is a widely used search engine, built and maintained by the National Biotechnology Information Center (NCBI) of the National Library of Medicine (NLM), which can provide more than 28 million academic biomedical publications.

This method is simple and feasible. The selected genes and bladder cancer will be searched as the keywords in the literature database, and then consult the literature to see if there is a confirmed relationship between the screened genes and the occurrence and development of bladder cancer.

Since there are many genes obtained, one-to-one comparison is neither feasible nor necessary. Therefore, the several genes with higher degree values were selected for literature search.

Biological Functions and Pathways Enrichment Analysis of Differentially Methylated Genes

In this study, using DAVID (Huang et al., 2009a; Huang et al., 2009b), a database used for annotation, visualization and integration of discoveries, we conducted a GO (Gene Ontology) biological functions enrichment analysis and a KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways enrichment analysis towards the list of candidate biomarker genes, with p controlled within 0.05, which could find out the biological characteristics involved by our biomarker genes.

The Construction of Prognostic Model and Survival Analysis

In order to be accurate, all cancer patients were divided into two data sets on average in this study, a training set and a test set. The training set was used for establishment of models and screening of prognostic markers while the test set was for follow-up validation of the prognostic model proposed by training set. The division of two sets should meet the following criteria: 1) All samples are divided into training set and test set randomly. 2) There were no significant differences in age distribution, staging, follow-up time and mortality between the two sets (Fisher's exact test or t test). That is to say, patients of all types were randomly assigned to the training and test sets, including patients with missing clinical information. Then, we used the samples in training set and the candidate prognostic markers obtained from PPI network to construct models to screen for specific prognostic markers in bladder cancer.

In the first step, we found out DNA methylation spectrum of candidate markers, as well as clinical phenotype information and follow-up information of the samples and established a univariate COX proportional hazard regression model for each gene, so as to assess the association between patient's survival and DNA methylation levels. Additionally, we also constructed univariate COX proportional hazard regression models to determine the clinical factors that significantly affect patient's survival. In the next step, genes and the clinical factors that significantly affect survival were introduced into the multivariate COX proportional hazard regression model to find independent prognostic factors (genes). For each gene i, the formulas of univariate and multivariate COX proportional hazard regression models were defined as follows:

$$\mathrm{h}(t,x)_i = h0(t)\,exp\left(\beta_{methy} methy_i\right) \quad (1)$$

$$\mathrm{h}(t,x)_i = h0(t)\,exp\left(\beta_{methy} methy_i + \sum \beta_{clinical} clinical\right) \quad (2)$$

In these two formulas, $methy_i$ is the DNA methylation level vector of Gene i in all samples, $clinical$ represents clinical attribute information, β_{methy}, $\beta_{clinical}$ are the coefficients of the regression model. The positive regression coefficient indicates that the increase of methylation level is related to the increase of death risk (risk gene), while the negative regression coefficient indicates that the increase of methylation level is related to the decrease of death risk (protective gene). Univariate and multivariate COX proportional hazard regression models were constructed using function coxph() in survival R package.

After univariate and multivariate COX proportional hazard regression analysis, independent prognostic markers that were still significant were used to calculate risk scores in the training set (prognostic index model). Risk score is a linear combination of DNA methylation level and regression coefficients of these markers, representing different risk levels of patients.

The formula is as follows:

$$\text{Risk Score} = \sum_{i=1}^{n} \beta_i X_i \quad (3)$$

In the formula, β_i is the COX regression coefficient of Gene i in the training set, X_i is the methylation level of Gene i, n is the number of genes that have a significant impact on survival. Next, taking the median risk score as the threshold, the patients in the training set were divided into a high-risk group and a low-risk group. The survival difference between the two groups was analyzed, the overall survival status of patients was estimated by Kaplan-Meier method and the statistical significance of the difference was determined by log-rank test. Functions survfit() and survdiff() in survival R package were used in the process.

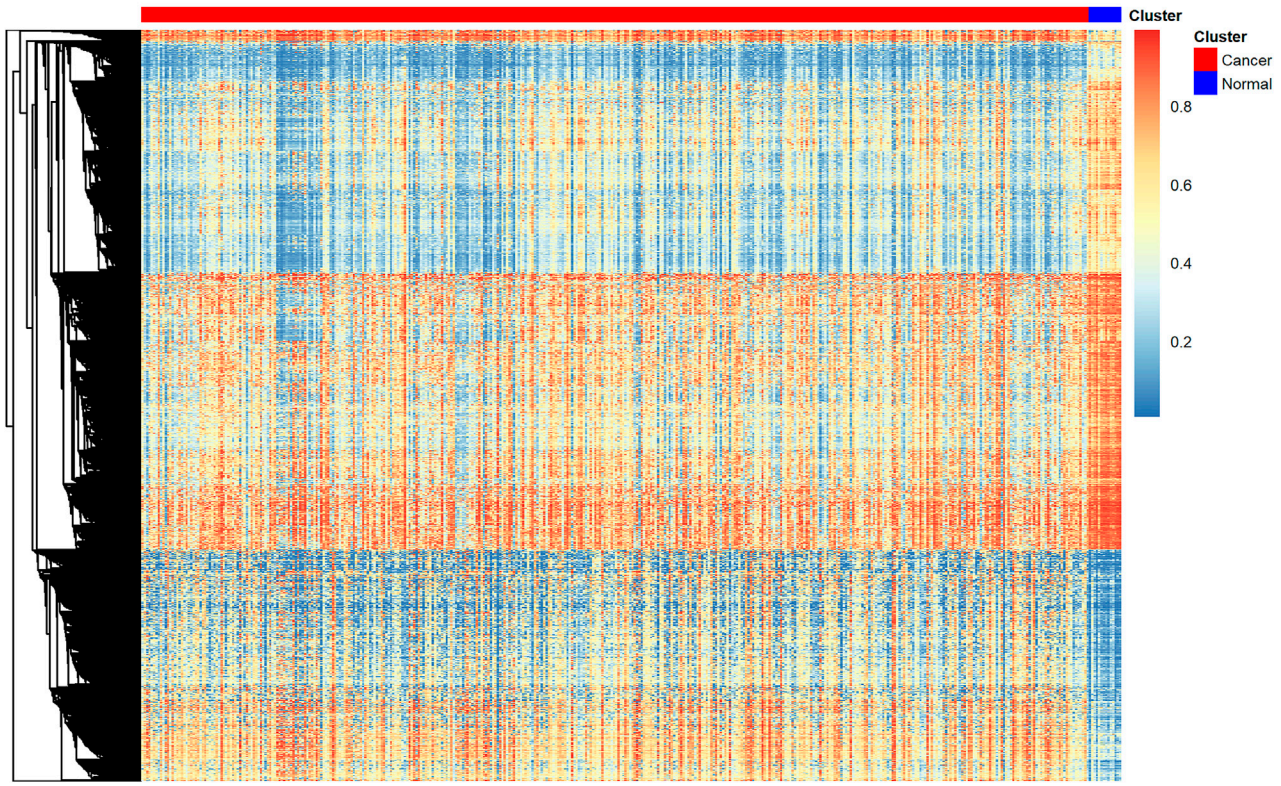


FIGURE 1 | Heat map of differentially methylated CpGs.

Then, the regression coefficients and the threshold of risk score from the training set were directly applied to the test set, and the patients in the test set were also divided into a high-risk group and a low-risk group. The prognostic difference between the two risk groups was assessed using the same method as in the training set. Besides, in order to consider the reproducibility of the prognostic biomarkers screened in this study. We searched the GEO database for an independent validation dataset GSE37816 which consisted of 18 bladder cancer patient samples and 6 normal control samples.

RESULTS

Identification of Differentially Methylated Genes

Through the processes mentioned above, we identified 9702 differentially methylated CpGs between bladder cancer patients and control samples. The differentially methylated CpGs were shown in the heat map (**Figure 1**). The heat map displays the methylation level of differentially methylated CpGs in cancer samples and adjacent control samples. We utilized the function pheatmap() in pheatmap package of R to create this graph.

By using heat map to display the complex data in the form of images, it is convenient for the observers to draw fast and accurate conclusions about the internal laws of the data.

In this figure, the rows represent CpG sites, the columns represent patients, and the colors represent the levels of DNA methylation. Red represents the higher methylation level; blue represents the lower methylation level. From this figure we can see that the differentially methylated CpGs are able to separate cancer samples and adjacent control samples clearly.

Further Potential Biomarkers of Bladder Cancer Were Obtained by Constructing Protein-Protein Interaction Network

The screening of cancer markers by differential methylation analysis alone may not be comprehensive and the collection of markers is huge which poses challenges for further screening of prognostic biomarkers, so we further used the physical interaction relationships of genes in the protein-protein interaction network to further screen candidate cancer markers. We used the obtained human protein-protein interaction information to construct a protein-protein interaction network (PPI). It can be seen that there were 13,368 nodes and 80,077 edges, indicating that they have interactions. Abnormalities in one gene can affect other genes linked to it, increasing the risk of diseases associated with that specific gene. Therefore, the genes linked to the differentially methylated genes we obtained previously in PPI may also have a potential impact on the occurrence and development of bladder

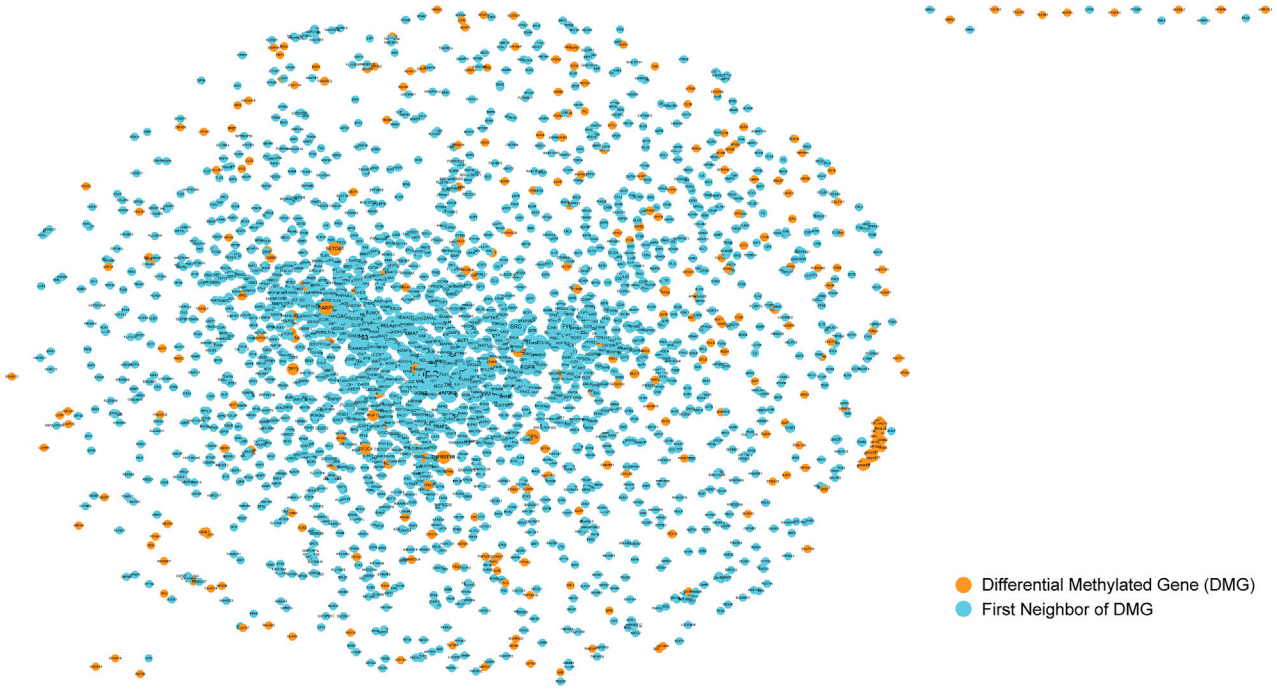


FIGURE 2 | Sub-network of protein-protein interaction network. The nodes colored orange represents the differential methylated genes, and the first neighbors of them were represented by blue nodes.

cancer. So, after that, the differentially methylated CpGs obtained in the previous step were mapped to gene symbols and mapped into the PPI. Then we selected these nodes and their one-step neighbor genes to construct a subnet(**Figure 2**). It can be seen that the subnet contains 2466 nodes and 21292 edges. The size of the node is represented by its degree, the greater the degree of the node, the larger the node. The nodes colored orange represents the differential methylated genes, and the first neighbors of them were represented by blue nodes. Because the nodes with larger degree connect more nodes in the network, they are particularly important. Therefore, nodes (genes) with degree ranked in top 10% of the subnet (246 genes, **Supplementary Material**) were selected for further analysis to obtain prognostic biomarkers for bladder cancer.

Bladder Cancer Related Genes Were Confirmed by Literature Review

Next, to verify that our method of screening for biomarkers for bladder cancer is reliable, we selected the top genes for literature mining and verification. To see if there is any evidence in the literature that these genes are indeed involved in the development and progression of bladder cancer. We found the following conclusions: First, TP53, a well-known tumor suppressor, is unsurprisingly associated with bladder cancer. According to a recent study, RB1 and TP53 co-mutations correlated strongly with genomic biomarkers of response to immunity checkpoint inhibitors in urothelial bladder cancer (Manzano et al., 2021). A meta-analysis provided further evidences that the expression of TP53 mutation was associated with the diagnosis efficiency of advanced bladder cancer. Higher expression of TP53 mutation was observed in the high stage of bladder cancer or the MIBC, and lower expression of TP53 mutation in the Ta stage of bladder cancer or the NMIBC. The expression level of TP53 mutation was probably a critical diagnosed biomarker in advanced bladder cancer (Liao et al., 2021).

A genome-wide study showed that YWHAZ is involved in the 8q22.3 amplicon, and its gene amplification is mainly detected in muscle invasive bladder cancer (MIBC). Immunohistochemical staining confirmed that the overexpression of YWHAZ is related to higher tumor stage, lymph node/vascular infiltration and mitotic activity. Univariate and multivariate analysis further indicated the prognostic potential of YWHAZ for more aggressive cancer types. The ectopic expression of YWHAZ in bladder cells with low endogenous YWHAZ levels enhances cell resistance to Adriamycin and cisplatin and ionizing radiation (Yu et al., 2019).

Invasive urothelial carcinoma of the bladder muscle (UCB) often recurs after radical bladder surgery. Changes in the expression of sex steroid hormone receptors are related to the prognosis of UCB tumors and may represent therapeutic goals. Erben et al. detected estrogen receptor 1 (ESR1) by quantitative reverse transcription polymerase chain reaction (RT-qPCR). They found that in patients with low AR-mRNA expression, the increase in ESR1 expression was related to the decrease in RFS and DSS. The expression status of ESR1 can further divide patients with low AR expression into subgroups with significantly reduced RFS and DSS (Erben et al., 2019).

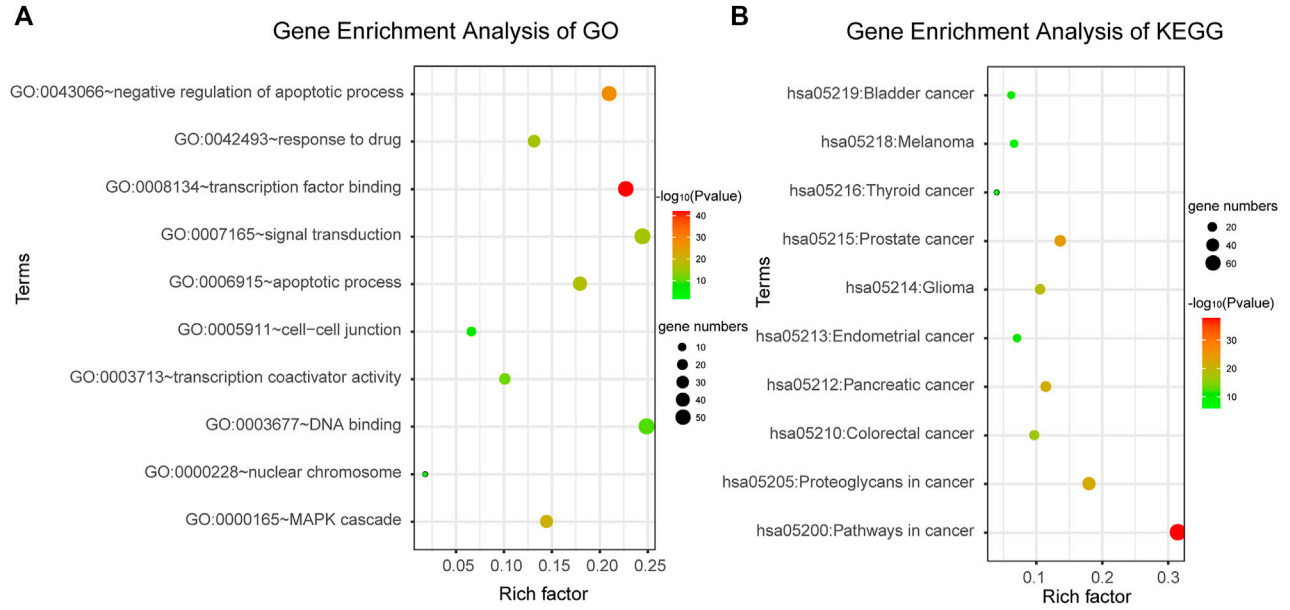


FIGURE 3 | Gene enrichment analysis. **(A)** Gene enrichment analysis of Gene Ontology. **(B)** Gene enrichment analysis of KEGG.

Besides, bladder cancer is the leading cause of morbidity and mortality worldwide. Currently, immunotherapy has become a valuable treatment for bladder cancer. Tumor mutational burden (TMB) is considered to be the most commonly used biomarker for predicting immunotherapy. Zhu et al. downloaded the somatic mutation data of bladder cancer from the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) data sets, and found that both cohorts contained 11 frequently mutated genes, including EP300 (E1A binding protein P300). The EP300 mutation was related to the increase of TMB, suggesting a good clinical prognosis. In addition, based on Gene Set Enrichment Analysis (GSEA) and CIBERSORT algorithm, they observed that the EP300 mutation up-regulated the signal pathway of the immune system and enhanced the anti-tumor immune response. In short, EP300 can be used as a biomarker for predicting immune response of bladder cancer (Zhu et al., 2020).

In Wang et al.'s study, compared with adjacent tissues, the P-SMAD2/3 protein was significantly increased ($p < 0.05$) in bladder cancer. the combination therapy of TGF-β compared with TGF, 1 + hepaCAM significantly down-regulated p-SMAD2/3, procaspase 3/8/9 and PARP, and induced bladder cancer cell apoptosis-β alone. The overexpression of hepaCAM prevented the translocation of p-SMAD2/3 from the cytoplasm to the nucleus in bladder cancer cells. Their findings revealed the important role of the p-SMAD2/3 pathway in hepaCAM-induced apoptosis of cancer cells, and provided valuable insights for current and future clinical trials of Ad-hepaCAM and p-SMAD2/3 (Wang et al., 2016).

The purpose of this section is to test whether the genes obtained from the previous operations are really related to the research problem selected in this topic, namely bladder cancer. Only when the results obtained were tested correctly, the next work would be meaningful. Through the literature review on these genes, it can be seen from the excavation that the genes analyzed in the above steps are meaningful and can be said to be "correct" to a certain extent. This has laid a solid and reliable foundation for the subsequent analysis of related genes. However, the relationship between the methylation levels of these genes and bladder cancer remains to be further studied, which is also the research content of this paper.

Pathways and Biological Functions Candidate Biomarker Genes Involved

DAVID bioinformatics tool was employed to complete the Gene Ontology and KEGG pathway enrichment analysis. From the results of enrichment, we can see that the candidate genes are involved in a number of cancer-related biological functions, such as negative regulation of apoptotic process, MAPK cascade, apoptotic process, signal transduction, response to drug, etc. (**Figure 3A**). And they are also involved in a number of cancer-related pathways, such as Pathways in cancer, Bladder cancer, Prostate cancer, Pancreatic cancer, etc. (**Figure 3B**). This suggested that the candidate genes we screened were highly likely to be cancer markers, including bladder cancer. Therefore, it was speculated that abnormalities in these genes may lead to dysregulation of related biological processes and pathways, thus inducing cancer. The results were shown in **Figure 3**, the color depth of each small bubble indicates the degree of enrichment. The redder the color, the higher the enrichment degree, and the greener the color, the lower the enrichment degree. Because there were too many significant terms, we only showed partial enrichment results.

TABLE 1 | Univariate Cox regression results for genes.

Factor	Coefficient	*p*_value	HR	Lower95	Upper95
ELF3	1.577	0.024	4.841	1.232	19.028
FASLG	2.099	0.003	8.155	2.085	31.899
PRKCZ	−3.640	0.0002	0.026	0.004	0.189
HIST3H3	1.363	0.027	3.910	1.164	13.138
BCL10	1.784	0.001	5.950	2.024	17.493

TABLE 2 | Univariate Cox regression results for clinical factors.

Factor	Coefficient	*p*_value	HR	Lower95	Upper95
Age	0.035	0.012	1.036	1.008	1.065
N	0.331	0.007	1.393	1.093	1.775
T	0.589	0.004	1.802	1.206	2.694
Stage	0.749	0.009	2.212	1.217	4.020

TABLE 3 | Multivariate Cox regression results.

Factor	Coefficient	*p*_value	HR	Lower95	Upper95
FASLG	2.099	0.003	8.155	2.085	31.899
PRKCZ	−3.640	0.0003	0.026	0.004	0.189

Identification and Analysis of Prognostic Markers in Bladder Cancer

Using 246 candidate biomarkers obtained from the network and according to the processes mentioned above, we identified 5 prognostic risk markers for BLCA from the univariate COX proportional hazard regression model (**Table 1**), including ELF3, FASLG, PRKCZ, HIST3H3, BCL10, as well as 4 clinical factors (**Table 2**), including age, N stage, T stage, and pathological stage. Further introduced the significant genes in the univariate analysis into the multivariate COX proportional hazard regression analysis, using the clinical factors that were significant in the above univariate analysis as covariates, 2 risk genes including FASLG and PRKCZ that independently affecting prognosis of patients were found in BLCA (**Table 3**).

Survival analysis of the two groups of patients which were separated by the prognostic index model (see *Materials and Methods*) showed that there were significant differences in survival between the two risk groups (**Figure 4A**). In order to verify the reliability of the prognostic risk markers selected in this study, the reserved test set from the previously obtained data set was used to verify the prognostic efficacy of the genes. Further validation based on the reserved test set using the method stated above showed that there were also significant differences in survival between the two groups (**Figure 4B**). As we can tell from the figures, the two groups of patients can be separated using the two-gene prognostic model. This suggests that the two-gene prognostic model built in this study is reliable and can be used to distinguish the high and low risk of patients.

In order to consider the reproducibility of the prognostic biomarkers identified in this study. We searched the GEO database for an independent validation dataset. However, we found only one dataset of Illumina HumanMethylation27 BeadChip data for bladder cancer, which contains patients' survival information. The dataset accession number was GSE37816 and consisted of 18 bladder cancer patient samples and 6 normal control samples. Because there are far fewer CpG sites in Illumina HumanMethylation27 BeadChip data than in Illumina HumanMethylation450 BeadChip which was used in our study, we only found cg10161121, the CpG site mapping to one prognostic marker gene FASLG in our study, which was the only biomarker corresponding to the biomarkers in our study. Although the data also included two CpG loci mapped by PRKCZ, neither of these loci was a prognostic biomarker screened in this study. Therefore, we simply used the methylation level of cg10161121 to test its prognostic efficacy. That is, patients were divided into two groups based on the median methylation level to check if there was a difference in survival. The results showed that the difference between the survival curves of the two groups only reached borderline significance (p = 0.0744) (**Figure 4C**). To see why, we also examined the representation of this gene in the TCGA dataset (which we originally used). We only used methylation level of FASLG gene to group patients, and the results showed that the p value between the two groups in the training set was 0.0722, which was also only reached borderline significance, while there was no significant difference between the two groups in the test set (p = 0.3502), the two survival curves seem roughly distinguishable, though. This also reflects the important role of PRKCZ gene in the prognostic index model. These results suggested that single genes are not sufficient for prognostic model and that combination of these two genes are needed for a robust prognostic factor development. The combination of these two genes (to calculate the prognostic index proposed in this study) could clearly separate patients into the high and low risk groups. In any case, our study was limited by the validation data set, further validation is still necessary.

Since the DNA methylation levels of genes may affect their expression levels, we also checked the expression levels of these two genes between bladder cancer and normal samples. And we found a significant difference in PRKCZ expression between cancer and normal samples (cancer mean 9.51, normal mean 7.75, p value = 0.0007). However, there was no significant difference in FASLG gene expression between cancer and normal samples (cancer mean 2.96, normal mean 3.12, p value = 0.7017), but we can see that its expression level in the normal samples is slightly higher than in the cancer patients.

In addition, we also checked if expression of these genes can be used as prognostic biomarkers. We used the median expression levels of these two genes to divide patients into two groups and test whether there was a difference in survival between the two groups. The results showed that FASLG could divide patients into two different risk groups (p = 0.0366) (**Figure 4D**), but PRKCZ could not (p = 0.9099). In fact, the occurrence and progression of cancer is not always caused by abnormal gene expression, but our results further demonstrate the reliability of the FASLG gene as a prognostic biomarker gene for bladder cancer.

After consulting the literature, FASLG (Fas ligand) has been verified to be relevant with prognosis of BLCA (Mizutani et al.,

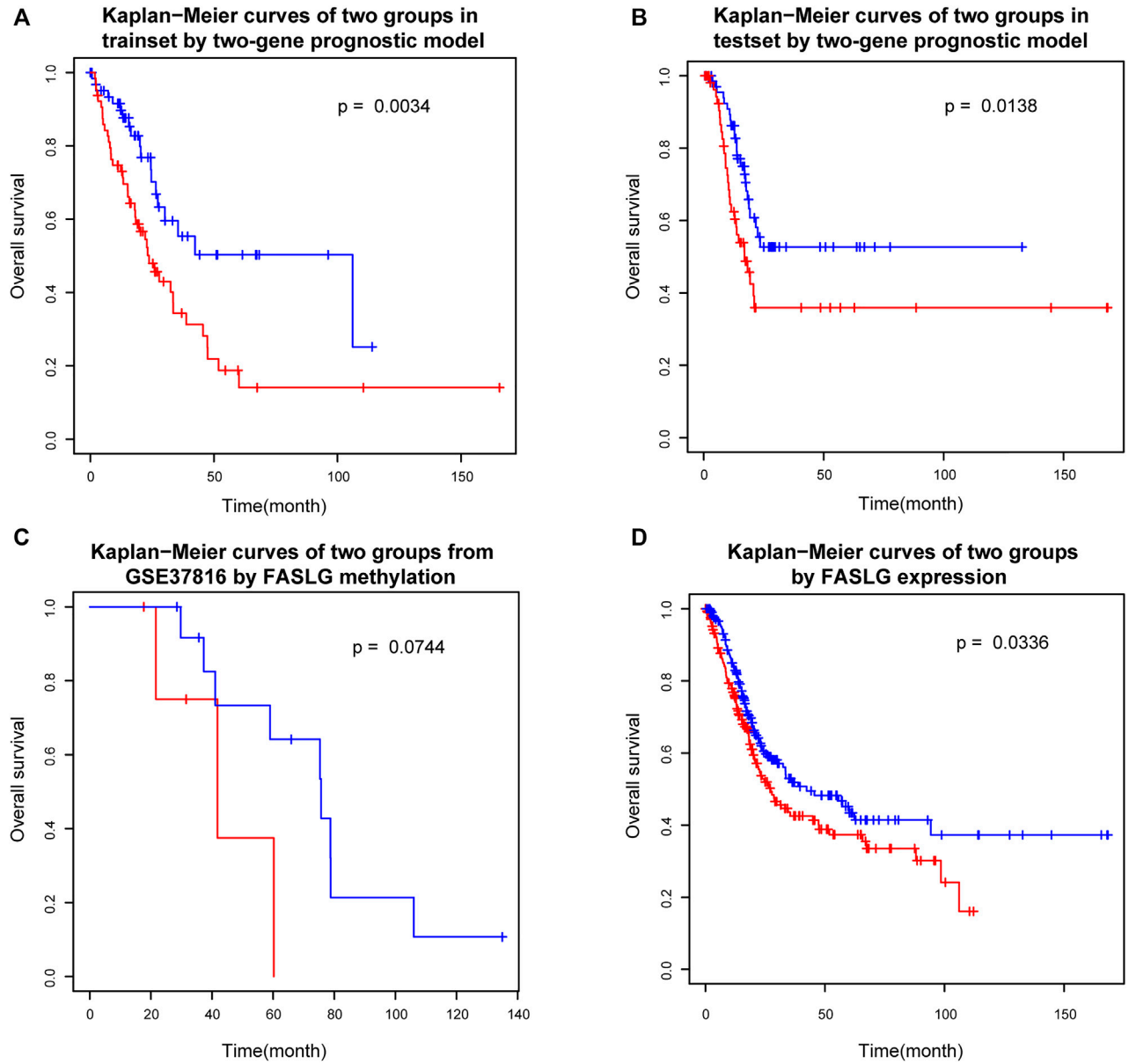


FIGURE 4 | Survival analysis of two groups. **(A)** Kaplan–Meier curves of two groups in trainset by two-gene prognostic model. **(B)** Kaplan–Meier curves of two groups in testset by two-gene prognostic model. **(C)** Kaplan–Meier curves of two groups from GSE37816 by FASLG methylation. **(D)** Kaplan–Meier curves of two groups by FASLG expression.

2002; Ben Bahria-Sediki et al., 2018). Which also confirmed that our screening method is effective. However, the prognostic value of DNA methylation levels in this gene has not been investigated. Perhaps the methylation of this gene is intrinsically related to other omics features that need further investigation. Another gene PRKCZ has not been tested to be efficient in treating this cancer, which may be potential targets for scientists and doctors to further research on it. In addition, we also checked whether DNA methylation of these two genes has been studied in other cancers or diseases, and whether it has prognostic value in other diseases. We found that the methylation of FASLG was studied in Acute Coronary Syndrome (Li et al., 2017) and gastric cancer (Moro et al., 2020). PRKCZ DNA methylation was identified as a potential epigenetic-sensitive target in acute coronary syndrome *via* network analysis (Infante et al., 2021). It was also suggested that the PRKCZ gene was the hypermethylated gene of type 2 diabetes mellitus (T2DM) and the hypermethylation PRKCZ gene may be involved in the pathogenesis of T2DM (Zou et al., 2013). But so far, no studies have been reported on the correlation between DNA methylation of these two genes and disease prognosis.

DISCUSSION

Cancer is a disease with high mortality rate and a serious threat to people's health. Previous studies focused only on the effect of genetic sequence changes on cancer, or malignancy. Recently, a relationship between cancer and the level of DNA methylation has been found. This paper studied the relationship between DNA methylation and cancer, and identified the prognostic biomarkers of bladder cancer from the perspective of DNA methylation, which can help doctors make more accurate prediction of patients' physical conditions after treatment, which is of great value. The results, however, do not seem to be encouraging, as they may only help predict patient survival. The hope is for prevention or treatment, and perhaps future studies will find a stronger link between DNA methylation and cancer.

In this article, the data used to complete the series of operations is reliable, which is very necessary, it is the basis of the work. In this study, we not only set up a training set, but also set up a test set and an independent validation set, which were used to check whether the selected biomarkers were true and effective. This method is more convincing and also ensures the scientific nature of the research. The graphical display of the results enables researchers to have a clearer understanding of the rules and changes of the data, which can help researchers find the information contained in the data more quickly.

Based on bladder cancer as the research object, this paper analyzed the spectrum based on DNA methylation of tumor prognosis biomarkers. Through a series of analysis, the valuable conclusion has been obtained. A two-gene prognostic model (the sum of regression coefficients and methylation levels) has been proposed, and their combination allows for high and low risk grouping of patients. Using only one of them could not achieve this grouping effect, even though each of them was prognostic. But it is a relative limitation, this experiment studied only one kind of cancer, whether the conclusion is suitable for other different kinds of tumors is unknown. When the study of tumor types enough, we would discover common patterns, and perhaps conclude that cancer in general was true, and that would be valuable. It is believed that if the relevant research work can be carried on in depth, people may discover the deeper link between cancer and genes, and may be able to look at and study cancer from a new angle, so as to achieve the ultimate goal of curing cancer. Of course, the process is always hard and long, but that's the beauty of science.

At the same time, in the process of research on this topic, the authors cannot help but think of other uses of DNA methylation. Considering its close relationship with tumorigenesis, perhaps the treatment of cancer is also involved in it. The research on this aspect is full of hope.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

GW conceived and designed the experiments. SZ and QZ conducted all the data processing work described in the section of methods and performed the analysis. SZ prepared and edited the manuscript. YL checked and proofread the entire manuscript. JZ and YD were responsible for the revision of the manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (62002057, 61771165, 62072095); Fundamental Research Funds for the Central Universities (2572020BH02); National Key R&D Program of China (2021YFC2100100); Heilongjiang Postdoctoral Fund (LBH-Q20135); Outstanding Youth Fund of the First Affiliated Hospital of Harbin Medical University (HYD2020JQ0008); Scientific Research Fund of Anhui Medical University (2019xkj023).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.817086/full#supplementary-material

Supplementary Table S1 | Genes with degree ranked in top 10% of the subnet.

REFERENCES

Bariol, C., Suter, C., Cheong, K., Ku, S.-L., Meagher, A., Hawkins, N., et al. (2003). The Relationship between Hypomethylation and CpG Island Methylation in Colorectal Neoplasia. *Am. J. Pathol.* 162 (4), 1361–1371. doi:10.1016/s0002-9440(10)63932-6

Baylin, S. B., and Jones, P. A. (2011). A Decade of Exploring the Cancer Epigenome - Biological and Translational Implications. *Nat. Rev. Cancer.* 11 (10), 726–734. doi:10.1038/nrc3130

Beggs, A. D., Jones, A., El-Bahwary, M., Abulafi, M., Hodgson, S. V., and Tomlinson, I. P. (2013). Whole-genome Methylation Analysis of Benign and Malignant Colorectal Tumours. *J. Pathol.* 229 (5), 697–704. doi:10.1002/path.4132

Ben Bahria-Sediki, I., Chebil, M., Sampaio, C., Martel-Frachet, V., Cherif, M., Zermani, R., et al. (2018). Prognostic Value of Soluble Death Receptor Ligands in Patients with Transitional Cell Carcinoma of Bladder. *Urol. Int.* 100 (4), 476–484. doi:10.1159/000488770

Boers, A., Wang, R., van Leeuwen, R. W., Klip, H. G., de Bock, G. H., Hollema, H., et al. (2016). Discovery of New Methylation Markers to Improve Screening for Cervical Intraepithelial Neoplasia Grade 2/3. *Clin. Epigenet.* 8, 29. doi:10.1186/s13148-016-0196-3

Calcagno, D. Q., Gigek, C. O., Chen, E. S., Burbano, R. R., and Smith Mde, A. (2013). DNA and Histone Methylation in Gastric Carcinogenesis. *Wjg* 19 (8), 1182–1192. doi:10.3748/wjg.v19.i8.1182

Cancer Genome Atlas Research Network (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Chen, S., Liu, Z., Li, M., Huang, Y., Wang, M., Zeng, W., et al. (2020). Potential Prognostic Predictors and Molecular Targets for Skin Melanoma Screened by Weighted Gene Co-expression Network Analysis. *Cgt* 20 (1), 5–14. doi:10.2174/1566523220666200516170832

Cheng, L., Qi, C., Yang, H., Lu, M., Cai, Y., Fu, T., et al. (2021). gutMGene: a Comprehensive Database for Target Genes of Gut Microbes and Microbial Metabolites. *Nucleic Acids Res.* doi:10.1093/nar/gkab786

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics* 34 (11), 1953–1956. doi:10.1093/bioinformatics/bty002

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a Comprehensive Database for Dysbiosis of the Gut Microbiota in Disorders and Interventions. *Nucleic Acids Res.* 48 (D1), D554–D560. doi:10.1093/nar/gkz843

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for Identifying Similar Diseases. *Mol. Ther. - Nucleic Acids* 18, 590–604. doi:10.1016/j.omtn.2019.09.019

Erben, P., Sikic, D., Sikic, D., Wirtz, R. M., Martini, T., Weis, C.-A., et al. (2019). Analysis of the Prognostic Relevance of Sex-Steroid Hormonal Receptor mRNA Expression in Muscle-Invasive Urothelial Carcinoma of the Urinary Bladder. *Virchows Arch.* 474 (2), 209–217. doi:10.1007/s00428-018-2496-9

Esteller, M. (2007). Cancer Epigenomics: DNA Methylomes and Histone-Modification Maps. *Nat. Rev. Genet.* 8 (4), 286–298. doi:10.1038/nrg2005

Feinberg, A. P., Ohlsson, R., and Henikoff, S. (2006). The Epigenetic Progenitor Origin of Human Cancer. *Nat. Rev. Genet.* 7 (1), 21–33. doi:10.1038/nrg1748

Gao, J., Zhang, L., Yu, G., Qu, G., Li, Y., and Yang, X. (2020). Model with the GBDT for Colorectal Adenoma Risk Diagnosis. *Curr. Bioinformatics* 15 (9), 971–979. doi:10.2174/1574893614666191120142005

Hlady, R. A., Zhou, D., Puszyk, W., Roberts, L. R., Liu, C., and Robertson, K. D. (2017). Initiation of Aberrant DNA Methylation Patterns and Heterogeneity in Precancerous Lesions of Human Hepatocellular Cancer. *Epigenetics* 12 (3), 215–225. doi:10.1080/15592294.2016.1277297

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res.* 37 (1), 1–13. doi:10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4 (1), 44–57. doi:10.1038/nprot.2008.211

Infante, T., Franzese, M., Ruocco, A., Schiano, C., Affinito, O., Pane, K., et al. (2021). ABCA1, TCF7, NFATC1, PRKCZ, and PDGFA DNA Methylation as Potential Epigenetic-Sensitive Targets in Acute Coronary Syndrome *via* Network Analysis. *Epigenetics*, 1–17. doi:10.1080/15592294.2021.1939481

Iqubal, A., Iqubal, M. K., Khan, A., Ali, J., Baboota, S., and Haque, S. E. (2020). Gene Therapy, A Novel Therapeutic Tool for Neurological Disorders: Current Progress, Challenges and Future Prospective. *Cgt* 20 (3), 184–194. doi:10.2174/1566523220999200716111502

Jaenisch, R., and Bird, A. (2003). Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals. *Nat. Genet.* 33 (Suppl. l), 245–254. doi:10.1038/ng1089

Jandrig, B., Seitz, S., Hinzmann, B., Arnold, W., Micheel, B., Koelble, K., et al. (2004). ST18 Is a Breast Cancer Tumor Suppressor Gene at Human Chromosome 8q11.2. *Oncogene* 23 (57), 9295–9302. doi:10.1038/sj.onc.1208131

Jones, P. A., and Baylin, S. B. (2007). The Epigenomics of Cancer. *Cell* 128 (4), 683–692. doi:10.1016/j.cell.2007.01.029

Kanathezath, A., Chembra, V., Padingare Variyath, S. K., and Nair, G. G. (2021). Identification of Biomarkers and Functional Modules from Genomic Data in Stage-wise Breast Cancer. *Cbio* 16 (5), 722–733. doi:10.2174/1574893615999200922123104

Karsli-Ceppioglu, S., Dagdemir, A., Judes, G., Ngollo, M., Penault-Llorca, F., Pajon, A., et al. (2014). Epigenetic Mechanisms of Breast Cancer: an Update of the Current Knowledge. *Epigenomics* 6 (6), 651–664. doi:10.2217/epi.14.59

Kawakami, T., Shiina, H., Igawa, M., Deguchi, M., Nakajima, K., Ogishima, T., et al. (2004). Inactivation of the hMSH3 Mismatch Repair Gene in Bladder Cancer. *Biochem. Biophysical Res. Commun.* 325 (3), 934–942. doi:10.1016/j.bbrc.2004.10.114

Lao, V. V., and Grady, W. M. (2011). Epigenetics and Colorectal Cancer. *Nat. Rev. Gastroenterol. Hepatol.* 8 (12), 686–700. doi:10.1038/nrgastro.2011.173

Lee, J.-H., Park, S.-J., Abraham, S. C., Seo, J.-S., Nam, J.-H., Choi, C., et al. (2004). Frequent CpG Island Methylation in Precursor Lesions and Early Gastric Adenocarcinomas. *Oncogene* 23 (26), 4646–4654. doi:10.1038/sj.onc.1207588

Li, E., and Zhang, Y. (2014). DNA Methylation in Mammals. *Cold Spring Harbor Perspect. Biol.* 6 (5), a019133. doi:10.1101/cshperspect.a019133

Li, J., Zhu, X., Yu, K., Jiang, H., Zhang, Y., Deng, S., et al. (2017). Genome-Wide Analysis of DNA Methylation and Acute Coronary Syndrome. *Circ. Res.* 120 (11), 1754–1767. doi:10.1161/circresaha.116.310324

Li, S. Y., Wu, H. C., Mai, H. F., Zhen, J. X., Li, G. S., and Chen, S. J. (2019). Microarray-based Analysis of Whole-genome DNA Methylation Profiling in Early Detection of Breast Cancer. *J. Cel Biochem* 120 (1), 658–670. doi:10.1002/jcb.27423

Liao, Y., Tang, H., Wang, M., Wang, K., Wang, Y., and Jiang, N. (2021). The Potential Diagnosis Role of TP53 Mutation in Advanced Bladder Cancer: A Meta-Analysis. *J. Clin. Lab. Anal.* 35 (5), e23765. doi:10.1002/jcla.23765

Liu, Y., Gu, Y., Su, M., Liu, H., Zhang, S., and Zhang, Y. (2019). An Analysis about Heterogeneity Among Cancers Based on the DNA Methylation Patterns. *BMC Cancer* 19 (1), 1259. doi:10.1186/s12885-019-6455-x

Liu, Z., Sun, T., Zhang, Z., Bi, J., and Kong, C. (2021). An 18-gene Signature Based on Glucose Metabolism and DNA Methylation Improves Prognostic Prediction for Urinary Bladder Cancer. *Genomics* 113 (1 Pt 2), 896–907. doi:10.1016/j.ygeno.2020.10.022

Luo, X., Zhang, T., Zhai, Y., Wang, F., Zhang, S., and Wang, G. (2021). Effects of DNA Methylation on TFs in Human Embryonic Stem Cells. *Front. Genet.* 12, 639461. doi:10.3389/fgene.2021.639461

Luo, Y., Wong, C.-J., Kaz, A. M., Dzieciatkowski, S., Carter, K. T., Morris, S. M., et al. (2014). Differences in DNA Methylation Signatures Reveal Multiple Pathways of Progression from Adenoma to Colorectal Cancer. *Gastroenterology* 147 (2), 418–429. e418. doi:10.1053/j.gastro.2014.04.039

Maekita, T., Nakazawa, K., Mihara, M., Nakajima, T., Yanaoka, K., Iguchi, M., et al. (2006). High Levels of Aberrant DNA Methylation in Helicobacter Pylori-Infected Gastric Mucosae and its Possible Association with Gastric Cancer Risk. *Clin. Cancer Res.* 12 (3 Pt 1), 989–995. doi:10.1158/1078-0432.CCR-05-2096

Manzano, R. G., Catalan-Latorre, A., and Brugarolas, A. (2021). RB1 and TP53 Co-mutations Correlate Strongly with Genomic Biomarkers of Response to Immunity Checkpoint Inhibitors in Urothelial Bladder Cancer. *BMC cancer* 21 (1), 432. doi:10.1186/s12885-021-08078-y

Mizutani, Y., Yoshida, O., Ukimura, O., Kawauchi, A., Bonavida, B., and Miki, T. (2002). Prognostic Significance of a Combination of Soluble Fas and Soluble Fas Ligand in the Serum of Patients with Ta Bladder Cancer. *Cancer Biother. Radiopharm.* 17 (5), 563–567. doi:10.1089/108497802760804790

Mo, F., Luo, Y., Fan, D., Zeng, H., Zhao, Y., Luo, M., et al. (2020). Integrated Analysis of mRNA-Seq and miRNA-Seq to Identify C-MYC, YAP1 and miR-3960 as Major Players in the Anticancer Effects of Caffeic Acid Phenethyl Ester in Human Small Cell Lung Cancer Cell Line. *Cgt* 20 (1), 15–24. doi:10.2174/1566523220666200523165159

Mohammed, M., Mwambi, H., and Omolo, B. (2021). Colorectal Cancer Classification and Survival Analysis Based on an Integrated RNA and DNA Molecular Signature. *Cbio* 16 (4), 583–600. doi:10.2174/1574893615999200711170445

Moro, H., Hattori, N., Nakamura, Y., Kimura, K., Imai, T., Maeda, M., et al. (2020). Epigenetic Priming Sensitizes Gastric Cancer Cells to Irinotecan and Cisplatin by Restoring Multiple Pathways. *Gastric Cancer* 23 (1), 105–115. doi:10.1007/s10120-019-01010-1

Øster, B., Thorsen, K., Lamy, P., Wojdacz, T. K., Hansen, L. L., Birkenkamp-Demtröder, K., et al. (2011). Identification and Validation of Highly Frequent CpG Island Hypermethylation in Colorectal Adenomas and Carcinomas. *Int. J. Cancer* 129 (12), 2855–2866. doi:10.1002/ijc.25951

Sahu, R., and Pattanayak, S. P. (2020). Strategic Developments & Future Perspective on Gene Therapy for Breast Cancer: Role of mTOR and Brk/PTK6 as Molecular Targets. *Cgt* 20 (4), 237–258. doi:10.2174/1566523220999200731002408

Shen, Z., and Zou, Q. (2020). Basic Polar and Hydrophobic Properties Are the Main Characteristics that Affect the Binding of Transcription Factors to Methylation Sites. *Bioinformatics* 36 (15), 4263–4268. doi:10.1093/bioinformatics/btaa492

Singh, G., Swain, A. C., and Mallick, B. (2021). Delineating Characteristic Sequence and Structural Features of Precursor and Mature Piwi-Interacting RNAs of

Epithelial Ovarian Cancer. *Cbio* 16 (4), 541–552. doi:10.2174/1574893615999200715164755

Su, G., Morris, J. H., Demchak, B., and Bader, G. D. (2014). Biological Network Exploration with Cytoscape 3. *Curr. Protoc. Bioinformatics* 47, 8–24. doi:10.1002/0471250953.bi0813s47

Tahara, T., Arisawa, T., Shibata, T., Nakamura, M., Yoshioka, D., Okubo, M., et al. (2009). Increased Number of Methylated CpG Islands Correlates with Helicobacter pylori Infection, Histological and Serological Severity of Chronic Gastritis. *Eur. J. Gastroenterol. Hepatol.* 21 (6), 613–619. doi:10.1097/meg.0b013e32830e28b2

Tanaka, E., Uchida, D., Shiraha, H., Kato, H., Ohyama, A., Iwamuro, M., et al. (2020). Promising Gene Therapy Using an Adenovirus Vector Carrying REIC/Dkk-3 Gene for the Treatment of Biliary Cancer. *Cgt* 20 (1), 64–70. doi:10.2174/1566523220666200309125709

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor Origin Detection with Tissue-specific miRNA and DNA Methylation Markers. *Bioinformatics* 34 (3), 398–406. doi:10.1093/bioinformatics/btx622

Uemura, N., Okamoto, S., Yamamoto, S., Matsumura, N., Yamaguchi, S., Yamakido, M., et al. (2001). Helicobacter pyloriInfection and the Development of Gastric Cancer. *N. Engl. J. Med.* 345 (11), 784–789. doi:10.1056/nejmoa001999

Wang, X., Chen, E., Tang, M., Yang, X., Wang, Y., Quan, Z., et al. (2016). The SMAD2/3 Pathway Is Involved in hepaCAM-Induced Apoptosis by Inhibiting the Nuclear Translocation of SMAD2/3 in Bladder Cancer Cells. *Tumor Biol.* 37 (8), 10731–10743. doi:10.1007/s13277-016-4821-8

White, J. (2020). PubMed 2.0. *Med. Reference Serv. Q.* 39 (4), 382–387. doi:10.1080/02763869.2020.1826228

Yalcin, D., and Otu, H. H. (2021). An Unbiased Predictive Model to Detect DNA Methylation Propensity of CpG Islands in the Human Genome. *Cbio* 16 (2), 179–196. doi:10.2174/1574893615999200724145835

Yang, Y., Wu, L., Shu, X., Lu, Y., Shu, X. O., Cai, Q., et al. (2018). Genetic Data from Nearly 63,000 Women of European Descent Predicts DNA Methylation Biomarkers and Epithelial Ovarian Cancer Risk. *Cancer Res.* 79 (3):505–517. doi:10.1158/0008-5472.CAN-18-2726

Yu, C. C., Li, C. F., Chen, I. H., Lai, M. T., Lin, Z. J., Korla, P. K., et al. (2019). YWHAZamplification/Overexpression Defines Aggressive Bladder Cancer and Contributes to Chemo-/radio-resistance by Suppressing Caspase-mediated Apoptosis. *J. Pathol.* 248 (4), 476–487. doi:10.1002/path.5274

Yu, Y., Cao, H., Zhang, M., Shi, F., Wang, R., and Liu, X. (2018). Prognostic Value of DNA Methylation for Bladder Cancer. *Clinica Chim. Acta* 484, 207–212. doi:10.1016/j.cca.2018.05.056

Zhai, Y., Chen, Y., Teng, Z., and Zhao, Y. (2020). Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions. *Front. Cel Dev. Biol.* 8, 591487. doi:10.3389/fcell.2020.591487

Zhang, R., Lai, L., He, J., Chen, C., You, D., Duan, W., et al. (2019). EGLN2 DNA Methylation and Expression Interact with HIF1A to Affect Survival of Early-Stage NSCLC. *Epigenetics*. 14 (2):118–129. doi:10.1080/15592294.2019.1573066

Zhu, G., Pei, L., Li, Y., and Gou, X. (2020). EP300 Mutation Is Associated with Tumor Mutation burden and Promotes Antitumor Immunity in Bladder Cancer Patients. *Aging* 12 (3), 2132–2141. doi:10.18632/aging.102728

Zhuang, J., Dai, S., Zhang, L., Gao, P., Han, Y., Tian, G., et al. (2020). Identifying Breast Cancer-Induced Gene Perturbations and its Application in Guiding Drug Repurposing. *Curr. Bioinformatics* 15 (9), 1075–1089. doi:10.2174/1574893615666200203104214

Zou, L., Yan, S., Guan, X., Pan, Y., and Qu, X. (2013). Hypermethylation of the PRKCZ Gene in Type 2 Diabetes Mellitus. *J. Diabetes Res.* 2013, 721493. doi:10.1155/2013/721493

Zuo, Y., Song, M., Li, H., Chen, X., Cao, P., Zheng, L., et al. (2020). Analysis of the Epigenetic Signature of Cell Reprogramming by Computational DNA Methylation Profiles. *Cbio* 15 (6), 589–599. doi:10.2174/1574893614666190919103752

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhang, Zhang, Zhang, Liang, Du and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

METHODS
published: 31 January 2022
doi: 10.3389/fcell.2021.827108

Revealing Cavin-2 Gene Function in Lung Based on Multi-Omics Data Analysis Method

Changsheng Li[1†], Jingyu Huang[1†], Hexiao Tang[1†], Bing Liu[2,3†] and Xuefeng Zhou[1]*

[1]*Department of Thoracic Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China,* [2]*Department of Pulmonary and Critical Care Medicine, Zhongnan Hospital of Wuhan University, Wuhan, China,* [3]*Wuhan Research Center for Infectious Diseases and Cancer, Chinese Academy of Medical Sciences, Wuhan, China*

Research points out that it is particularly important to comprehensively evaluate immune microenvironmental indicators and gene mutation characteristics to select the best treatment plan. Therefore, exploring the relevant genes of pulmonary injury is an important basis for the improvement of survival. In recent years, with the massive production of omics data, a large number of computational methods have been applied in the field of biomedicine. Most of these computational methods are developed for a certain type of diseases or whole diseases. Algorithms that specifically identify genes associated with pulmonary injury have not yet been developed. To fill this gap, we developed a novel method, named AdaRVM, to identify pulmonary injury-related genes in large scale. AdaRVM is the fusion of Adaboost and Relevance Vector Machine (RVM) to achieve fast and high-precision pattern recognition of pulmonary injury genetic mechanism. AdaRVM found that Cavin-2 gene has strong potential to be related to pulmonary injury. As we known, the formation and function of Caveolae are mediated by two family proteins: Caveolin and Cavin. Many studies have explored the role of Caveolin proteins, but people still knew little about Cavin family members. To verify our method and reveal the functions of cavin-2, we integrated six genome-wide association studies (GWAS) data related to lung function traits, four expression Quantitative Trait Loci (eQTL) data, and one methylation Quantitative Trait Loci (mQTL) data by Summary data level Mendelian Randomization (SMR). We found strong relationship between cavin-2 and canonical signaling pathways ERK1/2, AKT, and STAT3 which are all known to be related to lung injury.

Keywords: relevance vector machine, adaboost, lung cancer, pulmonary injury, cavin-2

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Fei Shen,
South China University of Technology, China
Tianyi Zhao,
Harbin Institute of Technology, China

****Correspondence:***
Xuefeng Zhou
snowingpeak@sina.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *01 December 2021*
Accepted: *15 December 2021*
Published: *31 January 2022*

Citation:
Li C, Huang J, Tang H, Liu B and Zhou X (2022) Revealing Cavin-2 Gene Function in Lung Based on Multi-Omics Data Analysis Method. Front. Cell Dev. Biol. 9:827108. doi: 10.3389/fcell.2021.827108

INTRODUCTION

The decline of lung function could predict mortality. Several studies have found that the decline of lung function is closely related to the occurrence of many diseases such as cardiovascular disease, lung cancer, and non-respiratory cancer (Duprez and Jacobs, 2018). Moreover, the measurement of lung function changes is the key to the diagnosis of many lung diseases such as Chronic Obstructive Pulmonary Disease (COPD) and Cystic Fibrosis (CF) (Shrine et al., 2019). These lung diseases especially COPD and CF are one of the causes of significant morbidity, mortality worldwide. For example, according to a study by (Ding et al., 2019), COPD is the fourth leading cause of death on Earth. In the diagnosis of these pulmonary diseases, the measurement of lung function indicators is

the main basis to judge the changes in the disease condition and degree. Therefore, it is necessary to understand the regulatory mechanism behind lung function to prevent and treat lung function decline.

The regulation of lung function is complicated and has not been revealed yet. Governing lung function is mediated by a variety of cells such as complex endothelial and epithelial cells, dendritic cells, alveolar macrophages, and fibroblasts (Suzuki et al., 2008). They work together to deal with lung injury by triggering inflammation and an immune response. Caveolae are commonly found in lung epithelial cells, endothelial cells, adipocytes and fibroblasts (Murata et al., 2007). Caveolae are flask—shaped invaginated structures on the plasma membrane of mammalian cells, and are involved in many important cellular functions, such as endocytosis and endocytosis transport, and cell signal transduction (Parton et al., 2018). Also, this structure has been proved to play an important role in regulating various lung diseases such as asthma, COPD, CF, and acute lung injury. The structure and function of Caveolae are mediated by Caveolins family proteins and Cavins family proteins (Parton et al., 2020). These proteins are highly expressed in a variety of lung cells. To sum up, further research of these proteins is helpful to reveal the mechanism of lung function injury.

Cavin family proteins include Cavin-1, Cavin-2, Cavin-3, and Cavin-4, which govern the expression and morphology of Caveolae. Cavin-2 protein is mainly responsible for controlling the change of the Caveolae shape. For instance, the over-expression of Cavin-2 could cause the deeper invagination of Caveolae. Cavin2, also known as Serum Deprivation Protein Response (SDPR) protein was described as a protein induced during serum deprivation (Chettimada et al., 2015). Subsequently researchers undertook an in-depth study of Cavin-2 and found that this protein is essential for cell proliferation, migration and invasion. SDPR has been reported as a tumor suppressor in a variety of cancers, such as breast, liver, stomach, and endometrial cancers. Studies have shown that SDPR works in cancers because of the SDPR gene inactivation by methylation (Ozturk et al., 2016). Additionally, the absence of Cavin-2 was found to result in the loss of Caveolae of endothelial cells in lung tissues (Hansen et al., 2013). However, little information was available on the role of Cavin-2 in the lung function injury.

Thousands of SNPs associated with complex traits have been identified through GWAS. Most of these identified loci are located in intergenic regions and how they affect phenotypes through genes or pathways is difficult to elucidate (Watanabe et al., 2017). One possible explanation is that these susceptible sites alter complex traits in individuals by regulating methylation levels of a gene or inhibiting its expression. As eQTL refers to some variation sites on chromosomes that specifically regulate mRNA and protein expression levels of a gene, some of the variants overlap with SNPs found by GWAS, suggesting that they may be involved in the regulation of gene expression (Li et al., 2015). Moreover, mQTL was applied to recognize SNP sites significantly correlated with the methylation level of a gene (McRae et al., 2018). DNA methylation usually acts as an inhibitor of gene transcription. It is clear that both mQTL and eQTL can change their expression patterns by mutating at a single locus. Therefore, the association between Cavin-2 gene expression and lung function injury can be found through the conjoint analysis of GWAS and two different types of QTL data.

Although the previous studies have found multiple genes that are related to pulmonary injury, the speed of revealing the disease-causing genes of pulmonary injury still does not meet the needs of treatment. With the continuous maturity of sequencing technology, multi omics data showed a blowout growth. This provides support for the application of computing methods in biomedical field (Zhao et al., 2020a; Zhao et al., 2021a). Computational methods have been applied to the association and interaction between mutation sites (Peng and Zhao, 2020), genes (McKinney et al., 2006; Zhao et al., 2020b), various RNAs (Asefpour Vakilian, 2020), proteins (Zhao et al., 2020c; Noé et al., 2020), metabolites (Zhao et al., 2020d), drugs (Tianyi et al., 2020) and diseases, and various databases have emerged (Zhao et al., 2021b). It has become a common method to mine disease pathogenesis on a large scale by integrating a variety of different data for modeling and analysis. However, most of these methods focus on one kind of disease or all diseases. Each disease has its own specificity, which makes us need to design algorithms to identify its related genes for the study of pulmonary injury. In this paper, we fused Adaboost with Relevance Vector Machine (RVM) to design a pipeline for the identification of pulmonary injury-related genes.

METHODS

Workflow

Firstly, we obtained 193 genes which are reported to be related to pulmonary injury in DisGeNET (Piñero et al., 2020). Then, we extracted the gene interaction information from String database (Szklarczyk et al., 2021). We extracted expression information of these genes from biogps database (Wu et al., 2016). Then, we did feature combination for each gene. Finally, we identified pulmonary injury-related genes by AdaRVM.

The whole workflow is shown in **Figure 1**.

Feature Combination

We obtained the expression of all genes in different tissues from Biogps database. We totally obtained gene expression in 13 tissues. For each gene, the feature of it would be:

$$Gene_1 = [E_1, E_2, ..., E_{13}] \quad (1)$$

E_i represents the expression of gene in ith tissue.

Then, we obtained gene interaction information from string database. We obtained 193 pulmonary injury-related genes in DisGeNET and found 1,019 genes can interact with at least one of them. The interactions between them could be the second feature of genes.

$$Gene_1 = [G_1, G_2, ..., G_{1212}] \quad (2)$$

G_i represents whether ith gene can interact with gene 1. If gene1 can interact with ith gene, $G_i = 1$, otherwise, $G_i = 0$.

Since the dimension of interaction feature for each gene is very high, we used Principal components analysis (PCA) to reduce the dimension. We put all genes together and used PCA to remain the

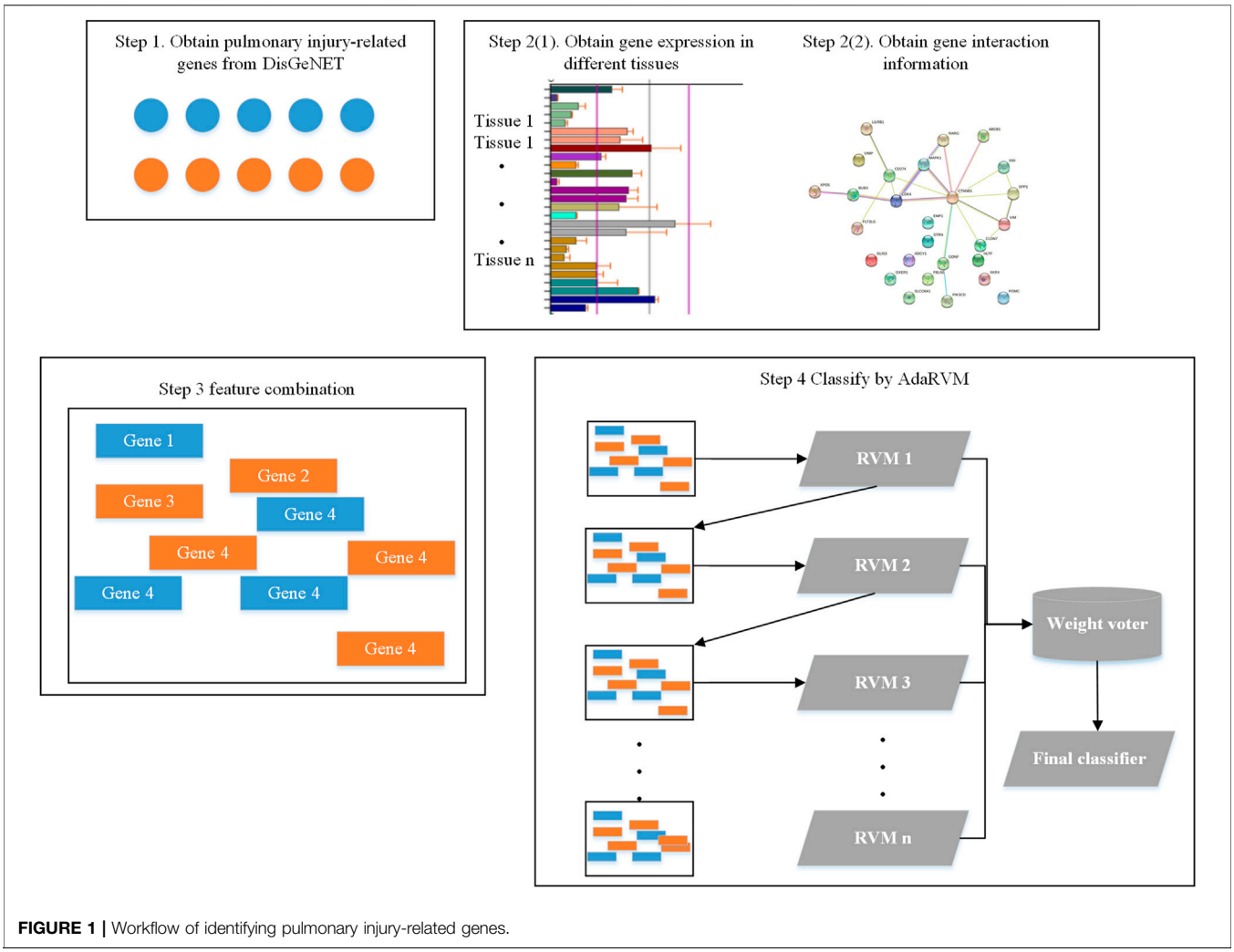


FIGURE 1 | Workflow of identifying pulmonary injury-related genes.

99% information. After PCA, the dimension of interaction feature for each gene is 237.

Finally, we combined interaction feature and expression feature together to get the final gene feature.

RVM

The RVM model can give the conditional probability distribution of the target variable t given an input vector x:

$$p(t|x,w,\beta) = N(t|y(x),\beta^{-1}) \tag{3}$$

$\beta = \sigma^{-2}$ is the noise accuracy.

The mean value is given by the linear model:

$$y(x) = \sum_{i=1}^{M} w_i \varphi_i(x) = w^T \varphi(x) \tag{4}$$

The basis function is given by the kernel function. Each data point in the training set is associated with a kernel function. The general expression is:

$$y(x) = \sum_{n=1}^{N} w_n k(x, x_n) + b \tag{5}$$

The formula here is similar to the predictive model formula in support vector regression.

Suppose the N observation data of the input vector x are aggregated into a data matrix X, the n th row is x_n^T, the objective value is $t = (t_1, \ldots, t_N)^T$, the likelihood function is:

$$p(t|X,w,\beta) = \prod_{n=1}^{N} p(t_n|x_n,w,\beta) \tag{6}$$

The prior distribution on the parameter vector w is introduced, and the Gaussian prior of zero mean is considered. However, the key difference of RVM is that a separate super parameter w_i is introduced for each weight parameter a_i, rather than a shared super parameter. The prior form of weight is:

$$p(w|a) = \prod_{i=1}^{M} N\left(w_i|0, a_i^{-1}\right) \tag{7}$$

a_i represents the accuracy of the corresponding parameter w_i.

When the model evidence for these hyperparameters is maximized, most of them tend to infinity, and the posterior probability distribution of the corresponding weight parameters is concentrated near 0. Then the associated basis function will have no effect on the prediction of the model, so it can be removed to form sparsity.

The posterior probability distribution of weights is still Gaussian, in the form:

$$p(w|t, X, a, \beta) = N\left(w|m, \sum\right) \tag{8}$$

where $m = \beta \sum \varphi^T t, \sum = (aI + \beta\varphi^T\varphi)^{-1}$.

After convergence, it can be found that part of the hyperparameter a_i values tend to infinity, and the weight parameter w_i corresponding to these hyperparameters has a posterior probability distribution with a mean and variance of 0. Then the corresponding parameters and basis function $\varphi_i(x)$ are erased from the model.

AdaRVM

Adaboost generally uses a single-layer decision tree as its weak classifier. The single-layer decision tree is the most simplified version of the decision tree. It has only one decision point. That is to say, if the training data has multi-dimensional features, the single-layer decision tree can only select one of the one-dimensional features to make decisions. In our method, AdaRVM uses RVM as a weak classifier.

There are four steps to achieve Adaboost. Suppose our training set sample is:

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \tag{9}$$

The output weight of the kth weak learner of the training set is:

$$D(k) = (w_{k1}, w_{k2}, \ldots, w_{km}) w_{1i} = \frac{1}{m} \tag{10}$$

Where m is the number of training samples.

The first step is to get the weighted error rate of the kth weak classifier $G_k(x)$ on the training set.

$$e_k = p(G_k(x_i) \neq y_i) = \sum_{i=1}^{m} w_{ki} I(G_k(x_i) \neq y_i) \tag{11}$$

The second step is to obtain the weight coefficient of the weak learner. For the binary classification problem, the weight coefficient of the kth weak classifier $G_k(x)$ is

$$a_k = \frac{1}{2}\log\frac{1 - e_k}{e_k} \tag{12}$$

It can be seen from the above formula that if the classification error rate e_k is larger, the corresponding weak classifier weight coefficient a_k is smaller. In other words, a weak classifier with a small error rate has a larger weight coefficient.

The third step is to update the sample weight D. Assuming that the weight coefficient of the sample set of the kth weak classifier is

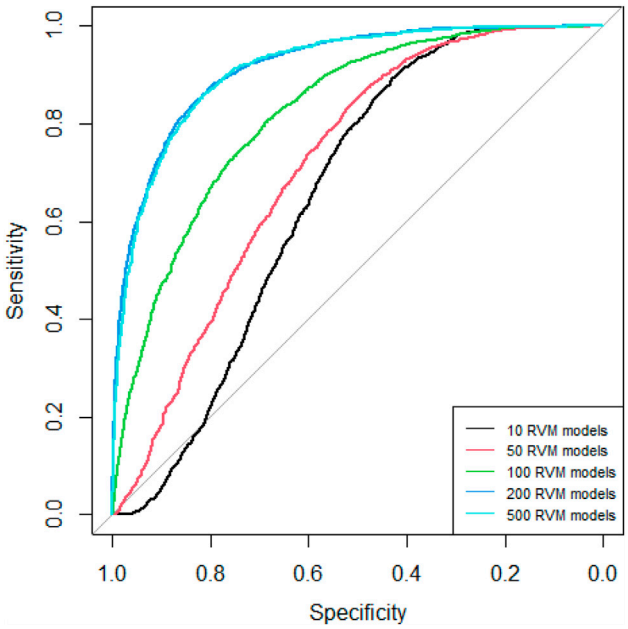


FIGURE 2 | ROC curves of different RVM models.

$D(k) = (w_{k1}, w_{k2}, \ldots, w_{km})\ w_{1i} = \frac{1}{m}$, the weight coefficient of the sample set of the corresponding k+1 weak classifier is:

$$w_{k+1,i} = \frac{w_{ki}}{Z_K} \exp\left(-a_k y_i G_k(x_i)\right) \tag{13}$$

Where Z_K is normalization factor

$$Z_k = \sum_{i=1}^{m} w_{ki} \exp\left(-a_k y_i G_k(x_i)\right) \tag{14}$$

The last step is weighted voting, and the final strong classifier is:

$$f(x) = gn\left(\sum_{k=1}^{K} a_k G_k(x)\right) \tag{15}$$

RESULTS

Method Evaluation

Since AdaRVM needs to construct multiple RVM models, the number of RVM models determines the accuracy of AdaRVM. Although in theory, the more the number of RVM models, the more accurate AdaRVM. But the huge number of RVM will cause the training time to rise sharply, and after reaching a certain amount, the accuracy of AdaRVM will be stable. Therefore, we need to try to build AdaRVM with multiple numbers of RVM models to know the most suitable number of models. We used 10, 50, 100, 200, 500 RVM models to construct AdaRVM respectively. We used 90% data to train the model and used the rest 10% data to test the performance of different AdaRVMs. The AUC and AUPR of these models are shown in **Figures 2**, **3**.

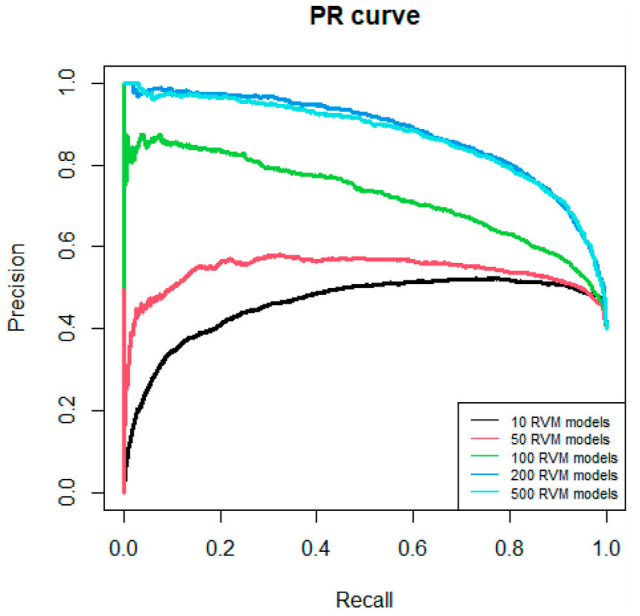


FIGURE 3 | PR curves of different RVM models.

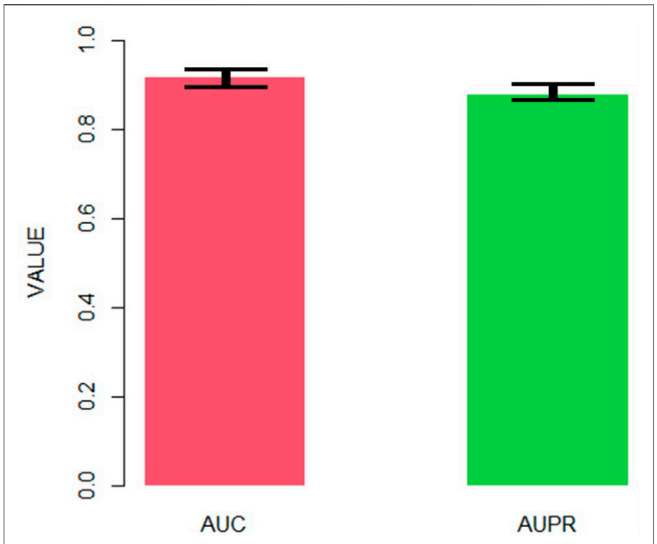


FIGURE 4 | The AUC and AUPR of AdaRVM in 10-cross validation.

As we can see in **Figures 2**, **3**, as the number of RVM models increases, the AUC and AUPR of AdaRVM have improved significantly. But as the number of RVM reached 200, both AUC and AUPR stabilized and remained at 0.91 and 0.88. Therefore, we used 200 RVM models to construct the final AdaRVM.

After obtaining the final AdaRVM, we need to test the performance of it. 10-cross validation has been applied to test the stability of AdaRVM.

Figure 4 shows the AUC and AUPR of AdaRVM in 10-cross validation. Since we tested AdaRVM 10 times, the mean AUC is 0.921 and the mean AUPR is 0.882. The standard deviation is 0.006 and 0.005, respectively.

The experiments showed the high accuracy of AdaRVM. At the same time, in 10 cross-validation experiments, AdaRVM has a small variance of AUC and AUPR, indicating that the performance of the method is very stable, and it is a reliable method for predicting pulmonary injury-related genes.

Revealing the Functions of Cavin-2

AdaRVM found that Cavin-2 has strong potential to be related to pulmonary injury. To verify our method and explore the function of cavin-2, we used Summary data level Mendelian Randomization (SMR) to integrate GWAS, eQTL, and mQTL to deeply reveal the role of cavin-2 in lung injury.

Lung function indicators contain the lung volume, the forced vital capacity (FVC), the forced expiratory volume in one second (FEV1), and the ratio of FEV1 to FVC (FEV1/FVC), and the peak expiratory flow (PEF). Lung volume represents the volume of gas passing through the lungs at different stages of the respiratory cycle. FVC refers to the maximum amount of air that can be exhaled as soon as possible after inhaling as much as possible. FEV1 is the volume of expiratory breath in the first second after the maximum inhalation. FEV1/FVC is a common indicator to judge airway obstruction and can reflect the type and degree of ventilation disorders. PEF is the maximum flow when exhaling forcefully. These indicators are important basis for the diagnosis of lung diseases. Additionally, when lung damage is caused by a variety of reasons, the interstitium secretes collagen for repair. The excessive repair can lead to the formation of cystic pulmonary fibrosis. Based on the above, we selected GWAS data related to the six lung function injury traits (lung volume, FVC, FEV1, FEV1/FVC, PEF, and CF) for the following analysis.

Four datasets of GWAS data were derived from the same study results (Shrine et al., 2019). These four data individually described FVC, FEV1, FEV1/FVC, and PEF traits. The acquisition of CF GWAS data depends on the BioBank Japan (BBJ) Project, a prospective BioBank that collected different kinds of biological samples from 12 medical institutions in Japan (Sakaue et al., 2021).

To focus the SNPs localized in the region of Cavin-2 gene, we searched the location of Cavin-2 gene in GeneCards website. The site contains detailed information about all type of genes such as protein-coding genes, RNA genes, gene clusters. The gene information showed in this website includes gene aliases, related position information in the genome, gene function, gene mapping, gene involved in pathways, and so on. According to the search results, Cavin-2 gene was located at 192,699,028- 192,711,981 on chromosome 2. Thus, we eliminated SNPs that did not exist in that range. Finally, six GWAS subsets of Cavin-2 gene SNPs were generated. Then we utilized these subsets for the next step of association analysis with eQTL and mQTL.

In our study, we employed four eQTL data including three blood eQTL and one lung tissue eQTL, and one mQTL data. Two datasets of eQTL were downloaded from Genotype-Tissue Expression (GTEx) website. The GTEx project collected and analyzed samples from 53 non-diseased tissues across nearly

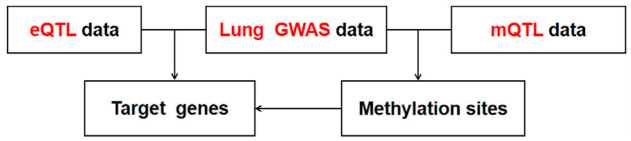


FIGURE 5 | The process of GWAS, eQTL, and mQTL conjoint analysis.

1,000 individuals to investigate tissue-specific gene expression patterns. Since we studied lung function traits, we selected one lung eQTL data, and one blood eQTL data. The remaining two eQTL datasets were obtained by RNA sequencing of the whole peripheral blood of 2,116 healthy adults from four Dutch cohort (CODAM, LLD, LLS, and RS). One eQTL data contains 23,060 gene level effects, while the other dataset contain 21,888 exon level effects (Zhernakova et al., 2017). Similarly, the mQTL dataset was downloaded from the same browser. This study analyzed the data from the five Dutch biobank study (CODAM, LLD, LLS, NTR, and RS) of 3,841 whole blood samples (Bonder et al., 2017). The mQTL dataset contained 272,037 independent cis-mQTL effects. The whole datasets used in this paper are listed in **Table 1**.

First, we conducted co-localization analysis on two datasets of GWAS and eQTL to find overlapping SNPs and generate a new data containing these SNPs (**Figure 5**). Then the Z-Score method was used to standardized GWAS and eQTL data for the subsequent analysis. Accoding to the Z-Score formula (① c = qnorm (1 − p÷2); Beta > 0, c = Z Score, Beta < 0; ② c = −Z Score; Z Score = Beta÷SE), the Z value could calculated through using any two of the three values. Beta/OR value refers to the value of regression coefficient, while SE value refers to the standard error of the regression coefficient. Z values could compute the T_{SMR} value using **Eq. 16**, and then calculated the *p* values (P_{SMR}) corresponding to each variation by chi-square test.

$$T_{SMR} = \frac{Z_{GWAS}^2 * Z_{eQTL}^2}{Z_{GWAS}^2 + Z_{eQTL}^2} \tag{16}$$

P_{SMR} value of 0.05 was taken as the statistically significant threshold. If the P_{SMR} value of some variations is less than 0.05, these variations may affect the pulmonary phenotype by affecting Cavin-2 gene expression. To determine the relationship between these variation points and the classical signal pathways related to lung injury, we performed further signal pathway analysis to verify whether the changes of Cavin-2 gene expression mediated by these variation points could interact with the components of ERK1/2, AKT, and STAT3 classical signal pathways related to lung injury, revealing the function of Cavin-2 gene in lung injury and further confirming the accuracy of our analysis results.

We first conducted the conjoint analysis between the pair of GWAS and eQTL data, and a total of 24 times SMR analyses were performed. First of all, we need to find the same mutation locus between GWAS dataset and eQTL dataset and then combine the information of the mutation locus in the two datasets to generate a new SMR dataset. As the GWAS data subset had a very small list of SNPs, the overlapping SNPs found were much rarer in the new SMR datasets. No more than 10 SNPs were found in all SMR datasets. There were 3 overlapping SNPs in the dataset, which was the dataset containing the most overlapping SNPs. These data sets generally contain only one overlapping susceptible site information.

In total, we discovered 3 susceptible loci in 24 SMR datasets. These SNPs were rs111946466, rs1128965, and rs6718527. These three sites were repeated in these SMR datasets. Half of the SMR datasets contained the rs1128965 information, suggesting this SNPs may closely associated with the Cavin-2 gene expression. Additionally, the rs6718527 site may exhibit a tissue-specific as it is identified only by three SMR datasets containing eQTL information for lung.

Then, we performed SMR analysis on these datasets to verify whether these three loci affect lung functional phenotypes by influencing Cavin-2 gene expression. The Z value of GWAS and eQTL in the dataset was used to calculate the T_{SMR} value of the mutation locus. Then, the Chi-square test of T_{SMR} was performed to obtain the P_{SMR} value. The *p* value was compared with the threshold value of 0.05. It was found that the four data sets had the susceptible site rs1128965. Besides, the P_{SMR} value obtained after analyzing the susceptible site was below 0.05. These four data sets are consistent with the lung function impairment phenotype of FVC and PEF. This suggests that the mutation locus can down-regulate the expression of the Cavin-2 gene, resulting in abnormal lung function of FVC and PEF.

The SMR dataset of the 6 mQTL also had SNPs information of three susceptible loci. The three SNPs include rs12477095, rs11681727, and rs58245883. These three susceptibility loci appeared in all six datasets. However, only the SMR dataset of FEV_1 trait identified rs12477095 as significant through SMR analysis. This suggests that this site may perform DNA methylation of Cavin-2, affecting the FEV_1 phenotype associated with lung function.

The Cavin-2 signal pathway was analyzed to further explain the association between the changes in gene expression mediated by Cavin-2-point mutation and ERK1/2, PI3K-AKT signal pathway, and STAT3. The extracellular signal-regulated kinase (ERK) pathway is activated by various extracellular factors, such as growth factors and hormones, to participate in cell proliferation and differentiation, stress response, etc. SDPR is a metastasis inhibitor that inhibits epithelial-mesenchymal transformation (EMT) and migration, promoting apoptosis by interacting with ERK analysis to inhibit the extracellular ERK pathway in breast tumors. Therefore, SDPR may inhibit the ERK pathway to regulate lung injury.

Additionally, AKT-related signaling pathways help regulate cell functions such as cell survival, proliferation, differentiation, and migration. They are also important compensation mechanisms for the body in responding to various harmful stimuli. The PI3K/AKT/eNOS signaling pathway is a signal transduction pathway that plays an essential regulatory role in endothelial cells. The endothelial nitric oxide synthase (eNOS) is an important downstream target of AKT and helps in regulating vascular growth and endothelial function. Cavin-2 regulates nitric oxide (NO) production in endothelial cells by controlling eNOS

TABLE 1 | The new dataset list for SMR analysis. Thirty SMR datasets were created for this analysis.

GWAS	eQTL	SMR	GWAS	eQTL	SMR
FEV_1	GTEx_Blood	FEV_1_GTEx_Blood	FVC	GTEx_Blood	FVC_GTEx_Blood
	GTEx_Lung	FEV_1_GTEx_Lung		GTEx_Lung	FVC_GTEx_Lung
	GENE_Blood	FEV_1_GENE_Blood		GENE_Blood	FVC_GENE_Blood
	EXON_Blood	FEV_1_EXON_Blood		EXON_Blood	FVC_EXON_Blood
FEV_1/FVC	GTEx_Blood	FEV_1/FVC_GTEx_Blood	PEF	GTEx_Blood	PEF_GTEx_Blood
	GTEx_Lung	FEV_1/FVC_GTEx_Lung		GTEx_Lung	PEF_GTEx_Lung
	GENE_Blood	FEV_1/FVC_GENE_Blood		GENE_Blood	PEF_GENE_Blood
	EXON_Blood	FEV_1/FVC_EXON_Blood		EXON_Blood	PEF_EXON_Blood
Lung volume	GTEx_Blood	LungVolume_GTEx_Blood	CF	GTEx_Blood	CF_GTEx_Blood
	GTEx_Lung	LungVolume_GTEx_Lung		GTEx_Lung	CF_GTEx_Lung
	GENE_Blood	LungVolume_GENE_Blood		GENE_Blood	CF_GENE_Blood
	EXON_Blood	LungVolume_EXON_Blood		EXON_Blood	CF_EXON_Blood

stability and activity. Cavin-2 can also help regulate lung injury by regulating eNOS associated with the PI3K-Akt signal pathway of lung injury.

As a stress-induced inflammatory signaling pathway, signal transducers and activators of transcription 3 (STAT3) play an essential role in regulating various biological behaviors of a normal organism, such as immune response, tissue repair, cell growth, etc. STAT3 is an indispensable key molecule in the tumorigenesis and tumor-associated inflammation process, which is promoted by chronic inflammation and abnormal activation of STAT3. This causes various diseases since signal transduction and STAT3 help protect respiratory epithelial cells during injury. Cavin-2 is abundant in lung epithelial cells and is closely associated with cell signal transduction. Cavin-2 helps regulate the function of the lung during an injury by affecting signal transduction in the STAT3 signaling pathway.

CONCLUSION

Gene mutations play an important role in the occurrence of pulmonary injury, and many mutated genes and genomes have been identified. With the widespread application of machine learning methods in the biomedical field, a large number of researchers use such methods to study disease-related biomolecules on a large scale. But no method has been developed specifically to study pulmonary injury-related genes. In this paper, we developed AdaRVM which used gene expression and gene interaction features to identify pulmonary injury-related genes. We extracted the expression of genes in 13 tissues and 1,212 genes interaction. Due to the high dimension of features, PCA was applied to reduce the dimension and remain the key information. Then, we have determined the method of constructing AdaRVM through 200 RVM models through many experimental attempts. Through the idea of Adaboost, the output results of 200 RVM models are aggregated to obtain the final prediction result.

The 10-cross validation showed the high precision of AdaRVM, with AUC of 0.921 and AUPR of 0.882, which proves the effectiveness of our method. To verify the accuracy of our results, we selected Cavin-2 gene which is predicted as pulmonary injury-related genes by AdaRVM to do case study. We investigated the relationship between the Cavin-2 gene and three signaling pathways associated with lung injury. ERK is a classic pathway of the mitogen-activated protein kinase (MAPK) signal transduction system. Typically, ERK is located in the cytoplasm and is activated only after phosphorylation. ERK regulates the activity of some transcription factors, such as STATs, Jun, Fos, and ATF2, through phosphorylation. These transcription factors further regulate the transcription of their respective target genes, leading to changes in the expression or activity of specific proteins, ultimately regulating cell function, and metabolism. ERK also regulates various activities such as cell growth, apoptosis, and embryogenesis. The ERK1/2 signaling pathway is essential in the EMT process and helps regulate cancer metastasis (Sheng et al., 2017). Typically, EMT occurs in epithelial cells, and various signals can induce it. Members of the transforming growth factor β family (TGFβ) are some important cytokines that can induce the transformation process of EMT. Besides, they can treat fibrotic diseases and tumor diseases. TGFβ signaling can induce EMT conversion through several different signaling mechanisms. The EMT induced by the transforming growth factor (TGF-β1) is important pathogenesis of silicosis fibrosis. TGF-β1 can induce EMT by activating the downstream ERK1/2 pathway and promoting silicosis fibrosis development. SDPR can act as a metastasis inhibitor and regulate the ERK pathway to inhibit EMT. Therefore, SDPR may inhibit the ERK pathway to regulate lung injury. The eNOS is a common downstream substrate of AKT, which is activated to promote eNOS protein expression. The eNOS is a kind of nitric oxide synthase isoenzyme, primarily distributed in vascular endothelium and catalyzes l-arginine to produce NO, which vasodilates and increases vascular permeability.

PI3K/AKT/eNOS is an important signal transduction pathway in endothelial cells. This pathway regulates inflammatory responses and the expression of inflammatory mediators, such as eNOS, primarily through downstream proteins of AKT (Everaert et al., 2010). Inactivation of eNOS is associated with oxidative stress, inflammatory response, and adhesion to the vascular endothelium. Endothelial dysfunction helps in atherosclerosis and ischemia-reperfusion injury. Endothelial dysfunction is characterized by reduced eNOS-induced NO bioavailability, which dilates blood vessels and inhibits inflammation. Cavin-2 also regulates NO production in endothelial cells by controlling the stability and

activity of eNOS, and cavin-2 knockdown cells produce much less NO than WT cells. The binding of estrogen, vascular endothelial growth factor, insulin, and other signal molecules with corresponding receptors on the cell surface can activate intracellular PI3K. Besides, the activated PI3K can generate PIP3 through phosphorylation. PIP3 binds to intracellular signals to promote PDK1 phosphorylation of AKT, activating the expression of its downstream target protein.

JAK-STAT comprises three parts: The receptor tyrosine kinase-associated receptor for receiving the signal, the tyrosine kinase JAK for transmitting the signal, and the transcription factor STAT for producing the effect (Johnson et al., 2018). Cytokines or other extracellular signal ligands on the cell membrane induce the aggregation of the corresponding receptors to form homologous or heterologous dimers. This brings the protein JAK together in the cytoplasm, leading to autophosphorylation. Activation of JAK leads to further phosphorylation of the tyrosine residues in the cytoplasmic region of the receptor and mediates STAT3 to bind to the phosphorylated tyrosine residues on the receptor through its Src homology 2 (SH2) domain, phosphorylating the tyrosine residues of STAT3. After phosphorylation, STAT3 is transferred to the nucleus in the form of a homologous or heterologous dimer and binds to the DNA promoter of the target gene to activate the transcription of the target gene. Interstitial lung disease (ILD) includes fibrotic lung diseases characterized by cell proliferation, interstitial inflammation, and fibrosis. The JAK/STAT pathway is activated in response to the interaction of many pro-fibrotic/pro-inflammatory cytokines such as IL-6, IL-11, and IL-13, elevated in various ILD. Similarly, several overexpressed growth factors in ILD, such as platelet-derived growth factor (PDGF), TGF-β1, and fibroblast growth factor (FGF), activate JAK/STAT via classical or non-classical signaling pathways. This suggests that JAK/STAT plays a dominant role in ILD.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CL wrote the article. JH designed the algorithm. HT performed the experiments. BL explained the mechanism of our results. XZ designed the whole work.

REFERENCES

Asefpour Vakilian, K. (2020). Machine Learning Improves Our Knowledge about miRNA Functions towards Plant Abiotic Stresses. *Sci. Rep.* 10 (1), 3041. doi:10.1038/s41598-020-59981-6

Bonder, M. J., Luijk, R., Luijk, R., Zhernakova, D. V., Moed, M., Deelen, P., et al. (2017). Disease Variants Alter Transcription Factor Levels and Methylation of Their Binding Sites. *Nat. Genet.* 49 (1), 131–138. doi:10.1038/ng.3721

Chettimada, S., Yang, J., Moon, H.-g., and Jin, Y. (2015). Caveolae, Caveolin-1 and Cavin-1: Emerging Roles in Pulmonary Hypertension. *Wjr* 5 (2), 126. doi:10.5320/wjr.v5.i2.126

Ding, H., Karunanithi, M., Ireland, D., McCarthy, L., Hakim, R., Phillips, K., et al. (2019). Evaluation of an Innovative mobile Health Programme for the Self-Management of Chronic Obstructive Pulmonary Disease (MH-COPD): Protocol of a Randomised Controlled Trial. *BMJ open* 9 (4), e025381. doi:10.1136/bmjopen-2018-025381

Duprez, D. A., and Jacobs, D. R. (2018). *Lung Function Decline and Increased Cardiovascular Risk: Quo Vadis*. Washington, DC: American College of Cardiology Foundation.

Everaert, B. R., Van Craenenbroeck, E. M., Hoymans, V. Y., Haine, S. E., Van Nassauw, L., Conraads, V. M., et al. (2010). Current perspective of pathophysiological and interventional effects on endothelial progenitor cell biology: focus on PI3K/AKT/eNOS pathway. *International journal of cardiology* 144 (3), 350–366. doi:10.1016/j.ijcard.2010.04.018

Hansen, C. G., Shvets, E., Howard, G., Riento, K., and Nichols, B. J. (2013). Deletion of Cavin Genes Reveals Tissue-specific Mechanisms for Morphogenesis of Endothelial Caveolae [J]. *Nat. Commun.* 4 (1), 1–13. doi:10.1038/ncomms2808

Johnson, D. E., O'Keefe, R. A., and Grandis, J. R. (2018). Targeting the IL-6/JAK/STAT3 signalling axis in cancer. Nature reviews. *Clinical oncology* 15 (4), 234–248. doi:10.1038/nrclinonc.2018.8

Li, H., Pouladi, N., Achour, I., Gardeux, V., Li, J., Li, Q., et al. (2015). eQTL Networks Unveil Enriched mRNA Master Integrators Downstream of Complex Disease-Associated SNPs. *J. Biomed. Inform.* 58, 226–234. doi:10.1016/j.jbi.2015.10.010

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine Learning for Detecting Gene-Gene Interactions. *Appl. Bioinformatics* 5 (2), 77–88. doi:10.2165/00822942-200605020-00002

McRae, A. F., Marioni, R. E., Shah, S., Yang, J., Powell, J. E., Harris, S. E., et al. (2018). Identification of 55,000 Replicated DNA Methylation QTL. *Sci. Rep.* 8 (1), 17605–17609. doi:10.1038/s41598-018-35871-w

Murata, T., Lin, M. I., Stan, R. V., Bauer, P. M., Yu, J., and Sessa, W. C. (2007). Genetic Evidence Supporting Caveolae Microdomain Regulation of Calcium Entry in Endothelial Cells. *J. Biol. Chem.* 282 (22), 16631–16643. doi:10.1074/jbc.m607948200

Noé, F., De Fabritiis, G., and Clementi, C. (2020). Machine Learning for Protein Folding and Dynamics. *Curr. Opin. Struct. Biol.* 60, 77–84. doi:10.1016/j.sbi.2019.12.005

Ozturk, S., Papageorgis, P., Wong, C. K., Lambert, A. W., Abdolmaleky, H. M., Thiagalingam, A., et al. (2016). SDPR Functions as a Metastasis Suppressor in Breast Cancer by Promoting Apoptosis. *Proc. Natl. Acad. Sci. USA* 113 (3), 638–643. doi:10.1073/pnas.1514663113

Parton, R. G., Tillu, V. A., and Collins, B. M. (2018). Caveolae. *Curr. Biol.* 28 (8), R402–R405. doi:10.1016/j.cub.2017.11.075

Parton, R. G., McMahon, K.-A., and Wu, Y. (2020). Caveolae: Formation, Dynamics, and Function. *Curr. Opin. Cel. Biol.* 65, 8–16. doi:10.1016/j.ceb.2020.02.001

Peng, J., and Zhao, T. (2020). Reduction in TOM1 Expression Exacerbates Alzheimer's Disease. *Proc. Natl. Acad. Sci. USA* 117 (8), 3915–3916. doi:10.1073/pnas.1917589117

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48 (D1), D845–D855. doi:10.1093/nar/gkz1021

Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., et al. (2021). A Cross-Population Atlas of Genetic Associations for 220 Human Phenotypes. *Nat. Genet.* 53 (10), 1415–1424. doi:10.1038/s41588-021-00931-x

Sheng, W., Chen, C., Dong, M., Wang, G., Zhou, J., Song, H., et al. (2017). Calreticulin promotes EGF-induced EMT in pancreatic cancer cells via Integrin/EGFR-ERK/MAPK signaling pathway. *Cell death and disease* 8 (10), e3147. doi:10.1038/cddis.2017.547

Shrine, N., Guyatt, A. L., Guyatt, A. L., Erzurumluoglu, A. M., Jackson, V. E., Hobbs, B. D., et al. (2019). New Genetic Signals for Lung Function Highlight Pathways and Chronic Obstructive Pulmonary Disease Associations across Multiple Ancestries. *Nat. Genet.* 51 (3), 481–493. doi:10.1038/s41588-018-0321-7

Suzuki, T., Chow, C. W., and Downey, G. P. (2008). Role of Innate Immune Cells and Their Products in Lung Immunopathology. *Int. J. Biochem. Cel. Biol.* 40 (6-7), 1348–1361. doi:10.1016/j.biocel.2008.01.003

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets. *Nucleic Acids Res.* 49 (D1), D605–D612. doi:10.1093/nar/gkaa1074

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying Drug–Target Interactions Based on Graph Convolutional Network and Deep Neural Network. *Brief. Bioinform.* 22 (2), 2141–2150. doi:10.1093/bib/bbaa044

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional Mapping and Annotation of Genetic Associations with FUMA. *Nat. Commun.* 8 (1), 1826. doi:10.1038/s41467-017-01261-5

Wu, C., Jin, X., Tsueng, G., Afrasiabi, C., and Su, A. I. (2016). BioGPS: Building Your Own Mash-Up of Gene Annotations and Expression Profiles. *Nucleic Acids Res.* 44 (D1), D313–D316. doi:10.1093/nar/gkv1104

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: A Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020). MRTFB Regulates the Expression of NOMO1 in colon. *Proc. Natl. Acad. Sci. USA* 117 (14), 7568–7569. doi:10.1073/pnas.2000499117

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2020). Identifying Protein Biomarkers in Blood for Alzheimer's Disease. *Front. Cel Dev. Biol.* 8, 472. doi:10.3389/fcell.2020.00472

Zhao, T., Hu, Y., and Cheng, L. (2020). Deep-DRM: A Computational Method for Identifying Disease-Related Metabolites Based on Graph Deep Learning Approaches. *Brief Bioinform.*, 22 (4), bbaa212. doi:10.1093/bib/bbaa212

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021). Prediction and Collection of Protein–Metabolite Interactions. *Brief. Bioinform.* 22 (5), bbab014. doi:10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021). SC2disease: a Manually Curated Database of Single-Cell Transcriptome for Human Diseases. *Nucleic Acids Res.* 49 (D1), D1413–D1419. doi:10.1093/nar/gkaa838

Zhernakova, D. V., Deelen, P., Vermaat, M., Van Iterson, M., Van Galen, M., Arindrarto, W., et al. (2017). Identification of Context-dependent Expression Quantitative Trait Loci in Whole Blood. *Nat. Genet.* 49 (1), 139–145. doi:10.1038/ng.3737

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Li, Huang, Tang, Liu and Zhou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 01 February 2022
doi: 10.3389/fcell.2022.845622

The Characterization of Structure and Prediction for Aquaporin in Tumour Progression by Machine Learning

Zheng Chen[1,2], *Shihu Jiao*[3], *Da Zhao*[1,2], *Quan Zou*[2,3], *Lei Xu*[4], *Lijun Zhang*[1]* and *Xi Su*[5]*

[1]School of Applied Chemistry and Biological Technology, Shenzhen Polytechnic, Shenzhen, China, [2]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [3]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [4]School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China, [5]Foshan Maternal and Child Health Hospital, Foshan, China

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Hua Tang,
Southwest Medical University, China
Xiaofeng Song,
Nanjing University of Aeronautics and Astronautics, China

***Correspondence:**
Lijun Zhang
c7zlj@szpt.edu.cn
Xi Su
xisu_fsfy@163.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *30 December 2021*
Accepted: *17 January 2022*
Published: *01 February 2022*

Citation:
Chen Z, Jiao S, Zhao D, Zou Q, Xu L, Zhang L and Su X (2022) The Characterization of Structure and Prediction for Aquaporin in Tumour Progression by Machine Learning. Front. Cell Dev. Biol. 10:845622. doi: 10.3389/fcell.2022.845622

Recurrence and new cases of cancer constitute a challenging human health problem. Aquaporins (AQPs) can be expressed in many types of tumours, including the brain, breast, pancreas, colon, skin, ovaries, and lungs, and the histological grade of cancer is positively correlated with AQP expression. Therefore, the identification of aquaporins is an area to explore. Computational tools play an important role in aquaporin identification. In this research, we propose reliable, accurate and automated sequence predictor iAQPs-RF to identify AQPs. In this study, the feature extraction method was 188D (global protein sequence descriptor, GPSD). Six common classifiers, including random forest (RF), NaiveBayes (NB), support vector machine (SVM), XGBoost, logistic regression (LR) and decision tree (DT), were used for AQP classification. The classification results show that the random forest (RF) algorithm is the most suitable machine learning algorithm, and the accuracy was 97.689%. Analysis of Variance (ANOVA) was used to analyse these characteristics. Feature rank based on the ANOVA method and IFS strategy was applied to search for the optimal features. The classification results suggest that the 26th feature (neutral/hydrophobic) and 21st feature (hydrophobic) are the two most powerful and informative features that distinguish AQPs from non-AQPs. Previous studies reported that plasma membrane proteins have hydrophobic characteristics. Aquaporin subcellular localization prediction showed that all aquaporins were plasma membrane proteins with highly conserved transmembrane structures. In addition, the 3D structure of aquaporins was consistent with the localization results. Therefore, these studies confirmed that aquaporins possess hydrophobic properties. Although aquaporins are highly conserved transmembrane structures, the phylogenetic tree shows the diversity of aquaporins during evolution. The PCA showed that positive and negative samples were well separated by 54D features, indicating that the 54D feature can effectively classify aquaporins. The online prediction server is accessible at http://lab.malab.cn/~acy/iAQP.

Keywords: cancer, random forest, anova, 3D structure, machine learning

INTRODUCTION

Water, as one of the most widely existing molecules, is the basic requirement for the development of organisms. Aquaporins (AQPs) are a large and evolutionarily conserved family of proteins that facilitate water absorption and flow across cytoplasmic compartments and cell membranes in microorganisms, animals, and plants. From a previous study, aquaporins, as water channel proteins, not only take part in water molecule transport but also respond to other small molecule transport, such as glycerol, urea, ammonia, and CO_2, which help those molecules cross cell membranes (Preston et al., 1992; Ma et al., 1997; Agre et al., 2002; Nielsen et al., 2002; Rojek et al., 2008). In the aquaporin family, some aquaporins are primarily water selective, such as AQP1, AQP2, AQP4, AQP5 and AQP8, while other parts of the aquaporins, such as AQP3, AQP7, AQP9, and AQP10, transport water, glycerol and other small solutes (Verkman, 2005). Aquaporins are small highly conserved membrane proteins that can selectively promote water molecule transportation through the cell membrane. Aquaporins (AQPs), with a molecular weight of 28 kDa, were first found in the membrane of human red blood cells (Agre et al., 2002). AQPs usually exist as tetramers; when water passes through these narrow channels, the conformation of AQPs can decide whether water passes through the cell membrane.

AQPs not only act as channels to take part in water and small molecule transport but are also widely related to a variety of pathophysiological statuses in cells. Evidence of AQPs in cell proliferation has aroused great interest in the research of AQPs in tumour progression (Levin and Verkman, 2006; Zhang et al., 2010; Jung et al., 2011; Nakahigashi et al., 2011; Di Giusto et al., 2012; Direito et al., 2016; De Ieso and Yool, 2018). At present, AQPs can be expressed in many types of tumours, including in the brain (Maugeri et al., 2016; Lan et al., 2017), breast (Jung et al., 2011), pancreas (Arsenijevic et al., 2019), colon (Nagaraju et al., 2016), skin (Hara-Chikuma and Verkman, 2008a), ovaries (Kasa et al., 2019) and lung (Chae et al., 2008). There was a positive correlation between the histological tumour grade and AQP expression, such as the expression of AQP4 in diffuse astrocytoma (Saadoun et al., 2002a; Kröger et al., 2004).

For colorectal cancer, the expression of AQP8 decreased (Fischer et al., 2001), while that of AQP1, AQP3 and AQP5 increased (Moon et al., 2003), indicating that AQPs can be expressed in tumours in humans. In general, AQP expression is upregulated in tumours. Therefore, many studies speculate that aquaporins allow water to penetrate, resulting in rapid tumour mass formation. In astrocytomas, the expression level of AQP4 is related to the amount of oedema but not to survival status (Saadoun et al., 2002a; Warth et al., 2007). Recent studies have indicated that AQPs, as prognostic markers, have a potential role in tumour-associated oedema because they participate in angiogenesis, tumour cell migration and proliferation (Saadoun et al., 2005a; Saadoun et al., 2005b; Hara-Chikuma and Verkman, 2006; Auguste et al., 2007; Hara-Chikuma and Verkman, 2008b). AQP1, AQP4 and AQP9 are expressed in brain tumours, and AQP4 expression increases with the severity of brain oedema (Saadoun et al., 2002a; Ding et al., 2010; Wang and Owler, 2011; Ding et al., 2013; Maugeri et al., 2016; Lan et al., 2017). In brain, lung, prostate and colon tumours, AQP1 with high expression participates in cell migration and tumour angiogenesis (Saadoun et al., 2002b; Saadoun et al., 2005b; Mobasheri et al., 2005; Kang et al., 2008). AQP3 has increased expression in ESCA, COAD, LUAD and LIHC (Marlar et al., 2017). AQP3 knockout mice can inhibit the development of skin tumours, and tumorigenesis can utilize ATP produced by AQP3-mediated glycerol transport (Hara-Chikuma and Verkman, 2008a). AQP5 is also related to the migration, metastasis, and poor prognosis of cancer cells in BRCA (breast cancer) (Jung et al., 2011; Lee et al., 2014; Jensen et al., 2016). AQP5-regulating miRNAs inhibit BRCA cell migration through exosome-mediated delivery (Park et al., 2020). Under exosome-mediated delivery, AQP5-regulated miRNAs inhibit BRCA cell migration (Park et al., 2020).

Aquaporins play a role in the development and prognosis of various cancers, so the machine learning recognition method of aquaporins is also one of the hot spots in cancer research. Machine learning methods are applied to establish a novel and efficient classification model of aquaporins and are helpful to accelerate the recognition of aquaporins. The amino acid sequence composition of the protein is considered to be a sequence feature of the protein (Tyagi et al., 2013).

There are two methods for protein classification methods, as follows: one is based on protein sequence information (Liu et al., 2020a; Zhang et al., 2021), and the other is based on protein structure features (Liu et al., 2019; Cai et al., 2020). The sequence-based protein classification method extracts features by using the amino acid composition, amino acid number and other sequence information of the protein sequence (Liu et al., 2014). These methods are efficient and useful in predicting a large number of protein sequence datasets (Lou et al., 2014). At present, there are various studies on the classification of protein sequences, such as using logistic regression and support vector machine (SVM) methods to predict DNA binding proteins (Shen and Zou, 2020; Liu et al., 2021a) by considering amino acid proportions, amino acid compositions, amino acid spatial asymmetric distributions and biological coding characteristics of evolutionary information (Szilágyi and Skolnick, 2006; Kumar et al., 2007). The protein classification method based on protein structure identifies proteins by using structure and sequence information (Liu et al., 2014). Previous studies have focused on positive electrostatic potential, protein surface, overall charge and positive patches (Shanahan et al., 2004; Bhardwaj et al., 2005), which have achieved excellent results. Under certain conditions, the prediction accuracy of three protein motifs (helix turning helix, helix hairpin helix, and helix loop helix) is 91.1%, which indicates that this method is efficient for protein determination (Cai et al., 2009).

In our work, to promote the rapid application of AQPs in cancer treatment, a powerful sequence-based analysis method to distinguish the AQPs and cross validation was applied for results demonstration (**Figure 1**). It is important to develop an effective model to predict AQPs. We propose a sequence-based AQP prediction model that performs stably on various classifiers. The AQP classification model uses the 188D feature extraction

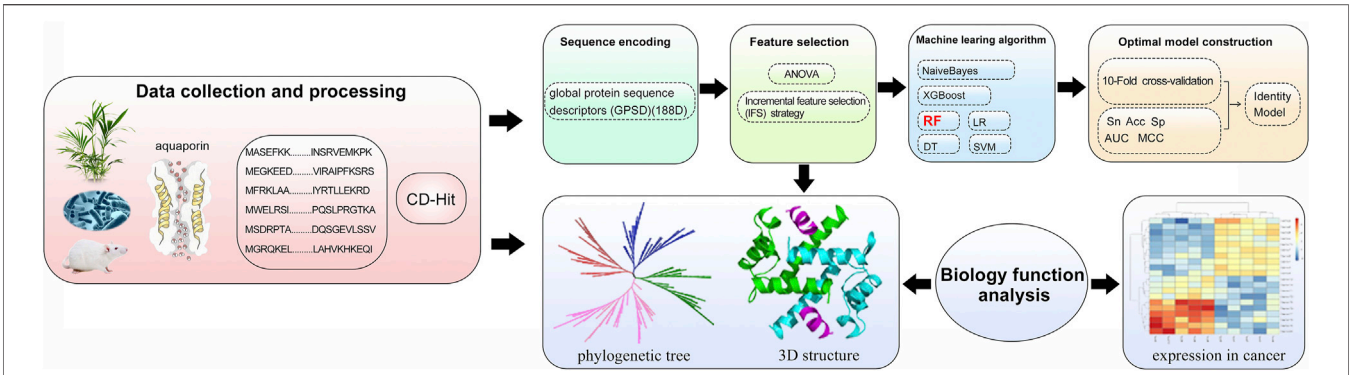


FIGURE 1 | The whole framework of the method iAQPs-RF to identify the aquaporins.

method, applies ANOVA to reduce the dimensionality, and uses different algorithms to optimize the AQP classification model. 188D is a characteristic of the frequency of continuous amino acid residues in proteins. ANOVA is used to prune features without affecting the accuracy of the predictor.

MATERIALS AND METHODS

Dataset

A high-quality dataset is essential for reliable and accurate predictor building (Su et al., 2021). Aquaporin was taken as the positive sample, and the protein sequence was collected from the protein database of the UniProt website (https://www.uniprot.org/) (Chen et al., 2016). Negative samples such as nonaquaporins were extracted from the Pfam database (http://pfam.xfam.org/). To ensure the reliability of the aquaporin dataset, we applied the following criteria to optimize the data: first, the sequences annotated as "prediction" were eliminated; second, we deleted the sequences of other protein fragments; through screening steps, 239 aquaporin sequences and 10,713 nonaquaporin sequences were obtained; third, the CD-HIT program (Fu et al., 2012) was used to eliminate redundant sequences and to avoid overestimating the prediction model (Zou et al., 2020). The cut-off of sequence identity is set to 90%. Finally, 151 aquaporins and 8,994 nonaquaporins were obtained to form the final dataset.

Features Extraction

One of the main factors for the performance accuracy of the prediction model is the quality of sample feature extraction. The prediction of the protein model mainly depends on the coding strategy of the protein sequence. According to the coding strategy of the protein sequence, the amino acid sequence can be transformed into a numerical vector (Liu et al., 2019; Muhammod et al., 2019; Zhu et al., 2019; Chen et al., 2020; Fu et al., 2020; Tang et al., 2020; Wang et al., 2020; Shao et al., 2021). In this paper, the global protein sequence descriptor (GPSD) method was used to represent the amino acid sequence. Global protein sequence descriptor (GPSD), known as 188 days method. This method mainly converts the sequence into a numerical vector according to the amino acid properties in the protein sequence and generates 188 features. These 188D features contain the information and properties of amino acid sequences [48,49]. According to the description of the GPSD method, the 188D features can be divided into two parts. The first part is the composition of amino acids. The first 20D features were obtained by calculating the frequency of amino acids in the protein sequence. The second part is to calculate the physicochemical properties of amino acids, which constitute 168 characteristics. Previous studies have provided detailed information on the eight physicochemical properties of amino acids (Lin et al., 2013; Liu et al., 2018; Li et al., 2019a). The protein sequence was encoded by CTD (C: composition, t: transition, D: distribution) mode to generate 21D features. Three groups were generated for 20 amino acids for each property. C is the occurrence frequencies (1 × 3D = 3D). T is the transition frequency (1 × 3D = 3D). D is the first, 25, 50, 75% and last position of a certain group in the peptide sequence (5 × 3 = 15D). Therefore, 8 * (3 + 3 + 15) = 168 features were produced for the CTD model.

Classifier

To find the most suitable machine learning algorithm, six commonly used classifiers are applied, including random forest (RF) (Ru et al., 2019), NaiveBayes, support vector machine (SVM) (Wei et al., 2018a; Dao et al., 2020a; Dao et al., 2020b; Wei et al., 2020), XGBoost (Yu et al., 2021a; Yang et al., 2021), logistic regression (LR) and decision tree (DT) (Li et al., 2019b). These efficient machine learning algorithms are usually used for feature analysis.

Feature Selection

For machine learning model building, features extracted from sequences always contain noise. A feature selection strategy to solve the information redundancy and overfitting problem can improve the feature representation ability (He et al., 2021). Analysis of variance (ANOVA) (Blanca et al., 2017; Wei et al., 2018b; Tang et al., 2018; Su et al., 2019a; Jung et al., 2019; Su et al., 2020; Liu et al., 2021b; Jin et al., 2021) has been used to analyse these characteristics and has been widely used in RNA, DNA and protein prediction. In this study, ANOVA is used to select the optimal features for model training. The feature subset with low redundancy is selected by ANOVA. We sort the original features

based on the ANOVA feature sorting algorithm and apply the IFS strategy to search the optimal feature subset.

Performance Standard

To evaluate the prediction accuracy of the model, the data of the following four formulas are usually used to solve the problem of classification prediction.

$$Acc = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$Sn = \frac{TP}{(TP + FN)}$$

$$Sp = \frac{TN}{(TN + FP)}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Accuracy (Acc), specificity (Sp), sensitivity (Sn) and Matthew correlation coefficient (MCC) were the commonly used evaluation parameters (Jiang et al., 2013; Wei et al., 2014; Wei et al., 2017a; Wei et al., 2017b; Wei et al., 2017c; Manavalan et al., 2019a; Manavalan et al., 2019b; Su et al., 2019b; Hong et al., 2019; Zeng et al., 2019; Zhang et al., 2020a; Liu et al., 2020b; Zeng et al., 2020; Yu et al., 2021b; Jin et al., 2021; Shao and Liu, 2021; Zhu et al., 2021). In the formulas, TP was the true positive number, TN was the true negative number, FP was the false-positive number and FN was the false negative number.

A receiver operating characteristic (ROC) curve was applied to study the prediction performance of the model. The area under the ROC curve (AUC) was used to assess the prediction performance of the model. AUC values of 0.5 and one represent random and perfect models, respectively (Zeng et al., 2017; Zeng et al., 2018; Dao et al., 2019; Feng et al., 2019; Lai et al., 2019; Lin et al., 2019; Zhu et al., 2019; Zhang et al., 2020b; Charoenkwan et al., 2020; Ding et al., 2020a; Ding et al., 2020b; Hasan et al., 2020; Huang et al., 2020; Jin et al., 2020; Li et al., 2020; Wang et al., 2020; Wu and Yu, 2021).

Construction of 3D Structure for AQPs

To verify the localization of aquaporins, the website (http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/) of the protein localization website and transmembrane prediction website (https://www.novopro.cn/tools/tmhmm.html) were applied to predict the subcellular localization and transmembrane structure of aquaporins. At the same time, Phyre2 software (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) was applied for the 3D structure prediction of aquaporins. The prediction results were visualized by PyMOL (version 2.5.1) software (https://pymol.org/2/).

Construction of Aquaporins Phylogenetic Tree

The phylogenetic tree of aquaporins was constructed to analyse the evolutionary diversity of the protein. Aquaporin sequence alignment results were analysed by MAFFT online software (https://mafft.cbrc.jp/alignment/server/) and used to construct a phylogenetic tree using IQ-TREE software (multicore version 1.6.12). The best fitting model for the phylogenetic tree was LG + F + R6 (Kalyaanamoorthy et al., 2017). The ultrafast bootstrap method was used for phylogenetic assessment, and 1,000 replicates per method were chosen in this work (Guindon et al., 2010; Minh et al., 2013; Hoang et al., 2018). The tree file was visualized by the iTOL website (https://itol.embl.de/).

TABLE 1 | Preliminary results of different feature descriptors using different classifiers.

188D_P: N = 1:1	Sn	Sp	Acc	MCC	AUROC
XGBoost	98	96.04	97.033	0.9416	0.9949
NaiveBayes	96.666	96.041	96.366	0.9279	0.9763
LR	97.999	94.748	96.377	0.9289	0.9857
DecisionTree	95.999	95.374	95.689	0.9155	0.9569
RF	98.666	96.707	97.689	0.9544	0.9987
SVM	95.332	96.04	95.7	0.9153	0.9917
188D_P: N = 1:2	Sn	Sp	Acc	MCC	AUROC
NaiveBayes	98	97.667	97.793	0.9531	0.9765
SVM	92.75	98.334	96.471	0.9224	0.9965
LR	96.083	98.001	97.359	0.9425	0.9958
RF	97.332	98.344	98.012	0.9564	0.9978
XGBoost	96.708	96.677	96.682	0.9284	0.9954
DecisionTree	92.041	93.687	93.141	0.8518	0.9286
188D_P: N = 1:3	Sn	Sp	Acc	MCC	AUROC
NaiveBayes	97.333	96.015	96.357	0.9102	0.9765
SVM	90.75	99.334	97.181	0.9244	0.9979
LR	95.417	98.445	97.682	0.9394	0.9952
RF	96.666	98.455	98.013	0.948	0.9979
XGBoost	95.374	98.011	97.343	0.9306	0.995
DecisionTree	93.999	96.697	96.024	0.8974	0.9535
188D_P: N = 1:4	Sn	Sp	Acc	MCC	AUROC
NaiveBayes	97.333	97.18	97.22	0.9206	0.9771
SVM	92.083	99.005	97.617	0.9256	0.9946
LR	93.457	97.844	96.953	0.9074	0.9942
RF	94.709	99.166	98.274	0.9461	0.9967
XGBoost	95.333	98.668	98.007	0.9391	0.9958
DecisionTree	92.708	98.841	97.614	0.9248	0.9578
188D_P: N = 1:5	Sn	Sp	Acc	MCC	AUROC
NaiveBayes	96.666	97.084	97.019	0.9022	0.9773
SVM	92.083	99.205	98.015	0.9292	0.9954
LR	93.999	98.143	97.46	0.9136	0.996
RF	94.667	99.338	98.562	0.9486	0.9975
XGBoost	95.333	98.94	98.341	0.9414	0.9963
DecisionTree	91.374	98.01	96.905	0.8924	0.9469
188D_P: N (151:8,994)	Sn	Sp	Acc	MCC	AUROC
XGBoost	86.084	99.934	99.703	0.9062	0.9989
NaiveBayes	96.666	97.977	97.955	0.6522	0.9793
LR	84.082	99.635	99.374	0.8158	0.9975
DecisionTree	72.208	99.365	98.918	0.6827	0.8579
RF	82.75	99.912	99.626	0.879	0.995
SVM	31.167	100	98.866	0.5503	0.9916

EXPERIMENT

Performance of Features Based on the 188-Dimensional Method (GPSD)

To select the best classifier for the AQP sequences, six widely used machine learning classifiers were employed to classify the features

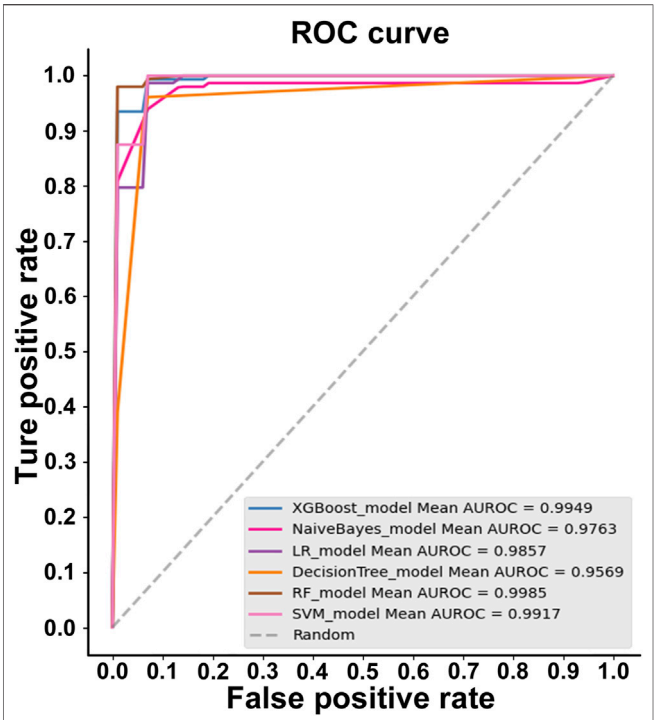


FIGURE 2 | ROC curves for the best performing feature with different classifiers.

of AQP sequences extracted by the 188-dimensional method (GPSD). For feature extraction by the 188-dimensional method (GPSD), we applied different ratios for the number of positive and negative samples (1:1, 1:2, 1:3, 1:4, 1:5, and 151: 8,994), and the results were classified by six machine learning classifiers (XGBoost, Naivebayes, LR, decision tree, RF and SVM). The results of all classifiers in the tenfold cross-validation were compared, and the comparison results are shown in **Table 1**.

The results of **Table 1** show that the different proportions of positive and negative samples indicated that P: N = 1:1 was the best ratio for the following analysis. Although the values of 1:2, 1:3, 1:4, 1:5 and 151:8,989 have higher values in SP and ACC, the values of Sn, MCC and AUROC are lower compared with P: N = 1:1. The increase in negative samples causes data imbalance and overfitting of the model. Therefore, the positive and negative sample ratio column of P: N = 1:1 is selected for model building.

For the AQP sequences (P: N = 1:1), random forest (RF) was the best algorithm, with the highest accuracy for the features extracted by the 188-dimensional method (GPSD) (AUC = 0.9987, Acc = 97.689%, MCC = 0.9544, Sn = 98.666%, Sp = 96.707%). XGBoost is the second algorithm with a slightly lower accuracy (AUC = 0.9949, Acc = 97.033%, MCC = 0.9416, Sn = 98%, Sp = 96.04%) compared with the random forest (RF) algorithm. The NaiveBayes, LR, DecisionTree and SVM algorithms have similar accuracies lower than the random forest (RF) algorithm for AQP sequence classification based on the 188-dimensional method (GPSD). The results in **Figure 2** indicated that RF was the best classifier with an accuracy of 0.9985, while the other classifiers of XGBoost, Naivebayes, LR, decision tree and SVM had accuracies of 0.9949, 0.9763, 0.9857, 0.9569 and 0.9917, respectively. In this study, six widely used classifiers are used for classification. The ROC of the RF classifier is 0.9985, which is relatively high. In general, regarding the evaluated accuracy of the AUC, Acc and MCC values, RF had the best performance in the AQP sequence classification results and was selected as the best classifier for model building.

Effect of Feature Selection Technologies

However, there are redundant or noisy features among the features extracted by the 188D method, which will affect the stability of the model. To overcome these effects, we use the ANOVA feature selection method to optimize these features. The optimized classification results of the feature selection method based on ANOVA are shown in **Table 2**. In addition, the optimal feature 54D is selected by combining ANOVA with an incremental feature selection (IFS) strategy, as shown in **Figure 3A**. The comparison results show that the accuracy of the optimal feature selected (ACC = 97.689) is slightly higher than that of the original feature (ACC = 97.356) (**Table 2**). Therefore, the ANOVA feature selection method was selected for feature optimization.

The PCA method was used to visually analyse the optimal feature (54D) after feature selection by the feature selection method (**Figure 3B**). **Figure 3B** indicates that positive and negative samples can almost be separated in the two-dimensional visualization diagram, which indicates that the 54D feature can effectively classify AQP proteins.

Feature Distribution Analysis

In this study, we performed feature analysis after feature selection. By analysing these 188D features, we determine the attribute information contained in these features. The results of feature analysis are shown in **Figure 3C**. According to the best feature analysis of the F-score value obtained by ANOVA, the features with an F-score value greater than 100 have a greater contribution to the classification. It can be seen from the figure that among the 188D features, the first is the 26th dimension feature, which is neutral/hydrophobic, followed by the 21st dimension feature, which is hydrophobic. The 26th dimension feature (neutral/hydrophobic) and 21st dimension feature (hydrophobic) signs showed that AQPs contained hydrophobic amino acids, which may be associated with the structural and functional properties of AQPs.

Structure Analysis of AQPs

Through feature selection, we know that hydrophobic features (the 26th dimension feature and 21st dimension feature) are the

TABLE 2 | ANOVA feature selection methods based on random forest.

ANOVA	Sn	Sp	Acc	MCC	AUROC
188D	98.666	96.04	97.356	0.9479	0.9991
ANOVA_ 54D	98.666	96.707	97.689	0.9544	0.997

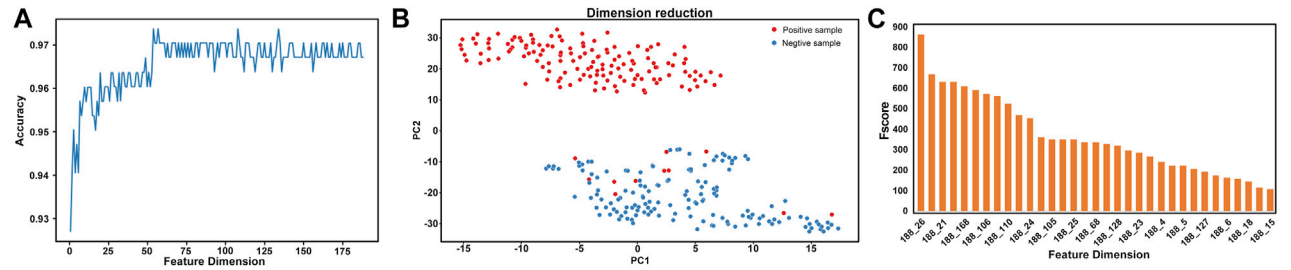


FIGURE 3 | Two-step feature selection result display **(A)** 10-fold CV and independent test accuracy of the RF classifier with the feature number varied **(B)** dimension reduction results based on the PCA method for the original data with a total of 188 dimensions **(C)** feature ranking of the F-score method obtained by ANOVA for the data with 188 features.

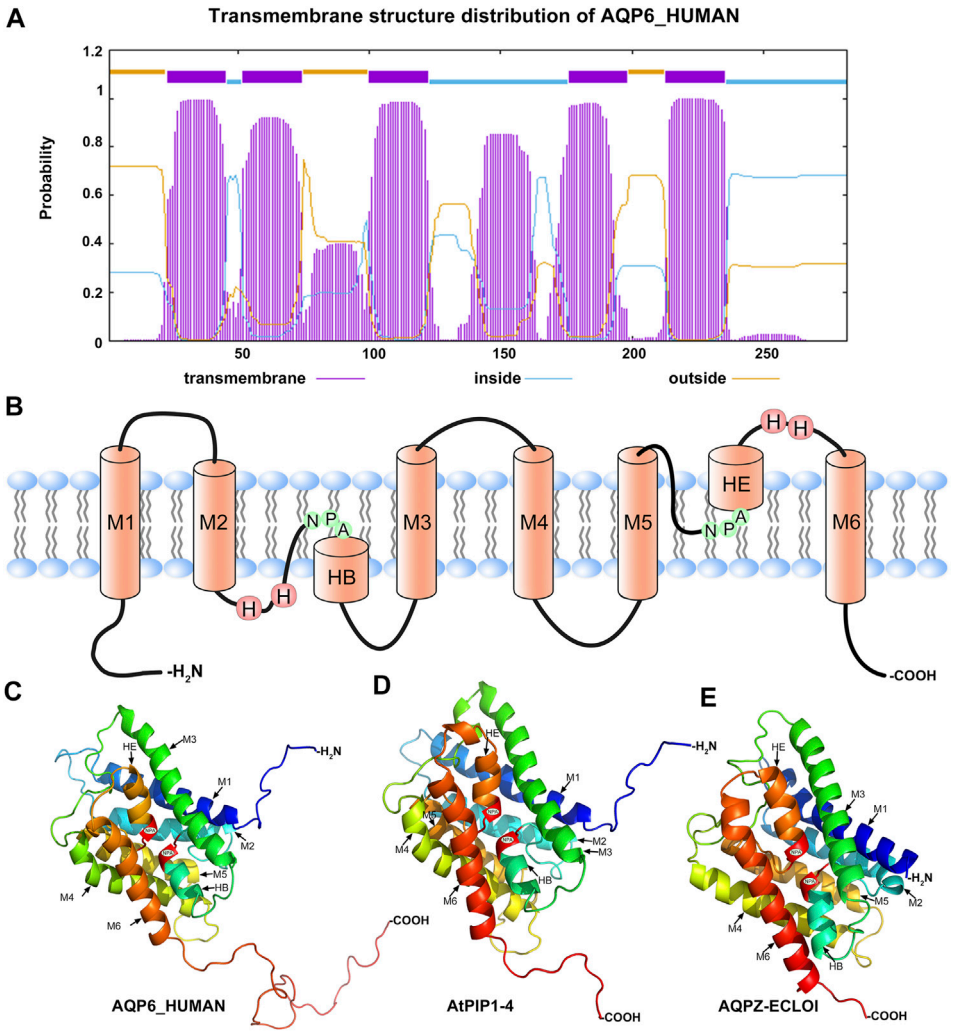


FIGURE 4 | The structure of AQPs **(A)** The prediction distribution of the transmembrane structure for AQP6_HUMAN **(B)** model of the structure of an AQP showing the principal features of the protein, NPA: asparagine-proline-alanine motifs; M1-M6: the transmembrane structure **(C–E)** 3D constructure of AQP6_HUMAN, AtPIP1-4 and AQPZ-ECLOI.

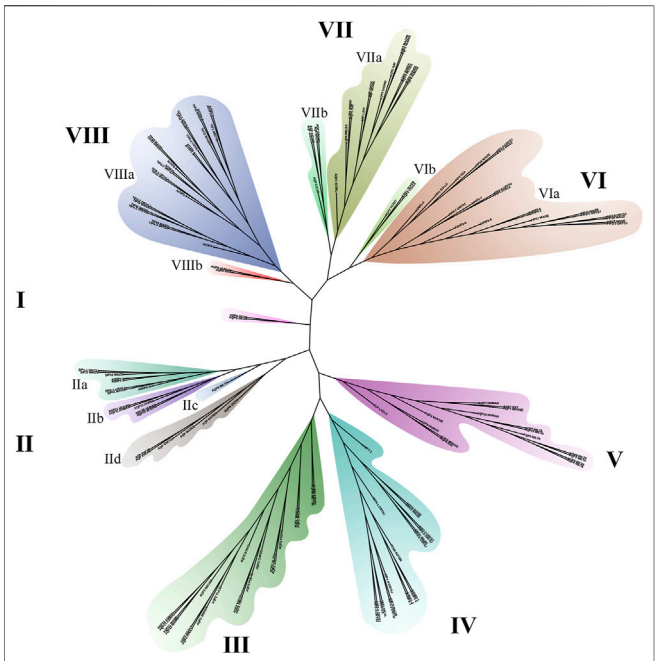


FIGURE 5 | Phylogenetic analysis of positive AQP proteins.

most significant features and make a great contribution to classification. Therefore, we analysed the protein localization of the AQP protein sequence, and the results showed that all AQP proteins were located on the cell membrane (**Supplement Table 1**). Cells are distinguished by a thin membrane. The core of the membrane is hydrophobic, which means it repels water. Many signals and nutrients cannot pass through the membrane itself but can pass through proteins across the membrane. Membrane proteins are essential for living cells, and plasma membrane proteins also have properties such as hydrophobicity, low solubility and low abundance. Therefore, the enrichment and classification extraction methods of soluble proteins cannot be used for plasma membrane proteins, mainly because the expression level of plasma membrane proteins in cells is very low, and they are highly hydrophobic in nature, which makes them easier to precipitate in aqueous solution and difficult to extract (Luche et al., 2003; Rawlings, 2016).

The Phyre2 website was used to analyse the transmembrane structure of HmAQP7. **Figure 2** shows that there are six α-helix transmembrane domains (**Figure 4A**): M1, M2, m3, M4, M5 and M6 (**Figure 4B**). A six-α-helix transmembrane domain forms a pore on the cell membrane to supply water molecules through the cell membrane. When the AQP protein folds, loops B (HB) and E (HE), which retain the lipophilic half helix, project to the protein molecular centre, making the highly conserved Asn-Pro-Asp (NPA) motif present the opposite direction, thus regulating the single file conductance of water and acting as a cation and proton exclusion filter (**Figures 4B–E**).

Evolution and Diversity

Aquaporin is a conserved membrane protein that contains highly conserved NAP domains and α-helical transmembrane domains in bacteria (**Figure 4E**), plants (**Figure 4D**) and humans (**Figure 4C**). To better verify the phylogenetic and evolutionary relationship of AQPs, 151 AQP protein sequences containing human, mouse, insect, fungus and bacteria were applied to construct a phylogenetic tree (**Figure 5**).

The results indicated that the 151 AQP protein sequences were divided into eight groups (**Figure 5**). The length of branches indicates the genetic relationship of AQP sequences. Among them, group III and group IV belong to plant and bacteria branches, respectively. Group II is the most complex branch, including the aquaporins of fungi, bacteria and animals. Among them, the VIa and VIIIa branches are plant subfamilies. AQPs of the VIIa and VIIb subfamilies belong to animals and insects, respectively. Group V contains one bacterial AQPZ and 15 animal AQPs, of which 7 belong to *Tardigrade*.

Expression of AQPs in Tumour Tissue

AQPs are considered to be important prognostic markers of cancers (Chow et al., 2020), so the expression of AQPs in cancer tissues is also crucial. **Figure 6** shows the expression level of AQP transcripts in 33 tumour tissues. AQP1_HUMAN has a high expression level in all tumour tissues and plays an important role in tumour angiogenesis and endothelial cell migration (Saadoun et al., 2005b). AQP3_HUMAN is expressed in almost all tumour tissues except ACC, LGG, UVM and AQP3_HUMAN-mediated glycerol transport, which allows the production of ATP for tumorigenesis. AQP3_HUMAN knockout mice can be resistant to carcinogen induction skin tumours (Hara-Chikuma and Verkman, 2008a). AQP3_HUMAN and AQP5_HUMAN were also expressed in COAD (Moon et al., 2003), while AQP5_HUMAN expression in human COAD is related to cell proliferation and metastasis. In BRCA, AQP5_HUMAN overexpression is associated with (Jung et al., 2011; Lee et al., 2014; Jensen et al., 2016) migration and poor prognosis in BRCA patients. Consistently, AQP5_HUMAN regulates miRNA migration through exosome-mediated (Park et al., 2020) and inhibits BRCA cell migration. AQP2_HUMAN, AQP12A_HUMAN, AQP12B_HUMAN and MIP had low expression levels in 33 tumour tissues, AQP4_HUMAN was highly expressed in GBM and LGG, and AQP9_HUMAN was highly expressed in LIHC.

Web Server Implementation

To facilitate the prediction of aquaporins, a user-friendly online server named iAQPs-RF is applied, which can be accessed from http://lab.malab.cn/~acy/iAQP. The protein sequences (FASTA format) were identified to determine whether aquaporins or non-aquaporins use the web server by users. First, the FASTA format protein sequences are enterd or pasted in the left blank box and the submit button is clicked; finally, the results are displayed on the right box. If you want to restart a new task, a clear button or the resubmit button was clicked to clear the sequences in the input box. Finally, new query protein sequences were allowed to enter the input box. The home page provides links of the contact information of authors and relevant data to download.

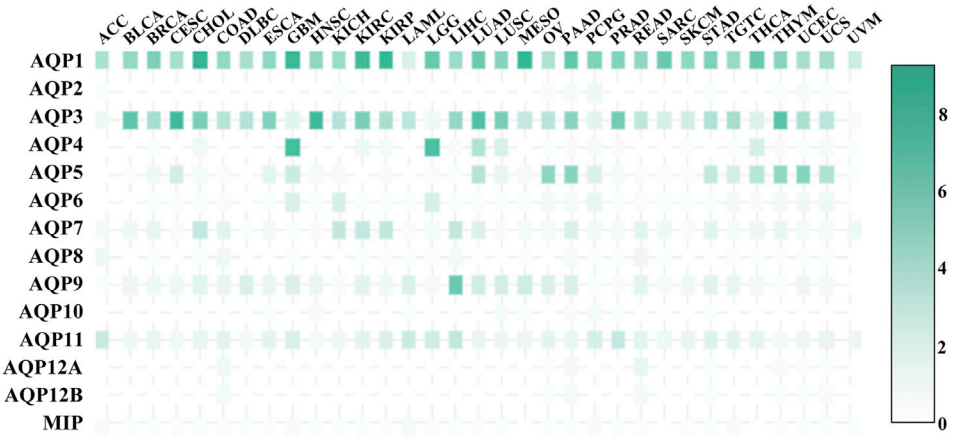


FIGURE 6 | Expression of AQPs in 33 human tumours.

CONCLUSION

The accurate identification of aquaporins by iAQPs can greatly promote the prediction of aquaporins and research on tumour diseases. In this study, we used the GPSD method to extract protein sequence features and the optimal random forest algorithm to construct new computational aquaporin identifier iAQPs-RF. Combined with the feature selection technique ANOVA, 54 optimal features are selected to build the predictor. According to the F-score value obtained by ANOVA, the 26th dimension feature and 21st dimension feature are ranked as the first and second dimension features among the 188 days features, respectively, and these two features possess neutral/hydrophobic characteristics. These two dimensional features make a great contribution to the classification of aquaporins. At the same time, through the location and 3D structure prediction of aquaporins protein, although the protein divided into eight groups and has diversity in evolution, all the proteins belong to plasma membrane proteins, and the protein sequence contains six α-helix transmembrane domains. The membrane proteins are hydrophobic and contain many hydrophobic amino acids (Luche et al., 2003; Rawlings, 2016), so these results are consistent with aquaporin classification.

The best CV evaluation accuracy of iAQPs-RF was 97.689%. At the same time, a network server is established. iAQPs-RF are expected to be a robust and reliable tool for aquaporin identification. Future work will focus on exploring deep learning to improve the performance of the model.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

LX, XS and LZ designed the research; ZC and SJ performed the research; ZC and DZ analyzed the data; ZC wrote the manuscript. All authors read and approved the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

FUNDING

The work was supported by the National Natural Science Foundation of China (No.62002244, No.62001311) and the Post-doctoral Foundation Project of Shenzhen Polytechnic China (No. 6020330002K).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.845622/full#supplementary-material

REFERENCES

Agre, P., King, L. S., Yasui, M., Guggino, W. B., Ottersen, O. P., Fujiyoshi, Y., et al. (2002). Aquaporin Water Channels - from Atomic Structure to Clinical Medicine. *J. Physiol.* 542, 3–16. doi:10.1113/jphysiol.2002.020818

Arsenijevic, T., Perret, J., Van Laethem, J.-L., and Delporte, C. (2019). Aquaporins Involvement in Pancreas Physiology and in Pancreatic Diseases. *Ijms* 20 (20), 5052. doi:10.3390/ijms20205052

Auguste, K. I., Jin, S., Uchida, K., Yan, D., Manley, G. T., Papadopoulos, M. C., et al. (2007). Greatly Impaired Migration of Implanted Aquaporin-4-deficient Astroglial Cells in Mouse Brain toward a Site of Injury. *FASEB j.* 21 (1), 108–116. doi:10.1096/fj.06-6848com

Bhardwaj, N., Langlois, R. E., Zhao, G., and Lu, H. (2005). Kernel-based Machine Learning Protocol for Predicting DNA-Binding Proteins. *Nucleic Acids Res.* 33 (20), 6486–6493. doi:10.1093/nar/gki949

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., and Bendayan, R. (2017). Non-normal Data: Is ANOVA Still a Valid Option? *Psicothema* 29 (4), 552–557. doi:10.7334/psicothema2016.383

Cai, L., Wang, L., Fu, X., Xia, C., Zeng, X., and Zou, Q. (2020). ITP-pred: an Interpretable Method for Predicting, Therapeutic Peptides with Fused Features Low-Dimension Representation. *Brief. Bioinform.* 22, bbaa367. doi:10.1093/bib/bbaa367

Cai, Y., He, J., Li, X., Lu, L., Yang, X., Feng, K., et al. (2009). A Novel Computational Approach to Predict Transcription Factor DNA Binding Preference. *J. Proteome Res.* 8 (2), 999–1003. doi:10.1021/pr800717y

Chae, Y. K., Woo, J., Kim, M.-J., Kang, S. K., Kim, M. S., Lee, J., et al. (2008). Expression of Aquaporin 5 (AQP5) Promotes Tumor Invasion in Human Non Small Cell Lung Cancer. *PLoS One* 3 (5), e2162. doi:10.1371/journal.pone.0002162

Charoenkwan, P., Yana, J., Schaduangrat, N., Nantasenamat, C., Hasan, M. M., Shoombuatong, W., et al. (2020). iBitter-SCM: Identification and Characterization of Bitter Peptides Using a Scoring Card Method with Propensity Scores of Dipeptides. *Genomics* 112 (4), 2813–2822. doi:10.1016/j.ygeno.2020.03.019

Chen, X.-X., Tang, H., Li, W.-C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. *Biomed. Res. Int.* 2016, 1–8. doi:10.1155/2016/1654623

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2020). iLearn: an Integrated Platform and Meta-Learner for Feature Engineering, Machine-Learning Analysis and Modeling of DNA, RNA and Protein Sequence Data. *Brief. Bioinformatics* 21 (3), 1047–1057. doi:10.1093/bib/bbz041

Chow, P. H., Bowen, J., and Yool, A. J. (2020). Combined Systematic Review and Transcriptomic Analyses of Mammalian Aquaporin Classes 1 to 10 as Biomarkers and Prognostic Indicators in Diverse Cancers. *Cancers* 12, 1911. doi:10.3390/cancers12071911

Dao, F.-Y., Lv, H., Wang, F., Feng, C.-Q., Ding, H., Chen, W., et al. (2019). Identify Origin of Replication in *Saccharomyces cerevisiae* Using Two-step Feature Selection Technique. *Bioinformatics (Oxford, England)* 35 (12), 2075–2083. doi:10.1093/bioinformatics/bty943

Dao, F.-Y., Lv, H., Yang, Y.-H., Zulfiqar, H., Gao, H., and Lin, H. (2020). Computational Identification of N6-Methyladenosine Sites in Multiple Tissues of Mammals. *Comput. Struct. Biotechnol. J.* 18, 1084–1091. doi:10.1016/j.csbj.2020.04.015

Dao, F.-Y., Lv, H., Zulfiqar, H., Yang, H., Su, W., Gao, H., et al. (2020). A Computational Platform to Identify Origins of Replication Sites in Eukaryotes. *Brief Bioinform* 22, 1940–1950. doi:10.1093/bib/bbaa017

De Ieso, M. L., and Yool, A. J. (2018). Mechanisms of Aquaporin-Facilitated Cancer Invasion and Metastasis. *Front. Chem.* 6, 135. doi:10.3389/fchem.2018.00135

Di Giusto, G., Flamenco, P., Rivarola, V., Fernández, J., Melamud, L., Ford, P., et al. (2012). Aquaporin 2-increased Renal Cell Proliferation Is Associated with Cell Volume Regulation. *J. Cell. Biochem.* 113 (12), 3721–3729. doi:10.1002/jcb.24246

Ding, T., Gu, F., Fu, L., and Ma, Y.-J. (2010). Aquaporin-4 in Glioma Invasion and an Analysis of Molecular Mechanisms. *J. Clin. Neurosci.* 17 (11), 1359–1361. doi:10.1016/j.jocn.2010.02.014

Ding, T., Zhou, Y., Sun, K., Jiang, W., Li, W., Liu, X., et al. (2013). Knockdown a Water Channel Protein, Aquaporin-4, Induced Glioblastoma Cell Apoptosis. *PLoS One* 8 (8), e66751. doi:10.1371/journal.pone.0066751

Ding, Y., Tang, J., and Guo, F. (2020a). Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. *Knowledge-Based Syst.* 204, 106254. doi:10.1016/j.knosys.2020.106254

Ding, Y., Tang, J., and Guo, F. (2020b). Identification of Drug-Target Interactions via Fuzzy Bipartite Local Model. *Neural Comput. Applic* 32, 10303–10319. doi:10.1007/s00521-019-04569-z

Direito, I., Madeira, A., Brito, M. A., and Soveral, G. (2016). Aquaporin-5: from Structure to Function and Dysfunction in Cancer. *Cell. Mol. Life Sci.* 73 (8), 1623–1640. doi:10.1007/s00018-016-2142-0

Feng, C.-Q., Zhang, Z.-Y., Zhu, X.-J., Lin, Y., Chen, W., Tang, H., et al. (2019). iTerm-PseKNC: a Sequence-Based Tool for Predicting Bacterial Transcriptional Terminators. *Bioinformatics (Oxford, England)* 35 (9), 1469–1477. doi:10.1093/bioinformatics/bty827

Fischer, H., Stenling, R., Rubio, C., and Lindblom, A. (2001). Differential Expression of Aquaporin 8 in Human Colonic Epithelial Cells and Colorectal Tumors. *BMC Physiol.* 1, 1. doi:10.1186/1472-6793-1-1

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., and Cd-Hit (2012). CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data. *Bioinformatics* 28 (23), 3150–3152. doi:10.1093/bioinformatics/bts565

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a Stacking and Pairwise Energy Content-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency. *Bioinformatics* 36 (10), 3028–3034. doi:10.1093/bioinformatics/btaa131

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* 59 (3), 307–321. doi:10.1093/sysbio/syq010

Hara-Chikuma, M., and Verkman, A. S. (2006). Aquaporin-1 Facilitates Epithelial Cell Migration in Kidney Proximal Tubule. *Jasn* 17 (1), 39–45. doi:10.1681/asn.2005080846

Hara-Chikuma, M., and Verkman, A. S. (2008). Aquaporin-3 Facilitates Epidermal Cell Migration and Proliferation during Wound Healing. *J. Mol. Med.* 86 (2), 221–231. doi:10.1007/s00109-007-0272-4

Hara-Chikuma, M., and Verkman, A. S. (2008). Prevention of Skin Tumorigenesis and Impairment of Epidermal Cell Proliferation by Targeted Aquaporin-3 Gene Disruption. *Mol. Cell Biol* 28 (1), 326–332. doi:10.1128/mcb.01482-07

Hasan, M. M., Schaduangrat, N., Basith, S., Lee, G., Shoombuatong, W., and Manavalan, B. (2020). HLPpred-Fuse: Improved and Robust Prediction of Hemolytic Peptide and its Activity by Fusing Multiple Feature Representation. *Bioinformatics (Oxford, England)* 36 (11), 3350–3356. doi:10.1093/bioinformatics/btaa160

He, S., Guo, F., Zou, Q., and HuiDing, H. (2021). MRMD2.0: A Python Tool for Machine Learning with Feature Ranking and Reduction. *Cbio* 15 (10), 1213–1221. doi:10.2174/1574893615999200503030350

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* 35 (2), 518–522. doi:10.1093/molbev/msx281

Hong, Z., Zeng, X., Wei, L., and Liu, X. (2019). Identifying Enhancer-Promoter Interactions with Neural Network Based on Pre-trained DNA Vectors and Attention Mechanism. *Bioinformatics* 36 (4), 1037–1043. doi:10.1093/bioinformatics/btz694

Huang, Y., Zhou, D., Wang, Y., Zhang, X., Su, M., Wang, C., et al. (2020). Prediction of Transcription Factors Binding Events Based on Epigenetic Modifications in Different Human Cells. *Epigenomics* 12 (16), 1443–1456. doi:10.2217/epi-2019-0321

Jensen, H. H., Login, F. H., Koffman, J. S., Kwon, T.-H., and Nejsum, L. N. (2016). The Role of Aquaporin-5 in Cancer Cell Migration: A Potential Active Participant. *Int. J. Biochem. Cell Biol.* 79, 271–276. doi:10.1016/j.biocel.2016.09.005

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting Human microRNA-Disease Associations Based on Support Vector Machine. *Ijdmb* 8 (3), 282–293. doi:10.1504/ijdmb.2013.056078

Jin, Q., Cui, H., Sun, C., Meng, Z., and Su, R. (2021). Free-form Tumor Synthesis in Computed Tomography Images via Richer Generative Adversarial Network. *Knowledge-Based Syst.* 218, 106753. doi:10.1016/j.knosys.2021.106753

Jin, S., Zeng, X., Xia, F., Huang, W., and Liu, X. (2020). Application of Deep Learning Methods in Biological Networks. *Brief Bioinform* 22 (2), 1902–1917. doi:10.1093/bib/bbaa043

Jung, H. J., Park, J.-Y., Jeon, H.-S., and Kwon, T.-H. (2011). Aquaporin-5: a Marker Protein for Proliferation and Migration of Human Breast Cancer Cells. *PLoS One* 6 (12), e28492. doi:10.1371/journal.pone.0028492

Jung, Y., Zhang, H., and Hu, J. (2019). Transformed Low-Rank ANOVA Models for High-Dimensional Variable Selection. *Stat. Methods Med. Res.* 28 (4), 1230–1246. doi:10.1177/0962280217753726

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nat. Methods* 14 (6), 587–589. doi:10.1038/nmeth.4285

Kang, S. K., Chae, Y. K., Woo, J., Kim, M. S., Park, J. C., Lee, J., et al. (2008). Role of Human Aquaporin 5 in Colorectal Carcinogenesis. *Am. J. Pathol.* 173 (2), 518–525. doi:10.2353/ajpath.2008.071198

Kasa, P., Farran, B., Prasad, G. L. V., and Nagaraju, G. P. (2019). Aquaporins in Female Specific Cancers. *Gene* 700, 60–64. doi:10.1016/j.gene.2019.03.032

Kröger, S., Wolburg, H., and Warth, A. (2004). Redistribution of Aquaporin-4 in Human Glioblastoma Correlates with Loss of Agrin Immunoreactivity from Brain Capillary Basal Laminae. *Acta neuropathologica* 107 (4), 311–318. doi:10.1007/s00401-003-0812-0

Kumar, M., Gromiha, M. M., and Raghava, G. P. (2007). Identification of DNA-Binding Proteins Using Support Vector Machines and Evolutionary Profiles. *BMC Bioinformatics* 8, 463. doi:10.1186/1471-2105-8-463

Lai, H.-Y., Zhang, Z.-Y., Su, Z.-D., Su, W., Ding, H., Chen, W., et al. (2019). iProEP: A Computational Predictor for Predicting Promoter. *Mol. Ther. - Nucleic Acids* 17, 337–346. doi:10.1016/j.omtn.2019.05.028

Lan, Y.-L., Wang, X., Lou, J.-C., Ma, X.-C., and Zhang, B. (2017). The Potential Roles of Aquaporin 4 in Malignant Gliomas. *Oncotarget* 8 (19), 32345–32355. doi:10.18632/oncotarget.16017

Lee, S. J., Chae, Y. S., Kim, J. G., Kim, W. W., Jung, J. H., Park, H. Y., et al. (2014). AQP5 Expression Predicts Survival in Patients with Early Breast Cancer. *Ann. Surg. Oncol.* 21 (2), 375–383. doi:10.1245/s10434-013-3317-7

Levin, M. H., and Verkman, A. S. (2006). Aquaporin-3-dependent Cell Migration and Proliferation during Corneal Re-epithelialization. *Invest. Ophthalmol. Vis. Sci.* 47 (10), 4365–4372. doi:10.1167/iovs.06-0335

Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepATT: a Hybrid Category Attention Neural Network for Identifying Functional Effects of DNA Sequences. *Brief. Bioinform.* 22, 1. doi:10.1093/bib/bbaa159

Li, M., Xu, H., and Deng, Y. (2019). Evidential Decision Tree Based on Belief Entropy. *Entropy* 21 (9), 897. doi:10.3390/e21090897

Li, Y., Niu, M., Zou, Q., and Elm-, M. H. C. (2019). ELM-MHC: An Improved MHC Identification Method with Extreme Learning Machine Algorithm. *J. Proteome Res.* 18 (3), 1392–1401. doi:10.1021/acs.jproteome.9b00012

Lin, C.-W., Chen, P.-N., Chen, M.-K., Yang, W.-E., Tang, C.-H., Yang, S.-F., et al. (2013). Kaempferol Reduces Matrix Metalloproteinase-2 Expression by Down-Regulating ERK1/2 and the Activator Protein-1 Signaling Pathways in Oral Cancer Cells. *PLoS One* 8 (11), e80883. doi:10.1371/journal.pone.0080883

Lin, H., Liang, Z.-Y., Tang, H., and Chen, W. (2019). Identifying Sigma70 Promoters with Novel Pseudo Nucleotide Composition. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (4), 1316–1321. doi:10.1109/tcbb.2017.2666141

Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an Updated Platform for Analyzing DNA, RNA and Protein Sequences at Sequence Level and Residue Level Based on Machine Learning Approaches. *Nucleic Acids Res.* 47 (20), e127. doi:10.1093/nar/gkz740

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., et al. (2014). iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition. *PLoS One* 9 (9), e106691. doi:10.1371/journal.pone.0106691

Liu, B., Zhu, Y., and Yan, K. (2020). Fold-LTR-TCP: Protein Fold Recognition Based on Triadic Closure Principle. *Brief. Bioinform.* 21 (6), 2185–2193. doi:10.1093/bib/bbz139

Liu, J., Su, R., Zhang, J., and Wei, L. (2021). Classification and Gene Selection of Triple-Negative Breast Cancer Subtype Embedding Gene Connectivity Matrix in Deep Neural Network. *Brief Bioinform* 22, bbaa395. doi:10.1093/bib/bbaa395

Liu, M.-L., Su, W., Wang, J.-S., Yang, Y.-H., Yang, H., and Lin, H. (2020). Predicting Preference of Transcription Factors for Methylated DNA Using Sequence Information. *Mol. Ther. - Nucleic Acids* 22, 1043–1050. doi:10.1016/j.omtn.2020.07.035

Liu, X.-J., Gong, X.-J., Yu, H., and Xu, J.-H. (2018). A Model Stacking Framework for Identifying DNA Binding Proteins by Orchestrating Multi-View Features and Classifiers. *Genes* 9 (8), 394. doi:10.3390/genes9080394

Liu, Y., Ouyang, X.-h., Xiao, Z.-X., Zhang, L., and Cao, Y. (2021). A Review on the Methods of Peptide-MHC Binding Prediction. *Cbio* 15 (8), 878–888. doi:10.2174/1574893615999200429122801

Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B., and Zhang, H. (2014). Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. *PLoS One* 9 (1), e86703. doi:10.1371/journal.pone.0086703

Luche, S., Santoni, V., and Rabilloud, T. (2003). Evaluation of Nonionic and Zwitterionic Detergents as Membrane Protein Solubilizers in Two-Dimensional Electrophoresis. *Proteomics* 3 (3), 249–253. doi:10.1002/pmic.200390037

Ma, T., Yang, B., and Verkman, A. S. (1997). Cloning of a Novel Water and Urea-Permeable Aquaporin from Mouse Expressed Strongly in colon, Placenta, Liver, and Heart. *Biochem. Biophysical Res. Commun.* 240 (2), 324–328. doi:10.1006/bbrc.1997.7664

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). mAHTPred: a Sequence-Based Meta-Predictor for Improving the Prediction of Anti-hypertensive Peptides Using Effective Feature Representation. *Bioinformatics* 35 (16), 2757–2765. doi:10.1093/bioinformatics/bty1047

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. *Mol. Ther. - Nucleic Acids* 16, 733–744. doi:10.1016/j.omtn.2019.04.019

Marlar, S., Jensen, H. H., Login, F. H., and Nejsum, L. N. (2017). Aquaporin-3 in Cancer. *Ijms* 18 (10), 2106. doi:10.3390/ijms18102106

Maugeri, R., Schiera, G., Di Liegro, C., Fricano, A., Iacopino, D., and Di Liegro, I. (2016). Aquaporins and Brain Tumors. *Ijms* 17 (7), 1029. doi:10.3390/ijms17071029

Minh, B. Q., Nguyen, M. A. T., and von Haeseler, A. (2013). Ultrafast Approximation for Phylogenetic Bootstrap. *Mol. Biol. Evol.* 30 (5), 1188–1195. doi:10.1093/molbev/mst024

Mobasheri, A., Airley, R., Hewitt, S., and Marples, D. (2005). Heterogeneous Expression of the Aquaporin 1 (AQP1) Water Channel in Tumors of the Prostate, Breast, Ovary, colon and Lung: a Study Using High Density Multiple Human Tumor Tissue Microarrays. *Int. J. Oncol.* 26 (5), 1149–1158. doi:10.3892/ijo.26.5.1149

Moon, C., Soria, J.-C., Jang, S. J., Lee, J., Hoque, M. O., Sibony, M., et al. (2003). Involvement of Aquaporins in Colorectal Carcinogenesis. *Oncogene* 22 (43), 6699–6703. doi:10.1038/sj.onc.1206762

Muhammod, R., Ahmed, S., Md Farid, D., Shatabda, S., Sharma, A., and Dehzangi, A. (2019). PyFeat: a Python-Based Effective Feature Generation Tool for DNA, RNA and Protein Sequences. *Bioinformatics (Oxford, England)* 35 (19), 3831–3833. doi:10.1093/bioinformatics/btz165

Nagaraju, G. P., Basha, R., Rajitha, B., Alese, O. B., Alam, A., Pattnaik, S., et al. (2016). Aquaporins: Their Role in Gastrointestinal Malignancies. *Cancer Lett.* 373 (1), 12–18. doi:10.1016/j.canlet.2016.01.003

Nakahigashi, K., Kabashima, K., Ikoma, A., Verkman, A. S., Miyachi, Y., and Hara-Chikuma, M. (2011). Upregulation of Aquaporin-3 Is Involved in Keratinocyte Proliferation and Epidermal Hyperplasia. *J. Invest. Dermatol.* 131 (4), 865–873. doi:10.1038/jid.2010.395

Nielsen, S., Frøkiær, J., Marples, D., Kwon, T.-H., Agre, P., and Knepper, M. A. (2002). Aquaporins in the Kidney: from Molecules to Medicine. *Physiol. Rev.* 82 (1), 205–244. doi:10.1152/physrev.00024.2001

Park, E. J., Jung, H. J., Choi, H. J., Jang, H. J., Park, H. J., Nejsum, L. N., et al. (2020). Exosomes Co-expressing AQP5-targeting miRNAs and IL-4 Receptor-binding Peptide Inhibit the Migration of Human Breast Cancer Cells. *FASEB j.* 34 (2), 3379–3398. doi:10.1096/fj.201902434R

Preston, G. M., Carroll, T. P., Guggino, W. B., and Agre, P. (1992). Appearance of Water Channels in Xenopus Oocytes Expressing Red Cell CHIP28 Protein. *Science* 256 (5055), 385–387. doi:10.1126/science.256.5055.385

Rawlings, A. E. (2016). Membrane Proteins: Always an Insoluble Problem? *Biochem. Soc. Trans.* 44, 790–795. doi:10.1042/BST20160025

Rojek, A., Praetorius, J., Frøkiaer, J., Nielsen, S., and Fenton, R. A. (2008). A Current View of the Mammalian Aquaglyceroporins. *Annu. Rev. Physiol.* 70, 301–327. doi:10.1146/annurev.physiol.70.113006.100452

Ru, X., Li, L., and Zou, Q. (2019). Incorporating Distance-Based Top-N-Gram and Random Forest to Identify Electron Transport Proteins. *J. Proteome Res.* 18 (7), 2931–2939. doi:10.1021/acs.jproteome.9b00250

Saadoun, S., Papadopoulos, M. C., Davies, D. C., Krishna, S., and Bell, B. A. (2002). Aquaporin-4 Expression Is Increased in Oedematous Human Brain Tumours. *J. Neurol. Neurosurg. Psychiatry* 72 (2), 262–265. doi:10.1136/jnnp.72.2.262

Saadoun, S., Papadopoulos, M. C., Davies, D. C., Bell, B. A., and Krishna, S. (2002). Increased Aquaporin 1 Water Channel Expression Inhuman Brain Tumours. *Br. J. Cancer* 87 (6), 621–623. doi:10.1038/sj.bjc.6600512

Saadoun, S., Papadopoulos, M. C., Hara-Chikuma, M., and Verkman, A. S. (2005). Impairment of Angiogenesis and Cell Migration by Targeted Aquaporin-1 Gene Disruption. *Nature* 434 (7034), 786–792. doi:10.1038/nature03460

Saadoun, S., Papadopoulos, M. C., Watanabe, H., Yan, D., Manley, G. T., and Verkman, A. S. (2005). Involvement of Aquaporin-4 in Astroglial Cell Migration and Glial Scar Formation. *J. Cel. Sci.* 118 (Pt 24), 5691–5698. doi:10.1242/jcs.02680

Shanahan, H. P., Garcia, M. A., Jones, S., and Thornton, J. M. (2004). Identifying DNA-Binding Proteins Using Structural Motifs and the Electrostatic Potential. *Nucleic Acids Res.* 32 (16), 4732–4741. doi:10.1093/nar/gkh803

Shao, J., and Liu, B. (2021). ProtFold-DFG: Protein Fold Recognition by Combining Directed Fusion Graph and PageRank Algorithm. *Brief Bioinform* 22 (3), bbaa192. doi:10.1093/bib/bbaa192

Shao, J., Yan, K., and Liu, B. (2021). FoldRec-C2C: Protein Fold Recognition by Combining Cluster-To-Cluster Model and Protein Similarity Network. *Brief Bioinform* 22 (3), bbaa144. doi:10.1093/bib/bbaa144

Shen, Z., and Zou, Q. (2020). Basic Polar and Hydrophobic Properties Are the Main Characteristics that Affect the Binding of Transcription Factors to Methylation Sites. *Bioinformatics* 36 (15), 4263–4268. doi:10.1093/bioinformatics/btaa492

Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2020). Empirical Comparison and Analysis of Web-Based Cell-Penetrating Peptide Prediction Tools. *Brief. Bioinform.* 21 (2), 408–420. doi:10.1093/bib/bby124

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A Deep forest Model to Predict Anti-cancer Drug Response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2019). Developing a Multi-Dose Computational Model for Drug-Induced Hepatotoxicity Prediction Based on Toxicogenomics Data. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (4), 1231–1239. doi:10.1109/tcbb.2018.2858756

Su, W., Liu, M.-L., Yang, Y.-H., Wang, J.-S., Li, S.-H., Lv, H., et al. (2021). PPD: A Manually Curated Database for Experimentally Verified Prokaryotic Promoters. *J. Mol. Biol.* 433 (11), 166860. doi:10.1016/j.jmb.2021.166860

Szilágyi, A., and Skolnick, J. (2006). Efficient Prediction of Nucleic Acid Binding Function from Low-Resolution Protein Structures. *J. Mol. Biol.* 358 (3), 922–933. doi:10.1016/j.jmb.2006.02.053

Tang, H., Zhao, Y.-W., Zou, P., Zhang, C.-M., Chen, R., Huang, P., et al. (2018). HBPred: a Tool to Identify Growth Hormone-Binding Proteins. *Int. J. Biol. Sci.* 14 (8), 957–964. doi:10.7150/ijbs.24174

Tang, Y.-J., Pang, Y.-H., Liu, B., and Idp-Seq2Seq (2020). IDP-Seq2Seq: Identification of Intrinsically Disordered Regions Based on Sequence to Sequence Learning. *Bioinformaitcs* 36 (21), 5177–5186. doi:10.1093/bioinformatics/btaa667

Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., and Raghava, G. P. S. (2013). In Silico models for Designing and Discovering Novel Anticancer Peptides. *Sci. Rep.* 3, 2984. doi:10.1038/srep02984

Verkman, A. S. (2005). More Than Just Water Channels: Unexpected Cellular Roles of Aquaporins. *J. Cel. Sci.* 118 (Pt 15), 3225–3232. doi:10.1242/jcs.02519

Wang, D., and Owler, B. K. (2011). Expression of AQP1 and AQP4 in Paediatric Brain Tumours. *J. Clin. Neurosci.* 18 (1), 122–127. doi:10.1016/j.jocn.2010.07.115

Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of Membrane Protein Types via Multivariate Information Fusion with Hilbert-Schmidt Independence Criterion. *Neurocomputing* 383, 257–269. doi:10.1016/j.neucom.2019.11.103

Warth, A., Simon, P., Capper, D., Goeppert, B., Tabatabai, G., Herzog, H., et al. (2007). Expression Pattern of the Water Channel Aquaporin-4 in Human Gliomas Is Associated with Blood-Brain Barrier Disturbance but Not with Patient Survival. *J. Neurosci. Res.* 85 (6), 1336–1346. doi:10.1002/jnr.21224

Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: A Sequence-Based Predictor for Identifying N6-Methyladenosine Sites Using Ensemble Learning. *Mol. Ther. - Nucleic Acids* 12, 635–644. doi:10.1016/j.omtn.2018.07.004

Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2020). Comparative Analysis and Prediction of Quorum-sensing Peptides Using Feature Representation Learning and Machine Learning Algorithms. *Brief. Bioinform.* 21 (1), 106–119. doi:10.1093/bib/bby107

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. *Ieee/acm Trans. Comput. Biol. Bioinf.* 11 (1), 192–201. doi:10.1109/tcbb.2013.146

Wei, L., Tang, J., and Zou, Q. (2017). Local-DPP: An Improved DNA-Binding Protein Prediction Method by Exploring Local Evolutionary Information. *Inf. Sci.* 384, 135–144. doi:10.1016/j.ins.2016.06.026

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017). A Novel Hierarchical Selective Ensemble Classifier with Bioinformatics Application. *Artif. Intelligence Med.* 83, 82–90. doi:10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved Prediction of Protein-Protein Interactions Using Novel Negative Samples, Features, and an Ensemble Classifier. *Artif. Intelligence Med.* 83, 67–74. doi:10.1016/j.artmed.2017.03.001

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a Sequence-Based Predictor Using Effective Feature Representation to Improve the Prediction of Anti-cancer Peptides. *Bioinformatics* 34 (23), 4007–4016. doi:10.1093/bioinformatics/bty451

Wu, X., and Yu, L. (2021). EPSOL: Sequence-Based Protein Solubility Prediction Using Multidimensional Embedding. *Bioinformatics (Oxford, England)* 37, 4314–4320. doi:10.1093/bioinformatics/btab463

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big Data Mining with Fusion of Multifarious Physical Examination Indicators. *Inf. Fusion* 75, 140–149. doi:10.1016/j.inffus.2021.02.015

Yu, L., Wang, M., Yang, Y., Xu, F., Zhang, X., Xie, F., et al. (2021). Predicting Therapeutic Drugs for Hepatocellular Carcinoma Based on Tissue-specific Pathways. *Plos Comput. Biol.* 17 (2), e1008696. doi:10.1371/journal.pcbi.1008696

Yu, X., Zhou, J., Zhao, M., Yi, C., Duan, Q., Zhou, W., et al. (2021). Exploiting XG Boost for Predicting Enhancer-Promoter Interactions. *Cbio* 15 (9), 1036–1045. doi:10.2174/1574893615666200120103948

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and Validation of Disease Genes Using HeteSim Scores. *Ieee/acm Trans. Comput. Biol. Bioinf.* 14 (3), 687–695. doi:10.1109/tcbb.2016.2520947

Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of Potential Disease-Associated microRNAs Using Structural Perturbation Method. *Bioinformatics* 34 (14), 2425–2432. doi:10.1093/bioinformatics/bty112

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a Network-Based Deep Learning Approach to In Silico Drug Repositioning. *Bioinformatics* 35 (24), 5191–5198. doi:10.1093/bioinformatics/btz418

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target Identification Among Known Drugs by Deep Learning from Heterogeneous Networks. *Chem. Sci.* 11 (7), 1775–1797. doi:10.1039/c9sc04336e

Zhang, D., Chen, H.-D., Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Zhang, Z.-Y., et al. (2021). iBLP: An XGBoost-Based Predictor for Identifying Bioluminescent Proteins. *Comput. Math. Methods Med.* 2021, 1–15. doi:10.1155/2021/6664362

Zhang, J., Chen, Q., and Liu, B. (2020). iDRBP_MMC: Identifying DNA-Binding Proteins and RNA-Binding Proteins Based on Multi-Label Learning Model and Motif-Based Convolutional Neural Network. *J. Mol. Biol.* 432 (22), 5860–5875. doi:10.1016/j.jmb.2020.09.008

Zhang, L., Xiao, X., and Xu, Z.-C. (2020). iPromoter-5mC: A Novel Fusion Decision Predictor for the Identification of 5-Methylcytosine Sites in Genome-wide DNA Promoters. *Front. Cell Dev. Biol.* 8, 614. doi:10.3389/fcell.2020.00614

Zhang, Z., Chen, Z., Song, Y., Zhang, P., Hu, J., and Bai, C. (2010). Expression of Aquaporin 5 Increases Proliferation and Metastasis Potential of Lung Cancer. *J. Pathol.* 221 (2), 210–220. doi:10.1002/path.2702

Zhu, X.-J., Feng, C.-Q., Lai, H.-Y., Chen, W., and Hao, L. (2019). Predicting Protein Structural Classes for Low-Similarity Sequences by Evaluating Different Features. *Knowledge-Based Syst.* 163, 787–793. doi:10.1016/j.knosys.2018.10.007

Zhu, Y., Li, F., Xiang, D., Akutsu, T., Song, J., and Jia, C. (2021). Computational Identification of Eukaryotic Promoters Based on Cascaded Deep Capsule Neural Networks. *Brief Bioinform* 22 (4), bbaa299. doi:10.1093/bib/bbaa299

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence Clustering in Bioinformatics: an Empirical Study. *Brief. Bioinform.* 21 (1), 1–10. doi:10.1093/bib/bby090

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Chen, Jiao, Zhao, Zou, Xu, Zhang and Su. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 01 February 2022
doi: 10.3389/fcell.2022.815843

Discovered Key CpG Sites by Analyzing DNA Methylation and Gene Expression in Breast Cancer Samples

Yan-Ni Cao[1], *Qian-Zhong Li*[1,2]* *and Yu-Xian Liu*[1]

[1]*Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot, China,* [2]*The State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Inner Mongolia University, Hohhot, China*

Breast cancer is the most common cancer in the world, and DNA methylation plays a key role in the occurrence and development of breast cancer. However, the effect of DNA methylation in different gene functional regions on gene expression and the effect of gene expression on breast cancer is not completely clear. In our study, we computed and analyzed DNA methylation, gene expression, and clinical data in the TCGA database. Firstly, we calculated the distribution of abnormal DNA methylated probes in 12 regions, found the abnormal DNA methylated probes in down-regulated genes were highly enriched, and the number of hypermethylated probes in the promoter region was 6.5 times than that of hypomethylated probes. Secondly, the correlation coefficients between abnormal DNA methylated values in each functional region of differentially expressed genes and gene expression values were calculated. Then, co-expression analysis of differentially expressed genes was performed, 34 hub genes in cancer-related pathways were obtained, of which 11 genes were regulated by abnormal DNA methylation. Finally, a multivariate Cox regression analysis was performed on 27 probes of 11 genes. Three DNA methylation probes (cg13569051 and cg14399183 of *GSN*, and cg25274503 of *CAV2*) related to survival were used to construct a prognostic model, which has a good prognostic ability. Furthermore, we found that the cg25274503 hypermethylation in the promoter region inhibited the expression of the *CAV2*, and the hypermethylation of cg13569051 and cg14399183 in the 5′UTR region inhibited the expression of *GSN*. These results may provide possible molecular targets for breast cancer.

Keywords: gene expression, DNA methylation, correlation, breast cancer, molecular targets

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Liang Yu,
Xidian University, China
Quan Zou,
University of Electronic Science and Technology of China, China

***Correspondence:**
Qian-Zhong Li
qzli@imu.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *15 November 2021*
Accepted: *06 January 2022*
Published: *01 February 2022*

Citation:
Cao Y-N, Li Q-Z and Liu Y-X (2022) Discovered Key CpG Sites by Analyzing DNA Methylation and Gene Expression in Breast Cancer Samples. Front. Cell Dev. Biol. 10:815843. doi: 10.3389/fcell.2022.815843

Abbreviations: ADMPs, abnormal DNA methylated probes; AUC, area under the receiver operating characteristic curve; CGI, CpG islands; DEGs, differentially expressed genes; FPKM, fragments per kilobase of exon model per million mapped fragments; GS, Gene Significance; MM, Module Membership; ROC, receiver operating characteristic; TSS, transcription start site; TTS, transcription termination site.

INTRODUCTION

Breast cancer is the most common malignant tumor in women and the main cause of cancer deaths in women worldwide. There were more than 2 million new breast cancer patients and more than 620,000 patients who died of breast cancer in 2018 (Bray et al., 2018). Since 2004, the incidence of breast cancer has increased slightly at a rate of about 0.3% per year (Siegel et al., 2020). At present, despite the use of many advanced treatment technologies to improve survival, but the quality of life for patients is poor, and for most patients, the finding of disease is in the late stage, or metastasis occurs at the late stage of diagnosis (Punglia et al., 2007). Therefore, it is particularly important to study breast cancer.

DNA methylation is a heritable and reversible epigenetic modification that can regulate gene expression without changing the DNA sequence. It mainly occurs at CpG sites and is considered to be a goalkeeper for long-term stable regulation of gene expression (Cedar and Bergman, 2009). DNA methylation of different regions has both positive and negative correlations with gene expression in breast cancer (Győrffy et al., 2016). Many studies have reported that DNA methylation in the promoter region is negatively correlated with gene expression (Li et al., 2016; Janostiak et al., 2018). Studies have also found that DNA methylation is positively correlated with gene expression levels in the gene body region (Yang et al., 2014). It has been reported that DNA methylation may play a key role in the process of carcinogenesis by down-regulating the expression of tumor suppressor genes (Jones and Baylin, 2007). The hypermethylation in the promoter region of tumor suppressor genes is related to gene inactivation and transcriptional inhibition, and the hypermethylation of CpG islands (CGI) in the promoter region is considered to be one of the earliest and most frequent changes in cancer (Baylin, 2005; Wittenberger et al., 2014). There are some studies have shown that the hypomethylation of the enhancer region is closely related to up-regulation for gene expression in breast cancer (Jin et al., 2019). The hypomethylation in the promoter region is related to the activation of oncogenes and metastasis-promoting genes and has been verified to play an important role in the occurrence, development, and metastasis of cancer (Stefanska et al., 2011; Jones, 2012; Nilsson et al., 2014). Thus, it is necessary to study the effect of DNA methylation in different regions on gene expression.

Studies have shown that abnormal DNA methylation is considered to be a key factor leading to the carcinogenesis of various tumors, including breast cancer (Kulis and Esteller, 2010; Karsli-Ceppioglu et al., 2014). For example, *PSAT1* methylation is associated with HR-positive, lymph node-positive breast cancer, and invasive lobular cancer. *GNE* methylation is associated with HR-negative breast cancer, while *CXCL14* methylation is associated with HER2-positive breast cancer (Bu et al., 2013). Studies have also found that *DACT2* promoter methylation is related to advanced tumor staging (Borgonio-Cuadra et al., 2018). *CRY2* is an independent indicator that reduces the risk of metastasis and recurrence in ER+ breast cancer patients (Liu et al., 2017). *DFNA5* methylation shows strong potential as a biomarker for breast cancer detection and prognosis (Croes et al., 2018). Therefore, it is very meaningful to find key genes regulated by DNA methylation in breast cancer.

Although there have been substantial advances in breast cancer treatment, the treatment of breast cancer is still limited due to the lack of precise breast cancer molecular targets (Tang et al., 2018). In this study, to find molecular targets of DNA methylation that affect breast cancer prognosis. Firstly, the correlations between gene expression values and abnormal DNA methylation in different functional regions of four types of genes were computed and analyzed. Secondly, breast cancer may not be caused by the regulation of a single gene but by the joint regulation of multiple genes. To find candidate molecular targets and describe the correlation patterns between genes, we used weighted gene co-expression network analysis (WGCNA) to construct a co-expressed gene network. Based on the analyses of the correlation between the module gene and the clinical characteristics of the samples, two modules that were strongly related to cancer were obtained, and the hub genes were selected by analyzing the importance of the gene in the module. We analyzed the pathways of these hub genes and selected genes enriched in key pathways as the key genes for our research. Then by analyzing the genes whose absolute value of the correlation between expression value of key genes and DNA methylation of different sites was greater than or equal to 0.6, we found that our results verify that promoter methylation was negatively correlated with gene expression, and gene body region methylation was positively correlated with gene expression. Finally, a multivariate Cox regression analysis was performed on the 27 probes, and a regression model was constructed using three probes. Survival analysis shows that the prognostic performance of the model is good. Consequently, the three probes may be molecular targets related to methylation in breast cancer.

MATERIALS AND METHODS

Data Sources

We downloaded the gene expression data [fragments per kilobase of exon model per million mapped fragments (FPKM) and COUNTS], DNA methylation data (HM450K), and clinical data (**Supplementary Table S1**) (hg38) of breast cancer and paracancerous tissues from the TCGA (The Cancer Genome Atlas) (https://tcga-data.nci.nih.gov/tcga/) database (**Supplementary Table S2**). We downloaded the human reference genome annotation file RefSeq gene (hg38) and the location file of CGI from UCSC (http:/genome.ucsc.edu/). The position file of the enhancer was obtained from the FANTOM5 (Function Annotation of The Mammalian Genome) (https://fantom.gsc.riken.jp/5/) database.

Data Preprocessing and Division of Different Regions

For gene annotation file, we retained 57,392 transcripts starting with the NM (the mature messenger RNA). We randomly reserve one of the transcripts with the same transcription start site (TSS),

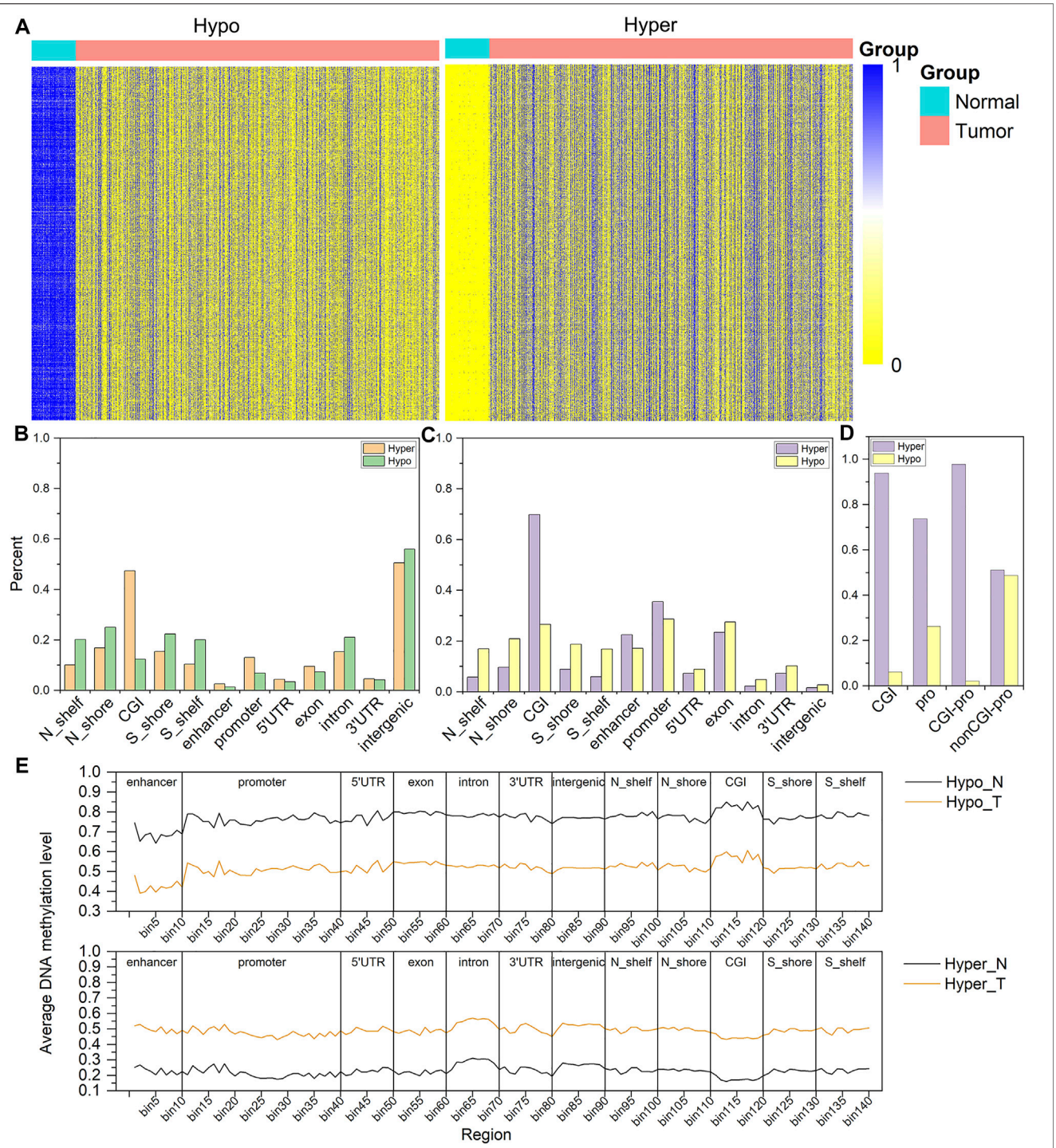


FIGURE 1 | Distribution map of DNA methylation in the whole genome. **(A)** is the methylation value of hypomethylation (Hypo) and hypermethylation (Hyper) probes in different samples, the rows are the methylation values of each probe, and the columns are the different samples. **(B)** is the density distribution of DNA methylation probes in each functional region of the genome. **(C)** is the density distribution of DNA methylation probes in each functional region of the genome normalized by length. **(D)** is the distribution ratio diagram of the hyper and hypo methylation probes of the CGI and the promoter region, **(E)** is the distribution of average DNA methylation levels in 12 regions of ADMPs in paracancerous tissues. (N) and breast cancer tissues (T). The abscissa bin1-bin140 is divided into 12 regions by 11 vertical lines, and the upper coordinate is the name of each region.

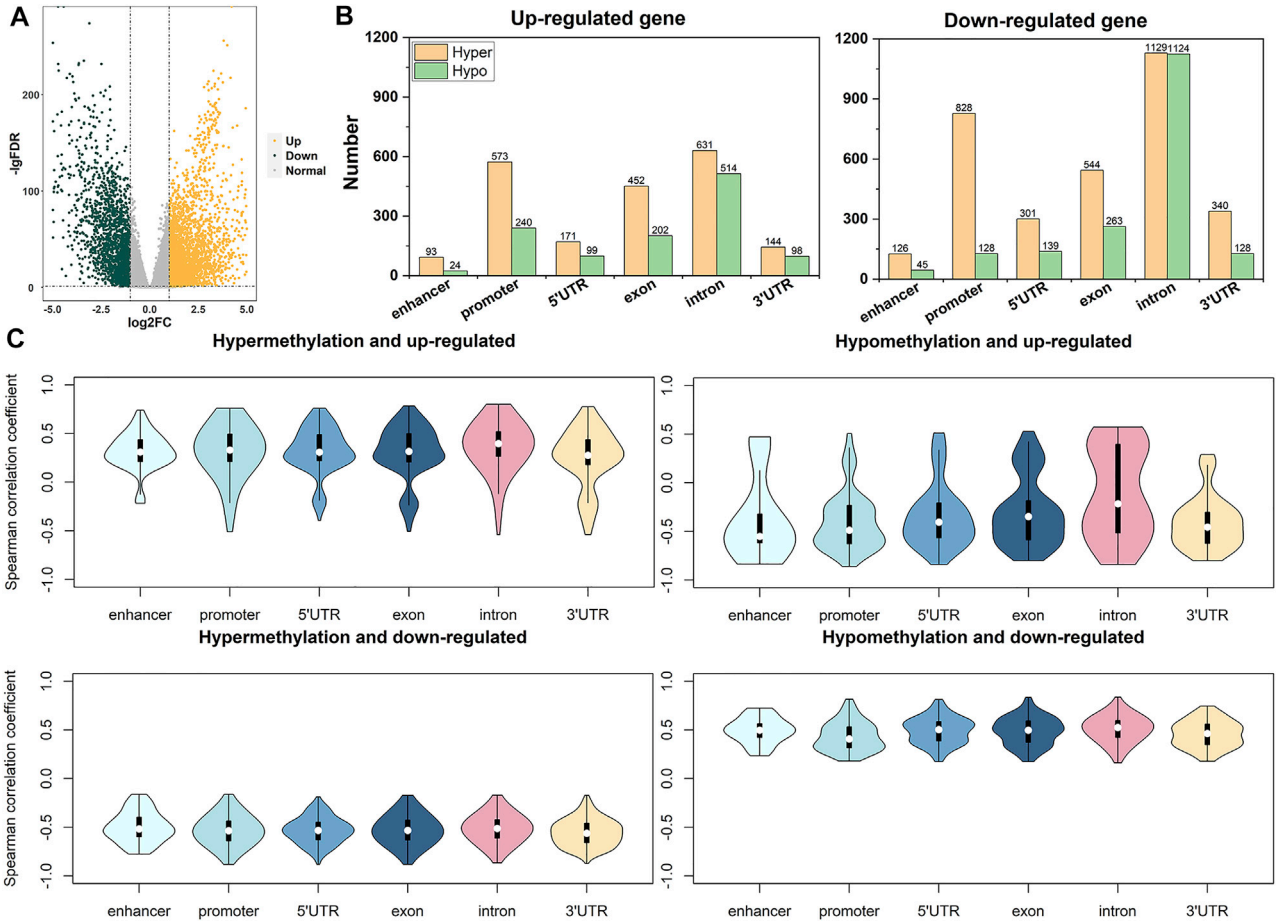


FIGURE 2 | The distribution of correlation coefficients between gene expression and DNA methylation. **(A)** is a volcano map for screening DEGs. **(B)** is the enrichment number of DNA methylation probes in each region of up-regulated and down-regulated genes. **(C)** is the correlation between DNA methylation and gene expression in each region of the four types of genes.

leaving 19,495 genes. Finally, the genes on chromosomes 1-22, X, and Y were retained, and a total of 19,484 genes encoding proteins were obtained for this study. We divided the genome into six regions {promoter (1,500 bp upstream and downstream of TSS), 5′UTR, exon, intron, 3′UTR, and intergenic region [from the transcription termination site (TTS) of one gene to the TSS of the next gene]}. In addition, we also selected the enhancer region and used the position of the enhancer to find the gene closest to it, and defined this gene as the target gene of the enhancer. After processing the location file of CGI, we obtained five regions [N_Shelf (2–4 kb upstream of CGI), N_Shore (0–2 kb upstream of CGI), CGI, S_Shore (0–2 kb downstream of CGI), and S_Shelf (2–4 kb downstream of CGI)]. Then we divided the promoter region into 30 windows in 100 bp, and each of the other 11 regions was divided into ten windows on average.

Calculation of Differentially Methylated CpG Sites and Average DNA Methylation Level

First, we integrated the downloaded DNA methylation data of 789 breast cancer and 96 paracancerous samples into a matrix. Then we used the Limma package for differential analysis (Ritchie et al., 2015). Finally, we selected the probe of $|\Delta\beta| \geq 0.2$, $p < 0.05$, adjusted p-value< 0.01, 14,855 hypermethylated probes and 11,056 hypomethylated probes were obtained (**Supplementary Table S3**). We collectively refer to hypermethylated probes and hypomethylated probes as abnormal DNA methylated probes (ADMPs).

$$\Delta\beta = \beta_{\omega,c} - \beta_{\omega,n} \quad (1)$$

here $\beta_{\omega,c}$ denotes the methylation level of the ω-th CpG site (probe) in the cancer sample, the $\beta_{\omega,n}$ denotes the methylation level of ω-th CpG site in the paracancerous sample.

We matched ADMPs to 12 different regions of the gene and calculated the DNA methylation level of each region for the gene. To better understand the abnormal DNA methylation characteristics of each region, we calculated the average DNA methylation level of each region according to the following formula:

$$\beta_b = \sum_{i=1}^{k} \beta_{b,i} / k$$

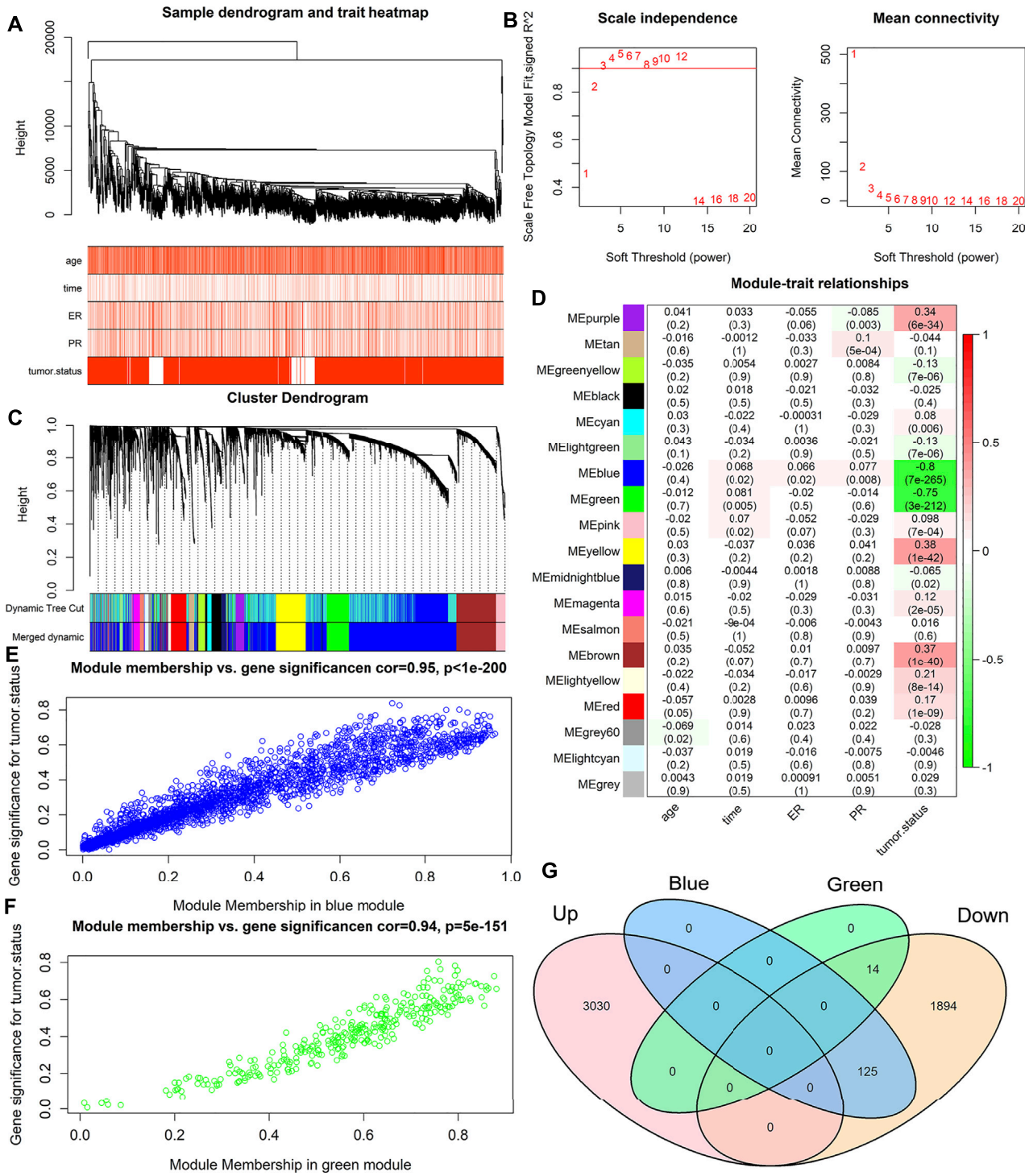


FIGURE 3 | WGCNA analysis of DEGs. **(A)** is a sample clustering tree diagram. **(B)** is the scale-free index of various soft threshold power, and the average connection degree of various soft threshold power. **(C)** is the module gene color block of cluster analysis. The first line is the module generated by the first clustering, and the second line is the merged module (when the similarity between different modules reaches 0.8, they are merged into one module). **(D)** is the heat map of the correlation coefficient between module genes and clinical features. **(E)** is the scatter plot of the distribution of the Module Membership (MM) and the Gene Significance (GS) in the blue module. **(F)** is the scatter plot of the distribution of MM and GS in the green module. **(G)** is the Venn diagram of the hub gene and the expression up-regulated gene and the expression down-regulated gene.

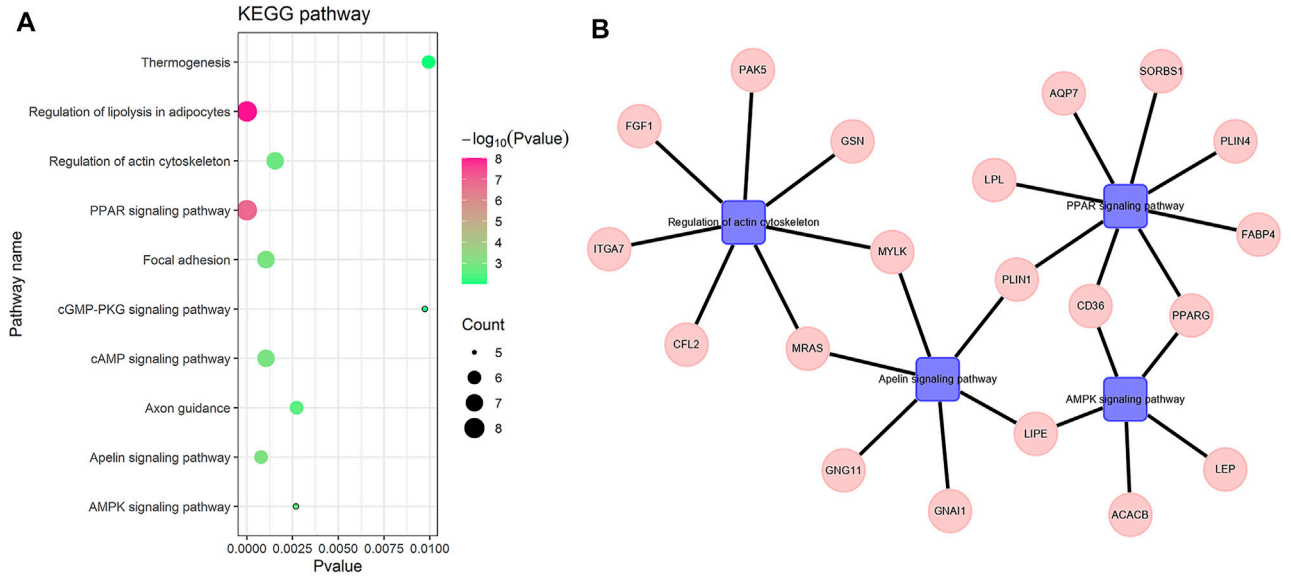


FIGURE 4 | Hub gene-enriched KEGG pathway. **(A)** is the bubble chart of the hub gene-enriched pathway. **(B)** is the pathway-enriched gene, the square is the pathway name, and the circle is the gene name.

$$\beta_{b,i} = \sum_{\omega=1}^{m} \beta_{\omega} / m \tag{2}$$

here β_b denotes the average DNA methylation level of the bth bin, $\beta_{b,i}$ denotes the methylation level of the bth bin in the ith gene, and k represents the number of genes whose DNA methylation level is not 0. β_{ω} denotes the DNA methylation level of the ω-th CpG site. m represents the number of probes falling into the bth bin of the ith gene.

Analysis of Differentially Expressed Genes

We used the DESeq2 package to process gene expression data (Love et al., 2014). First, the expression data were integrated and normalized into a matrix, and $\log_2$FC (foldchange = cancer/normal) >1, $p < 0.05$, adjusted p-value< 0.05, was used as the threshold. Finally, 5,063 differentially expressed genes (DEGs) were obtained, of which 3,030 genes were up-regulated, and 2,033 genes were down-regulated (**Supplementary Table S4**).

Selection of Co-Expressed Genes and Hub Genes

We used the WGCNA package to calculate the Pearson correlation coefficient between DEGs, and construct a similarity matrix. To better satisfy the structure of the scale-free network, the similarity matrix was transformed into a connectivity matrix through suitable soft thresholding. The topological overlap matrix (TOM) was calculated through the connectivity matrix, and finally, the degree of dissimilarity matrix was obtained through 1-TOM. Through the dissimilarity matrix, genes could be easily clustered to obtain different gene modules. When selecting the hub genes, first, we calculated the logarithm of the p-value after linear regression between the gene expression value and the clinical characteristics. It represents the relationship between gene expression and clinical characteristics in the module, that is, the Gene Significance (GS). If the absolute value of GS for a gene is greater, the biological significance of the gene is greater. Second, the Pearson correlation coefficient between the gene expression value and the characteristic gene of a given module was calculated, that is, the Module Membership (MM). The larger the MM value of a gene, the more important the gene is in the module (Langfelder and Horvath, 2008).

Correlation Analysis

We used 75 breast cancer samples with DNA methylation and gene expression in both cancerous and para-cancerous tissues. Due to the characteristics of DNA methylation and gene expression data, we used Spearman correlation to calculate the correlation between DNA methylation and gene expression, as shown in **Eq. 3**.

$$r_{\omega} = 1 - \frac{6\sum_{j=1}^{n}\left(rg_{\beta_{\omega,j}} - rg_{e_{i,j,\omega}}\right)^2}{n(n^2-1)} \tag{3}$$

here $rg_{\beta_{\omega,j}}$ denotes the order of the DNA methylation value of ω-th probe in the jth sample, $rg_{e_{i,j,\omega}}$ denotes the order of the gene expression value of the ith gene where the ω-th probe is located for the jth sample, r_{ω} denotes the Spearman correlation coefficient between the DNA methylation value of the ω-th probe and the expression value of its gene.

Prognostic Model Construction

We selected 11 genes because they are strongly correlated with tumor status, they are all enriched in pathways related to cancer, and their gene expression values are strongly correlated with DNA methylation. Then we used the methylation level of 27 CpG sites on

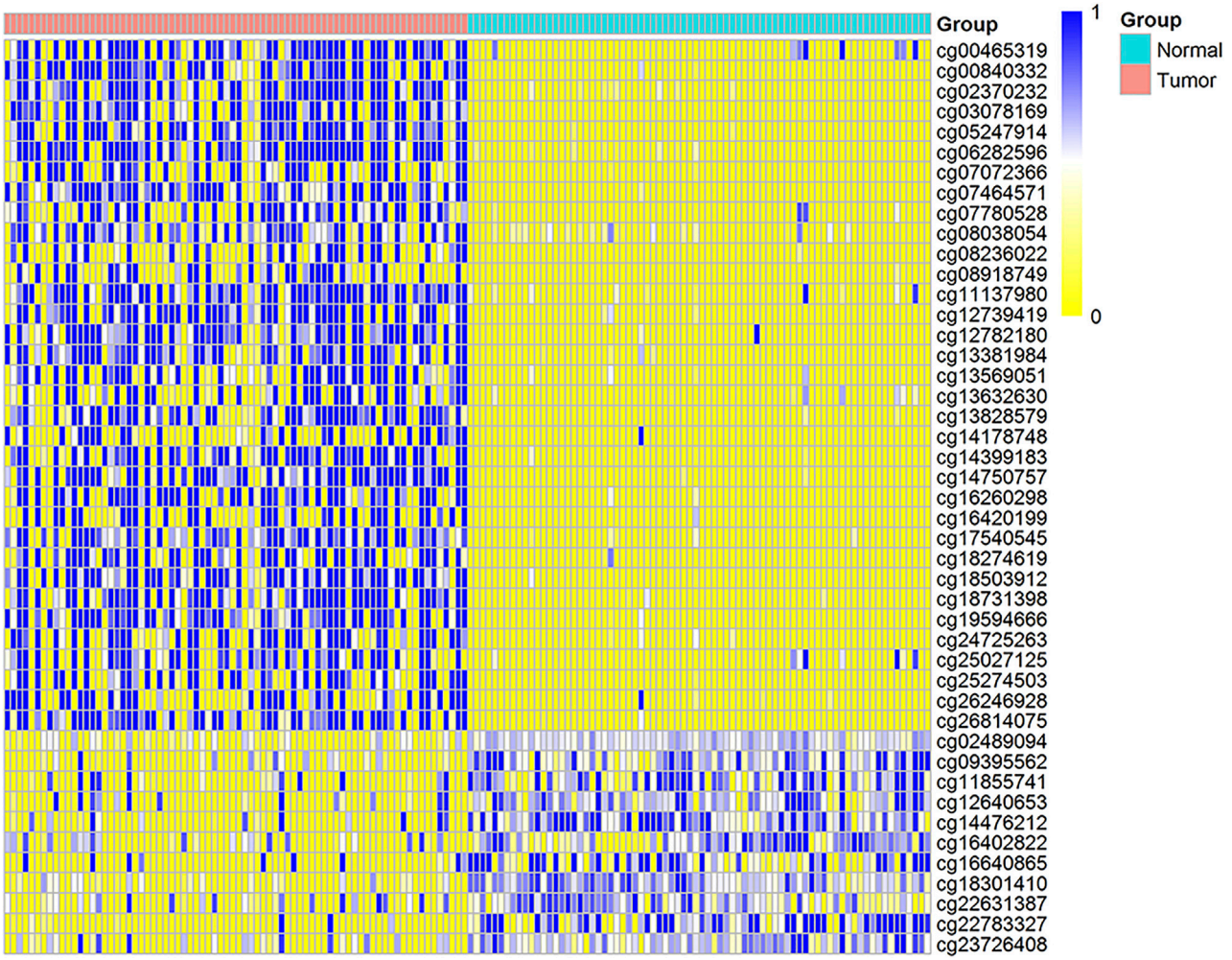


FIGURE 5 | The DNA methylation distribution map of probes for 17 key gene, the distribution heat map of the DNA methylation value of 45 CpG sites in 150 matched cancers and paracancerous samples. The row represents the DNA methylation value of each CpG site, and the column represents each sample. Blue indicates the trend of DNA methylation value close to 1, and yellow indicates that the DNA methylation value is close to 0.

the 11 genes and survival information to establish a multivariate Cox proportional hazard regression model, in which significant CpG sites were regarded as typical CpG sites. Finally, three methylation sites (cg25274503, cg13569051, and cg14399183) in the risk ratio model were determined. A risk score was established, the coefficient was weighted by the Cox model, and the risk score was calculated according to the following formula:

$$Risk\ Score = \sum Coef \times \beta \tag{4}$$

where $Coef$ represents the regression coefficient of the CpG site on the prognostic risk score, and β is the methylation level of CpG site.

RESULTS

Genome-Wide DNA Methylation Analysis

To understand the distribution of ADMPs in the breast cancer genome, we calculated the difference in DNA methylation. The 14,855 hypermethylated probes and 11,056 hypomethylated probes were obtained (**Figure 1A**). The ADMPs were matched to 12 regions of the genome (**Figures 1B,C,E**). In both sides of CGI, intergenic regions, and intron regions, the degree of enrichment for hypomethylated probes is higher than that for hypermethylated probes (**Figure 1B**). The hypermethylated probes are highly enriched in CGI and promoter regions. Then we normalized the number of probes distributed in each region according to the length of each region, and the distribution was shown in **Figure 1C**. The enrichment degree of hypermethylated probes is higher than that of the hypomethylated probes in CGI, enhancer regions, and promoter regions. In other regions, the enrichment degree of hypomethylated probes is higher than that of hypermethylated probes. By comparing **Figures 1B,C**, we found that hypermethylated probes are more significantly enriched in CGI and promoter regions after normalization by length. We found that the hypermethylated probes in CGI and promoter regions are highly enriched both before and after normalization. So, we took the intersection

TABLE 1 | Table of ADMPs significantly correlated with gene expression.

Style	Gene	cg id	Chr	Position	Location	Spearman correlation coefficient	*p* value
hyper	*CAV2*	cg12739419	chr7	116500539	Promoter, 1st intron, S_shore	–0.7466	0
		cg16260298	chr7	116500288	Promoter, 1st intron, CGI	–0.6980	0
		cg25274503	chr7	116500074	Promoter, 1st intron, CGI	–0.6353	0
	CFL2	cg25027125	chr14	34713595	3rd intron, N_shore	–0.7061	0
	FXYD1	cg03078169	chr19	35138887	Promoter, N_shelf, N_shore	–0.6494	0
		cg05247914	chr19	35138797	Promoter, N_shelf	–0.7427	0
		cg07780528	chr19	35139430	Promoter, N_shelf	–0.6864	0
		cg17540545	chr19	35139451	Promoter, N_shelf, N_shore	–0.7986	0
		cg18503912	chr19	35139375	Promoter, N_shelf, N_shore	–0.8249	0
	GNG11	cg08038054	chr7	93921469	Promoter	–0.6857	0
	GSN	cg13569051	chr9	121289425	5′UTR, 10th intron	–0.6657	0
		cg13828579	chr9	121306136	12th intron	–0.6756	0
		cg14399183	chr9	121286030	5′UTR, 10th intron	–0.7008	0
	LEP	cg00840332	chr7	128241216	Promoter, CGI	–0.6269	9.38e-18
		cg07464571	chr7	128240948	Promoter, CGI	–0.6032	3.10e-16
		cg12782180	chr7	128240879	Promoter, CGI	–0.6268	9.46e-18
		cg13381984	chr7	128241291	Promoter, 5′UTR, 1st exon, CGI	–0.6302	5.57e-18
		cg19594666	chr7	128241227	Promoter, CGI	–0.6331	3.53e-18
		cg26814075	chr7	128241245	Promoter, CGI	–0.6136	6.97e-17
	MGLL	cg18274619	chr3	127776009	Enhancer, 4th intron	–0.7373	0
	MYLK	cg00465319	chr3	123620721	3rd intron	–0.6799	0
		cg18731398	chr3	123695886	16th intron	–0.7483	0
	SEMA3G	cg11137980	chr3	52435210	5′UTR, 1st exon	–0.6044	0
	SORBS1	cg02370232	chr10	95415608	18th intron	–0.6763	0
		cg06282596	chr10	95415722	18th intron	–0.6528	0
hypo	*FXYD1*	cg22783327	chr19	35142354	5th intron, N_shore	0.6777	0
	PDE2A	cg16640865	chr11	72590514	24th exon, CGI	0.6494	0

between the hypermethylated probes in the CGI and the hypermethylated probes in the promoter region (**Figure 1D**) and found that most of the hypermethylated probes in the promoter region are located on the CGI. And we found that the number of hypermethylated probes is almost the same as the number of hypomethylated probes in the promoter region without CGI. In the promoter region with CGI, hypermethylated probes is as high as 97.87%. It can be speculated that the enrichment of hypermethylated probes in the promoter region is caused by the enrichment of CGI. Our results also confirmed that hypermethylated probes are mainly concentrated in CGI of the promoter regions.

We further explored the distribution of DNA methylation levels in the 12 regions (**Figure 1E**). In the paracancerous tissues, the DNA hypomethylated values are mainly concentrated in 0.65–0.85, the DNA hypermethylated values in each region are mainly concentrated in 0.15–0.3. In cancer tissues, the DNA hypomethylated values in each region are mainly concentrated in 0.4–0.55, the DNA hypermethylated values are mainly concentrated in 0.45–0.55. Regardless of whether the DNA methylation level of the cancer genome is a hypermethylated probe or a hypomethylated probe, the methylation level of each region of the gene is about 0.5.

Analysis of the Correlation Between DNA Methylation and Gene Expression in Different Functional Regions of DEGs

To understand the regulatory effect of DNA methylation on gene expression, we calculated the DEGs between breast cancer tissues and paracancerous tissues, of which 3,030 genes are up-regulated, and 2,033 genes are down-regulated (**Figure 2A**). We computed the ADMPs in various functional regions for up-regulated genes and down-regulated genes, as shown in **Figure 2B**. The total number of ADMPs in up-regulated genes is less than that in down-regulated genes. The number of up-regulated genes is about 1.5 times that of down-regulated genes, which further shows that ADMPs like to be enriched in down-regulated genes. The number of probes enrichment in the 5′UTR and 3′UTR of the up-regulated genes is almost the same, and the same is true in the down-regulated genes. The number of hypermethylated probes in the promoter and exon regions of up-regulated genes is about twice the number of hypomethylated probes, and the same pattern is observed in the exons of down-regulated genes. However, the number of hypermethylated probes in the promoter region of down-regulated genes is 6.5 times that of hypomethylated probes, which is quite different from up-regulated genes. From the up- and down-regulation of gene expression and the hyper- and hypo-methylation of probes, it can be divided into four categories (hypermethylated up-regulated genes, hypomethylated up-regulated genes, hypermethylated down-regulated genes, and hypomethylated down-regulated genes). We calculated the Spearman correlation coefficients between the gene expression value and each CpG probe falling into the functional region of the differentially expressed gene in these four types of genes. The correlation coefficients with $p < 0.05$ were selected for display, as shown in **Figure 2C**. We can see that the expression of

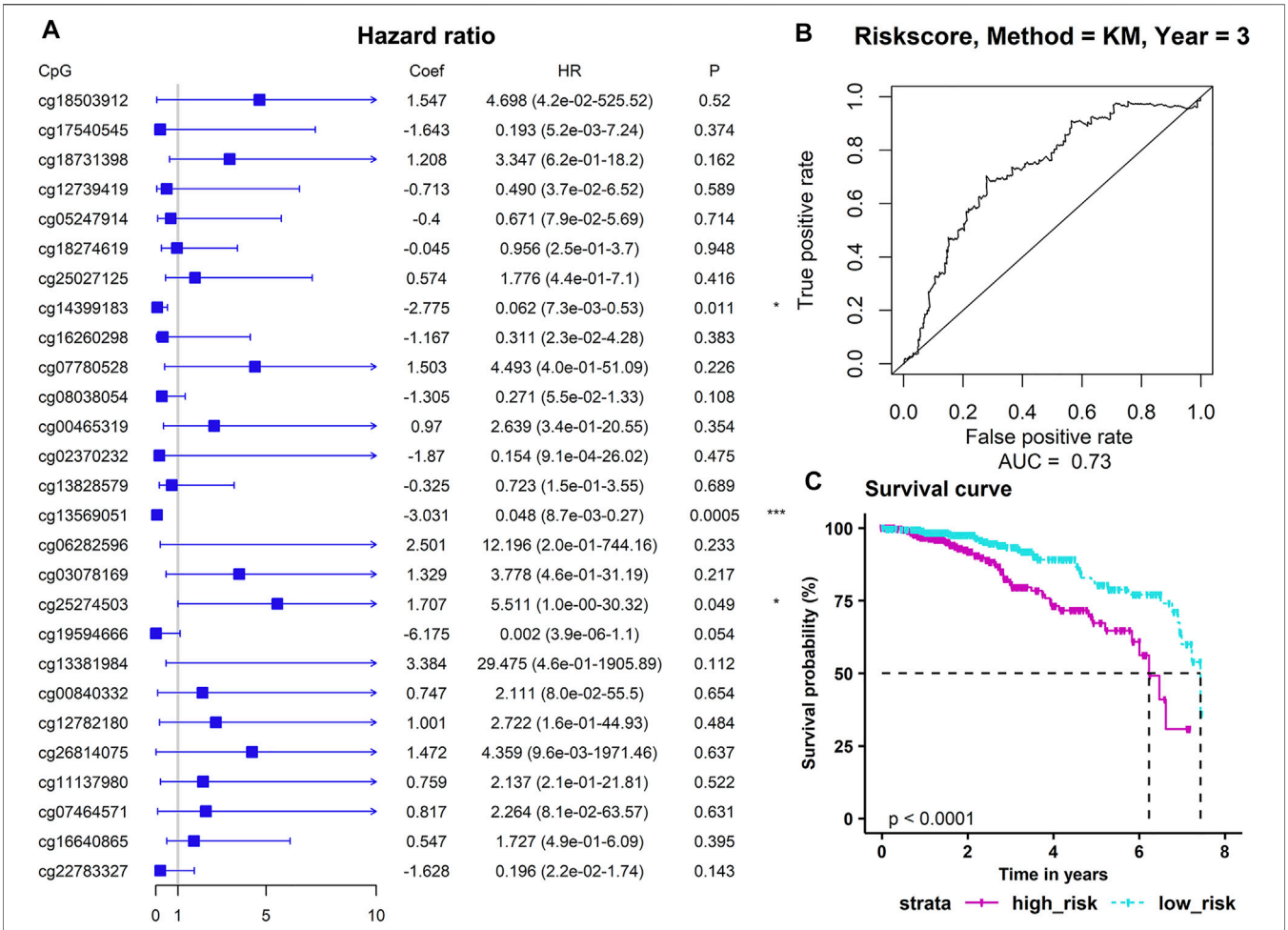


FIGURE 6 | Construction of a prognostic model. **(A)** The results of multivariate Cox regression analysis. *p*-values were shown as: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. **(B)** The sensitivity and specificity of the survival time of the sample. **(C)** The abscissa indicates the survival time (years), and the column indicates the survival rate.

hypomethylated up-regulated genes and hypermethylated down-regulated genes are negatively correlated with the DNA methylation level of the probes. The expression of hypermethylated up-regulated genes and hypomethylated down-regulated genes are positively correlated with the DNA methylation level of the probes. In summary, the abnormal DNA methylation of down-regulated genes is highly enriched. The most obvious manifestations are the promoter region. Among them, the number of hypermethylated probes in the promoter region of down-regulated genes is much higher than that of hypomethylated probes.

Discovery of Modular Genes Related to Breast Cancer and Screening of Hub Genes

The genes that affect breast cancer are not single, so we need to look for co-expressed gene clusters. We used 5,063 DEGs to find gene clusters related to breast cancer. First, we performed hierarchical clustering on all samples of DEGs and removed 14 outlier samples (**Figure 3A**). Then we constructed a scale-free network for the gene expression values of the remaining samples and chose three as the best soft threshold (**Figure 3B**). We set the minimum number of genes in each gene module to 30 and obtained 19 gene modules by clustering and merging similar modules (**Figure 3C**). At the same time, it can be seen that the gene expression is relatively independent between the modules (**Supplementary Figure S1**). Finally, the Pearson correlation analysis was carried out between 19 gene modules and clinical traits. **Figure 3D** shows that some module genes are strongly correlated with tumor status. Among them, MEblue and MEgreen show a strong negative correlation with tumor status, and the correlation coefficients are −0.8 (p = 7e-265) and −0.75 (p = 3e-212), respectively.

Because of the strong negative correlation between the above two gene modules and the tumor status, we calculated the GS and MM of each gene in blue and green modules. The results show that the MM and GS of genes in the blue module and the green module are highly linearly correlated (**Figures 3E,F**). Then we selected genes with MM value greater than or equal to 0.8 and GS value greater than or equal to 0.6 as hub genes that were highly

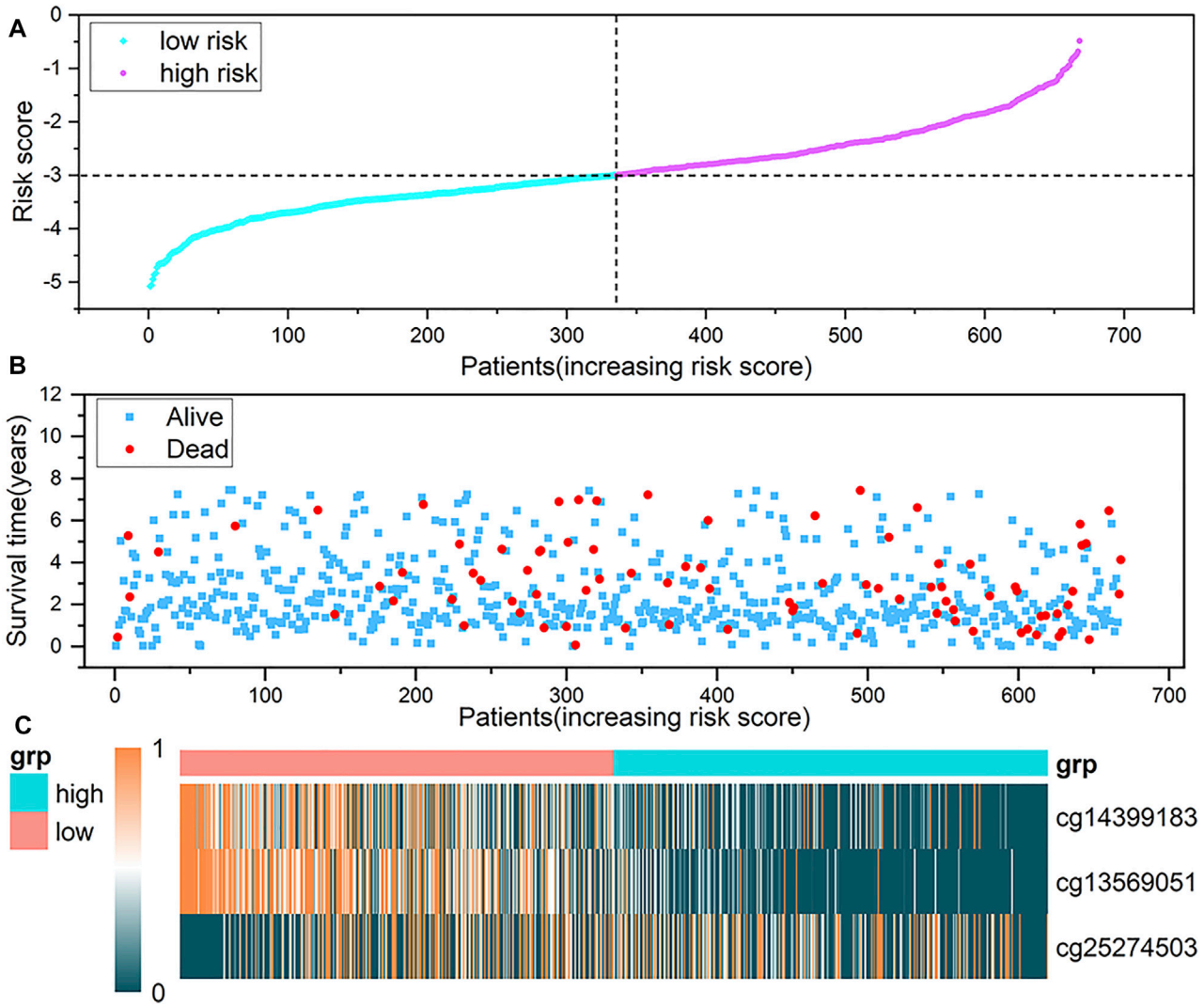


FIGURE 7 | Analysis of risk scoring model. **(A)** represents the characteristic risk score distribution of the three methylation probes. **(B)** is a scatter plot of the patient's survival time. **(C)** is the heat map of DNA methylation level, and a row represents a CpG probe, a column represents a patient. The three graphs have the same abscissa, and the abscissa is the patient in ascending order of risk score.

correlated with clinical features (**Supplementary Table S5**). 14 hub genes were obtained in the green module, and 125 hub genes were obtained in the blue module, and these hub genes were all down-regulated (**Figure 3G**). In summary, it is indicated that the down-regulation of the 139 hub genes we obtained may be related to the occurrence of breast cancer.

Pathway Analysis of Hub Gene

To understand the biological functions of the hub genes, we used GenCLiP3 (Wang et al., 2019) to analyze the pathways of these genes. We chose the KEGG pathway with p-value ≤ 0.01, Hit ≥ 5. It can be seen from **Figure 4** that these genes are mainly enriched in the PPAR signaling pathway, which can regulate lipid metabolism, adipogenesis, maintain metabolic homeostasis and inflammatory gene expression, and have anti-cancer effects in a variety of tumors (Ahmadian et al., 2013). The hub genes are also enriched in the AMPK signaling pathway. Because of AMPK's role in regulating energy homeostasis, AMPK is considered to be a potential target for developing new therapies for obesity, type 2 diabetes, metabolic syndrome, and cancer (Mihaylova and Shaw, 2011). The hub genes are also enriched in the Apelin signaling pathway, which is related to different key physiological processes, such as cell proliferation and energy metabolism regulation. On the other hand, it also involves a variety of pathologies, including diabetes, obesity, cardiovascular disease, and cancer (Antushevich and Wójcik, 2018). These genes are also enriched in the regulation of actin cytoskeleton pathway, which is mainly responsible for mediating various important cellular processes, including cell migration, proliferation, and survival (Saarikangas et al., 2010). These results indicate that hub genes are enriched in a variety of cancer-related pathways, further explaining the importance of these hub genes.

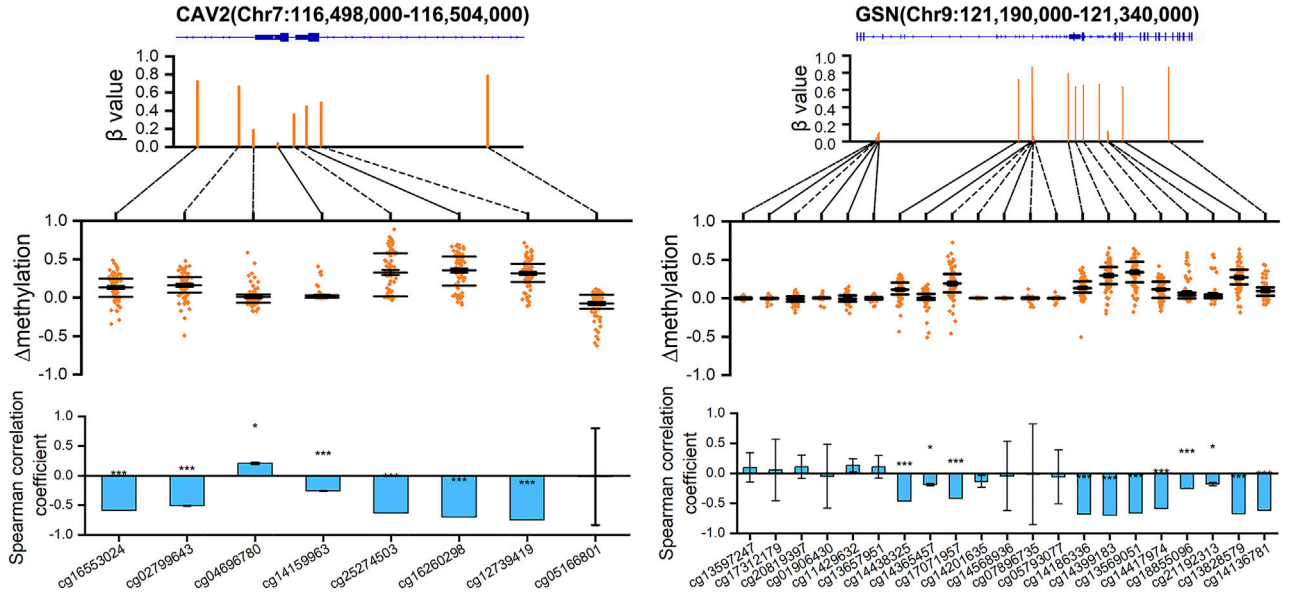


FIGURE 8 | DNA methylation distribution map of *CAV2* and *GSN*. The scatter graph shows the *Δβ* level, the histogram shows the correlation coefficient between the DNA methylation of each probe and the gene expression value (*p*-values were shown as: **p* < 0.05; ***p* < 0.01; ****p* < 0.001).

Correlation Analysis Between Key Gene Expression and DNA Methylation

We have obtained 34 genes enriched in the key pathways and called them key genes. To understand the influence mechanism of abnormal DNA methylation on the key genes whose expression is down-regulated, we matched probes for abnormal DNA methylation corresponding to the 34 genes. In the end, only 17 key genes have ADMPs. Among them, 34 hypermethylated probes were matched on 14 genes, 11 hypomethylated probes were matched on six genes, and there were both hypermethylated probes and hypomethylated probes on three genes, *MGLL*, *FXYD1*, and *MYLK*. **Figure 5** shows the DNA methylation value of the 45 probes on the cancer samples and matched paracancerous samples. We computed the Spearman correlation coefficients between the expression values of these 17 genes and the DNA methylation of 45 probes (**Supplementary Table S6**) and found that the hypomethylated probes for the key genes are mainly located in the intron and exon regions, and the DNA methylation of probes are positively correlated with gene expression. Then, we used 0.6 as the threshold, and the absolute value of the correlation coefficient greater than or equal to 0.6 was considered a significant correlation ($p < 0.05$). Finally, the DNA methylation values of the 27 probes are significantly correlated with the expression of 11 genes (**Table 1**). Among the 27 probes, the 25 probes are hypermethylated, and the expression of the 25 probes are significantly negatively correlated with their genes. Due to the overlap of gene annotation file regions, a probe may be in a different region. We see that most of these hypermethylated probes are located in the promoter, CGI, and intron, and the CGI are all contained in the promoter. Some hypermethylated probes are located in the 5′UTR region, the first exon, and enhancer regions. The cg16640865 and cg22783327 are hypomethylated, located in the 24th exon of *PDE2A* and the 5th intron of *FXYD1*, respectively. The results show that the CGI hypermethylation in the promoter for most key genes can inhibit gene expression. The hypomethylation of the 24th exon of *PDE2A* and the 5th intron of *FXYD1* also have an inhibiting effect on gene expression. This is consistent with the conclusion that the promoter methylation is negatively correlated with gene expression, and the methylation of the gene body region is positively correlated with gene expression.

The Construction and Evaluation of Prognostic Model

We used the above 27 ADMPs that were significantly related to key hub gene expression for regression analysis (**Figure 6A**). Three ADMPs (cg13569051, cg14399183, and cg25274503) were obtained using multivariate Cox analysis, and a Cox proportional hazard model was constructed. The formula of our model is: Risk Score = −3.0311 × β value of cg13569051 + (−2.7747) × β value of cg14399183 + 1.7067 × β value of cg25274503. According to the risk scoring formula, we performed receiver operating characteristic (ROC) analysis on the risk score of each sample, and the area under the curve is AUC = 0.73 (**Figure 6B**), which shows that the model is good for the prognosis. Then, we divided the patients into a high-risk group and a low-risk group using the median of the risk score as the dividing line. Through Kaplan-Meier survival analysis, we found that the survival time of the high-risk group was significantly lower than that of the low-risk group ($p < 0.0001$) (**Figure 6C**).

In addition, to observe whether the DNA methylation level of the probe in the model changes with the risk scoring model system, we sorted the samples according to the risk score. **Figures 7A,B** show a scatter plot of risk score distribution and patient status, where high risk is associated with more deaths. The heat map shows the methylation status of the three methylation probes between the high-risk group and the low-risk group (**Figure 7C**). The

methylation levels of cg14399183 and cg13569051 decrease as the risk increases. The methylation level of cg25274503 increases with increasing risk. The cg14399183 and cg13569051 probes located in the 5′UTR of the *GSN* gene are protective factors for breast cancer, while the cg25274503 probe located in the promoter of the *CAV2* gene is a risk factor for breast cancer.

DNA Methylation of *CAV2* and *GSN*

According to the above analysis, we can see the importance of *CAV2* and *GSN* for breast cancer. Therefore, we visualized the DNA methylation values of all eight probes on *CAV2* and all 21 probes on *GSN* ($n = 75$) (**Figure 8**). We calculated the $\Delta\beta$ value of each probe in the 75 samples and determined that the probes on *CAV2* and *GSN* genes are hypermethylated probes except for the median of $\Delta\beta$ about 0. Then we calculated the correlation between the DNA methylation value of each probe and the gene expression level. We see that all hypermethylated probes are significantly negatively correlated with the down-regulated *CAV2* and *GSN* in breast cancer. These results indicate that the down-regulation of the two genes, *CAV2* and *GSN*, is caused by hypermethylation of important DNA methylation sites.

DISCUSSION

In this study, based on analysis of the correlation between abnormal DNA methylation in six different regions and gene expression of DEGs, co-expression analysis, and KEGG pathway analysis, 34 key hub genes co-expressed and strongly correlated with cancer status were obtained. The 27 ADMPs that were significantly related to gene expression were obtained. Based on three methylation probes (cg13569051, cg14399183, and cg25274503) in the *CAV2* and *GSN* genes, a risk scoring model with good prognostic performance was constructed. It is further confirmed that the three probes can be used as molecular targets for breast cancer.

Our results showed that in breast cancer samples, the hypermethylation of cg25274503 in the promoter was significantly negatively correlated with the down-regulation of *CAV2*. This rule is also reflected in other studies. For example, the hypermethylation of CGI silences the *CAV2* gene, which can be used as an obvious marker of breast cancer (Uehiro et al., 2016). At the same time, the hypermethylation of cg13569051 and cg14399183 in the 5′UTR was significantly negatively correlated with the down-regulation of *GSN*. Similar findings have been reported in previous studies. For example, *GSN* is down-regulated in gastric cancer cell lines, and promoter DNA methylation is involved in this process (Wang et al., 2017). In addition, the *CAV1* gene is highly methylated and lowly expressed (Li et al., 2016). In ER breast cancer patients, *FOXA1* hypermethylation is associated with the down-regulation of gene expression (Espinal et al., 2017). The silencing of gene expression by *PAQR3* promoter hypermethylation may play an important role in breast cancer (Nowak et al., 2017). These reflect the negative correlation between DNA methylation in the promoter region and gene expression. However, in previous studies on breast cancer, there was almost no discovery of the relationship between the DNA methylation of cg25274503 and *CAV2* expression and the DNA methylation of cg13569051 and cg14399183 and the *GSN* expression.

In fact, *CAV2* is a gene encoding caveolin 2, which is involved in basic cell functions, including signal transduction, lipid metabolism, control of cell growth, and apoptosis; it may have tumor suppressor effects (Fujimoto et al., 2001). In all types of lung cancer, *CAV2* is dysregulated at the RNA and protein levels (Wikman et al., 2004). Experiments have verified that the *CAV2* gene transcription is down-regulated in mice and humans with obstructive bladder disease (Thangavel et al., 2019). Studies have confirmed that compared with the corresponding normal tissues, the mRNA level of *CAV2* in human breast cancer tissues is significantly down-regulated ($p < 0.001$) (Sagara et al., 2004). *GSN* is an actin binding protein, a key regulator of actin filament assembly and disassembly, and is involved in cell movement, shape, and metabolism (Feldt et al., 2019). Studies have shown that *GSN* gene transcription is down-regulated in breast cancer of humans and some animals, and the activation of *GSN* may be a protective factor in the treatment of cancer cells against cancer (Mielnicki et al., 1999; Shahrokh et al., 2019). The secreted *GSN* inhibits the invasion and migration of colon cancer cells (Chen et al., 2019). The expression of *GSN* in bladder cancer is higher than that in normal tissues, and the prognosis of bladder cancer patients whose gene expression of *GSN* is up-regulated is worse (Yang et al., 2020). *GSN* is overexpressed in HCC tissues, and high *GSN* expression is significantly associated with advanced Edmondson grade, encapsulation, and multiple tumors (Zhang et al., 2020). These different studies show that the role of *GSN* in cancer depends on the type of cancer studied. It has been suggested that *GSN* has both the functions of tumor suppressor genes and oncogenes (Feldt et al., 2019). It can be seen from the above research that *CAV2* and *GSN* have important roles in a variety of cancers. Our study also found that *CAV2* and *GSN* were down-regulated in breast cancer, which is consistent with the results of previous studies.

In summary, our research has found two key genes (*CAV2* and *GSN*) related to breast cancer that may be regulated by DNA methylation and discovered three DNA methylation probes (cg13569051, cg14399183, and cg25274503). The risk scoring model was constructed by the three probes has a good prognostic ability. Therefore, these DNA methylation probes may be used as molecular targets for the prognosis of breast cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Gene expression data, DNA methylation data and clinical data from TCGA database (https://tcga-data.nci.nih.gov/tcga/); the genomic data from the UCSC database (http://genome.ucsc.edu/cgi-bin/hgTables); Ensembl (http://www.ensembl.org/Homo_sapiens/); the location file of CGI from UCSC (http:/genome.ucsc.edu/); the location file of an enhancer from the FANTOM5 (Function Annotation of The Mammalian Genome) (https://fantom.gsc.riken.jp/5/).

AUTHOR CONTRIBUTIONS

Y-NC conceived the research and participated in data analysis, visualization, and wrote the manuscript. Q-ZL designed this idea

and were involved in the discussion and revision of the article. Y-XL participated in data analysis, result discussion, and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (nos. 32160216, 31870838, and 61861035).

ACKNOWLEDGMENTS

We thank all colleagues in our laboratory for their generous suggestions and support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.815843/full#supplementary-material

REFERENCES

Ahmadian, M., Suh, J. M., Hah, N., Liddle, C., Atkins, A. R., Downes, M., et al. (2013). PPARγ Signaling and Metabolism: the Good, the Bad and the Future. *Nat. Med.* 19, 557–566. doi:10.1038/nm.3159

Antushevich, H., and Wójcik, M. (2018). Review: Apelin in Disease. *Clinica Chim. Acta* 483, 241–248. doi:10.1016/j.cca.2018.05.012

Baylin, S. B. (2005). DNA Methylation and Gene Silencing in Cancer. *Nat. Rev. Clin. Oncol.* 2 (Suppl. 1), S4–S11. doi:10.1038/ncponc0354

Borgonio-Cuadra, V. M., Miranda-Duarte, A., Rojas-Toledo, X., Garcia-Hernandez, N., Alfredo Sierra-Ramirez, J., Cardenas-Garcia, M., et al. (2018). Association between Promoter Hypermethylation of the DACT2 Gene and Tumor Stages in Breast Cancer. *J. BUON.* 23, 361–365.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J. Clinicians* 68, 394–424. doi:10.3322/caac.21492

Bu, D., Lewis, C. M., Sarode, V., Chen, M., Ma, X., Lazorwitz, A. M., et al. (2013). Identification of Breast Cancer DNA Methylation Markers Optimized for Fine-Needle Aspiration Samples. *Cancer Epidemiol. Biomarkers Prev.* 22, 2212–2221. doi:10.1158/1055-9965.EPI-13-0208

Cedar, H., and Bergman, Y. (2009). Linking DNA Methylation and Histone Modification: Patterns and Paradigms. *Nat. Rev. Genet.* 10, 295–304. doi:10.1038/nrg2540

Chen, Z., Li, K., Yin, X., Li, H., Li, Y., Zhang, Q., et al. (2019). Lower Expression of Gelsolin in Colon Cancer and its Diagnostic Value in Colon Cancer Patients. *J. Cancer* 10, 1288–1296. doi:10.7150/jca.28529

Croes, L., Beyens, M., Fransen, E., Ibrahim, J., Vanden Berghe, W., Suls, A., et al. (2018). Large-scale Analysis of DFNA5 Methylation Reveals its Potential as Biomarker for Breast Cancer. *Clin. Epigenet.* 10, 51. doi:10.1186/s13148-018-0479-y

Espinal, A. C., Buas, M. F., Wang, D., Cheng, D. T.-Y., Sucheston-Campbell, L., Hu, Q., et al. (2017). FOXA1 Hypermethylation: Link between Parity and ER-Negative Breast Cancer in African American Women? *Breast Cancer Res. Treat.* 166, 559–568. doi:10.1007/s10549-017-4418-y

Feldt, J., Schicht, M., Garreis, F., Welss, J., Schneider, U. W., and Paulsen, F. (2019). Structure, Regulation and Related Diseases of the Actin-Binding Protein Gelsolin. *Expert Rev. Mol. Med.* 20, e7. doi:10.1017/erm.2018.7

Fujimoto, T., Kogo, H., Ishiguro, K., Tauchi, K., and Nomura, R. (2001). Caveolin-2 Is Targeted to Lipid Droplets, a New "Membrane Domain" in the Cell. *J. Cell Biol.* 152, 1079–1086. doi:10.1083/jcb.152.5.1079

Győrffy, B., Bottai, G., Fleischer, T., Munkácsy, G., Budczies, J., Paladini, L., et al. (2016). Aberrant DNA Methylation Impacts Gene Expression and Prognosis in Breast Cancer Subtypes. *Int. J. Cancer* 138, 87–97. doi:10.1002/ijc.29684

Janostiak, R., Vyas, M., Cicek, A. F., Wajapeyee, N., and Harigopal, M. (2018). Loss of C-KIT Expression in Breast Cancer Correlates with Malignant Transformation of Breast Epithelium and Is Mediated by KIT Gene Promoter DNA Hypermethylation. *Exp. Mol. Pathol.* 105, 41–49. doi:10.1016/j.yexmp.2018.05.011

Jin, W., Li, Q.-Z., Zuo, Y.-C., Cao, Y.-N., Zhang, L.-Q., Hou, R., et al. (2019). Relationship between DNA Methylation in Key Region and the Differential Expressions of Genes in Human Breast Tumor Tissue. *DNA Cell Biol.* 38, 49–62. doi:10.1089/dna.2018.4276

Jones, P. A., and Baylin, S. B. (2007). The Epigenomics of Cancer. *Cell* 128, 683–692. doi:10.1016/j.cell.2007.01.029

Jones, P. A. (2012). Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and beyond. *Nat. Rev. Genet.* 13, 484–492. doi:10.1038/nrg3230

Karsli-Ceppioglu, S., Dagdemir, A., Judes, G., Ngollo, M., Penault-Llorca, F., Pajon, A., et al. (2014). Epigenetic Mechanisms of Breast Cancer: an Update of the Current Knowledge. *Epigenomics* 6, 651–664. doi:10.2217/epi.14.59

Kulis, M., and Esteller, M. (2010). DNA Methylation and Cancer. *Adv. Genet.* 70, 27–56. doi:10.1016/B978-0-12-380866-0.60002-2

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Li, Z., Heng, J., Yan, J., Guo, X., Tang, L., Chen, M., et al. (2016). Integrated Analysis of Gene Expression and Methylation Profiles of 48 Candidate Genes in Breast Cancer Patients. *Breast Cancer Res. Treat.* 160, 371–383. doi:10.1007/s10549-016-4004-8

Liu, L., Shen, H., and Wang, Y. (2017). CRY2 Is Suppressed by FOXM1 Mediated Promoter Hypermethylation in Breast Cancer. *Biochem. Biophysical Res. Commun.* 490, 44–50. doi:10.1016/j.bbrc.2017.06.003

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Mielnicki, L. M., Ying, A. M., Head, K. L., Asch, H. L., and Asch, B. B. (1999). Epigenetic Regulation of Gelsolin Expression in Human Breast Cancer Cells. *Exp. Cell Res.* 249, 161–176. doi:10.1006/excr.1999.4461

Mihaylova, M. M., and Shaw, R. J. (2011). The AMPK Signalling Pathway Coordinates Cell Growth, Autophagy and Metabolism. *Nat. Cell Biol.* 13, 1016–1023. doi:10.1038/ncb2329

Nilsson, E., Jansson, P. A., Perfilyev, A., Volkov, P., Pedersen, M., Svensson, M. K., et al. (2014). Altered DNA Methylation and Differential Expression of Genes Influencing Metabolism and Inflammation in Adipose Tissue from Subjects with Type 2 Diabetes. *Diabetes* 63, 2962–2976. doi:10.2337/db13-1459

Nowak, E. M., Poczęta, M., and Bednarek, I. (2017). DNA Methyltransferase Inhibitors Influence on the DIRAS3 and STAT3 Expression and *In Vitro* Migration of Ovarian and Breast Cancer Cells. *Ginekol Pol.* 88, 543–551. doi:10.5603/GP.a2017.0099

Punglia, R. S., Morrow, M., Winer, E. P., and Harris, J. R. (2007). Local Therapy and Survival in Breast Cancer. *N. Engl. J. Med.* 356, 2399–2405. doi:10.1056/NEJMra065241

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Saarikangas, J., Zhao, H., and Lappalainen, P. (2010). Regulation of the Actin Cytoskeleton-Plasma Membrane Interplay by Phosphoinositides. *Physiol. Rev.* 90, 259–289. doi:10.1152/physrev.00036.2009

Sagara, Y., Mimori, K., Yoshinaga, K., Tanaka, F., Nishida, K., Ohno, S., et al. (2004). Clinical Significance of Caveolin-1, Caveolin-2 and HER2/neu mRNA Expression in Human Breast Cancer. *Br. J. Cancer* 91, 959–965. doi:10.1038/sj.bjc.6602029

Shahrokh, S., Mansouri, V., and Razzaghi, M. (2019). Assessment of the SRC Inhibition Role in the Efficacy of Breast Cancer Radiotherapy. *J. Lasers Med. Sci.* 10, S18–S22. doi:10.15171/jlms.2019.S4

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer Statistics, 2020. *CA: A Cancer J. Clinicians* 70, 7–30. doi:10.3322/caac.21590

Stefanska, B., Huang, J., Bhattacharyya, B., Suderman, M., Hallett, M., Han, Z.-G., et al. (2011). Definition of the Landscape of Promoter DNA Hypomethylation in Liver Cancer. *Cancer Res.* 71, 5891–5903. doi:10.1158/0008-5472.CAN-10-3823

Tang, J., Kong, D., Cui, Q., Wang, K., Zhang, D., Gong, Y., et al. (2018). Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis. *Front. Oncol.* 8, 374. doi:10.3389/fonc.2018.00374

Thangavel, C., Gomes, C. M., Zderic, S. A., Javed, E., Addya, S., Singh, J., et al. (2019). NF-κB and GATA-Binding Factor 6 Repress Transcription of Caveolins in Bladder Smooth Muscle Hypertrophy. *Am. J. Pathol.* 189, 847–867. doi:10.1016/j.ajpath.2018.12.013

Uehiro, N., Sato, F., Pu, F., Tanaka, S., Kawashima, M., Kawaguchi, K., et al. (2016). Circulating Cell-free DNA-Based Epigenetic Assay Can Detect Early Breast Cancer. *Breast Cancer Res.* 18, 129. doi:10.1186/s13058-016-0788-z

Wang, H.-C., Chen, C.-W., Yang, C.-L., Tsai, I.-M., Hou, Y.-C., Chen, C.-J., et al. (2017). Tumor-Associated Macrophages Promote Epigenetic Silencing of Gelsolin through DNA Methyltransferase 1 in Gastric Cancer Cells. *Cancer Immunol. Res.* 5, 885–897. doi:10.1158/2326-6066.CIR-16-0295

Wang, J.-H., Zhao, L.-F., Wang, H.-F., Wen, Y.-T., Jiang, K.-K., Mao, X.-M., et al. (2019). GenCLiP 3: Mining Human Genes' Functions and Regulatory Networks from PubMed Based on Co-occurrences and Natural Language Processing. *Bioinformatics* 36, 1973–1975. doi:10.1093/bioinformatics/btz807

Wikman, H., Seppänen, J. K., Sarhadi, V. K., Kettunen, E., Salmenkivi, K., Kuosma, E., et al. (2004). Caveolins as Tumour Markers in Lung Cancer Detected by Combined Use of cDNA and Tissue Microarrays. *J. Pathol.* 203, 584–593. doi:10.1002/path.1552

Wittenberger, T., Sleigh, S., Reisel, D., Zikan, M., Wahl, B., Alunni-Fabbroni, M., et al. (2014). DNA Methylation Markers for Early Detection of Women's Cancer: Promise and Challenges. *Epigenomics* 6, 311–327. doi:10.2217/epi.14.20

Yang, J.-L., Wang, C. C. N., Cai, J.-H., Chou, C.-Y., Lin, Y.-C., and Hung, C.-C. (2020). Identification of GSN and LAMC2 as Key Prognostic Genes of Bladder Cancer by Integrated Bioinformatics Analysis. *Cancers* 12, 1809. doi:10.3390/cancers12071809

Yang, X., Han, H., De Carvalho, D. D., Lay, F. D., Jones, P. A., and Liang, G. (2014). Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell* 26, 577–590. doi:10.1016/j.ccr.2014.07.028

Zhang, Y., Luo, X., Lin, J., Fu, S., Feng, P., Su, H., et al. (2020). Gelsolin Promotes Cancer Progression by Regulating Epithelial-Mesenchymal Transition in Hepatocellular Carcinoma and Correlates with a Poor Prognosis. *J. Oncol.* 2020, 1–10. doi:10.1155/2020/1980368

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Cao, Li and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 11 February 2022
doi: 10.3389/fcell.2021.811957

Multi-Omics Prognostic Signatures Based on Lipid Metabolism for Colorectal Cancer

*YuanLin Sun, Bin Liu, YuJia Chen, YanPeng Xing and Yang Zhang**

Department of Gastrointestinal Surgery, The First Hospital of Jilin University, Changchun, China

Background: The potential biological processes and laws of the biological components in malignant tumors can be understood more systematically and comprehensively through multi-omics analysis. This study elaborately explored the role of lipid metabolism in the prognosis of colorectal cancer (CRC) from the metabonomics and transcriptomics.

Methods: We performed K-means unsupervised clustering algorithm and t test to identify the differential lipid metabolites determined by liquid chromatography tandem mass spectrometry (LC-MS/MS) in the serum of 236 CRC patients of the First Hospital of Jilin University (JLUFH). Cox regression analysis was used to identify prognosis-associated lipid metabolites and to construct multi-lipid-metabolite prognostic signature. The composite nomogram composed of independent prognostic factors was utilized to individually predict the outcome of CRC patients. Glycerophospholipid metabolism was the most significant enrichment pathway for lipid metabolites in CRC, whose related hub genes (GMRHGs) were distinguished by gene set variation analysis (GSVA) and weighted gene co-expression network analysis (WGCNA). Cox regression and least absolute shrinkage and selection operator (LASSO) regression analysis were utilized to develop the prognostic signature.

Results: Six-lipid-metabolite and five-GMRHG prognostic signatures were developed, indicating favorable survival stratification effects on CRC patients. Using the independent prognostic factors as variables, we established a composite nomogram to individually evaluate the prognosis of CRC patients. The AUCs of one-, three-, and five-year ROC curves were 0.815, 0.815, and 0.805, respectively, showing auspicious prognostic accuracy. Furthermore, we explored the potential relationship between tumor microenvironment (TME) and immune infiltration. Moreover, the mutational frequency of TP53 in the high-risk group was significantly higher than that in the low-risk group ($p < 0.001$), while in the coordinate mutational status of TP53, the overall survival of CRC patients in the high-risk group was significantly lower than that in low-risk group with statistical differences.

Conclusion: We identified the significance of lipid metabolism for the prognosis of CRC from the aspects of metabonomics and transcriptomics, which can provide a novel perspective for promoting individualized treatment and revealing the potential molecular biological characteristics of CRC. The composite nomogram including a six-lipid-metabolite prognostic signature is a promising predictor of the prognosis of CRC patients.

Keywords: colorectal cancer, prognosis, multi omics, lipid metabolism, prognostic signature, tumor immune infiltration, tumor microenvironment, somatic mutation

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Zhi Li,
The First Affiliated Hospital of China Medical University, China
Sha Cao,
Indiana University Bloomington, United States

****Correspondence:***
Yang Zhang
yangzhang@jlu.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *09 November 2021*
Accepted: *08 December 2021*
Published: *11 February 2022*

Citation:
Sun Y, Liu B, Chen Y, Xing Y and Zhang Y (2022) Multi-Omics Prognostic Signatures Based on Lipid Metabolism for Colorectal Cancer. Front. Cell Dev. Biol. 9:811957. doi: 10.3389/fcell.2021.811957

INTRODUCTION

The morbidity and mortality of CRC remained stubbornly high in the world, respectively ranking the third and fourth (Gao et al., 2016; Zhang et al., 2017). The progress of CRC was associated with the hereditary factor, age, chronic inflammation, lifestyle such as smoking and drinking, dietary habits, and environmental factors (Aran et al., 2016). At present, the improvement of the prognosis of CRC patients was still fairly limited, mainly because a substantial portion of CRC patients were often asymptomatic in the early stage. However, the mechanism of occurrence, development, and invasion of CRC was still well unclear. Therefore, to find the molecular mechanism and favorable prognostic monitoring index for the occurrence, development and invasion of CRC was the hotspot of the current research.

At present, a large number of studies have shown that abnormal lipid metabolism was closely correlated with the occurrence and development of tumor. The accelerated process of malignant transformation and cancer cell proliferation required more energy, which could induce the dysfunction of lipid metabolism to allow cancer cells to survive (Santos and Schulze, 2012). It has been reported that higher level of plasma total cholesterol (TC) was associated with an increased incidence of CRC (Zalloua et al., 2019), and the absorption of polyunsaturated fatty acids in the diet would moderate the level of TC, thereby reducing the risk of CRC (Muka et al., 2016). In some prospective studies, the level of TC was positively correlated with the occurrence of breast cancer, prostate cancer, and colon cancers. Contrarily, TC level was negatively correlated with the occurrence of hepatocellular cancer, gastric cancer, and lung cancer (Kitahara et al., 2011). The previous studies reported that the levels of low-density lipoprotein (LDL) level in nearly two-fifths of CRC patients presented rising trend, while the high-density lipoprotein (HDL) level in most patients (88.3%) were normal (Liao et al., 2015; Rodriguez-Broadbent et al., 2017). In view of the fact that the dysfunction of lipid metabolism played an important role in genesis and development of tumor, the biological activity of lipid metabolism has always been the key research area screening tumor therapeutic targets. ACOT8 was a kind of peroxisome lipolysis-related enzyme that catalyzed the decomposition of acyl CoA into free fatty acids and CoA for β-oxidation. It has been suggested that ACOT8 was overexpressed in certain kinds of cancers (Ramakrishna et al., 2010; Hung et al., 2014). What is more, downregulated expression of FASN *via* RNAi technique had a significant effect on lipid metabolism inhibition and TG storage of metastatic prostate cancer cells in human lymph nodes (De Schrijver et al., 2003). Since the survival of tumor cells mainly depended on FASN-mediated *de novo* synthesis of fatty acids, FASN was considered as the one of the important targets for human cancer therapy (Lupu and Menendez, 2006).

Metabonomics, an important branch of systematic biology, was mainly utilized to study the changes of metabolites in the dynamic process of metabolism to uncover the metabolic features of life activities. The results obtained from the examination for the metabolites could reflect the pathophysiological state of the organism more accurately and directly. Nowadays, blood and urine were the main research objects, and the accuracies of test results were guaranteed by removing impure metabolites (such as protein or saccharide). The applications of metabonomics (such as LC-MS (Yin et al., 2006), GC-MS (Tian et al., 2016), and CE-MS (DeLaney et al., 2018) and so on) in cancer research were intended to improve the diagnosis and prognosis of cancer, and was employed to distinct potential cancer biomarkers. Tenori et al. found that the recurrence of breast cancer was associated with the decrease of histidine level and the increase of blood glucose level and blood lipid level through the metabonomic analysis for the serum of patients with early breast cancer and metastatic breast cancer (Tenori et al., 2015). Lodi et al. published an NMR-based metabonomic study, which analyzed serum and urine samples from myeloma patients with different stages and metabolic patterns associated with disease progression to identify useful markers for myeloma patients. In this way, patients are grouped based on recurrence or remission (Lodi et al., 2013).

In this study, we firstly utilized LC-MS/MS to determine differential lipid metabolites altas based on the serum samples from 236 CRC patients. The functional enrichment analysis and multi-lipid metabolite prognostic signature analysis revealed the potential biological function and prognostic significance of differential lipid metabolites. Moreover, the prognosis of CRC patients was systematically evaluated by a composite nomogram. In the meantime, aiming at the most significant functional enrichment pathway enriching differential lipid metabolites, namely, glycerophospholipid metabolic pathway, we analyzed the prognostic value of GMRHG from the transcriptional level, and constructed the five-GMRHG prognostic signature. In addition, based on the five-GMRHG prognostic signature, we further expanded the significance of GMRHG in tumor immune infiltration, tumor mutation landscape, and antineoplastic therapy.

METHODS

Sample Source

In this study, the serum samples were collected from 236 patients (all patients voluntarily signed informed consent forms after being notified full details of the study) hospitalized in the JLUFH from 2008 to 2013, who were pathologically diagnosed with primary CRC, had not been treated with hormonal therapy or chemoradiotherapy, and did not have acute inflammatory reactions. The exclusion criteria were (Zhang et al., 2017) congenital disease, metabolic disease, hematological disease, chronic inflammatory disease, severe cerebro-cardiovascular disease, respiratory disease, liver and renal disease, mental disease (Gao et al., 2016); women who were pregnant or could not rule out possibility of pregnancy and breastfeeding (Aran et al., 2016); alcoholics, drug addicts,

long-term use of proton pump inhibitors and hormones or non-steroidal anti-inflammatory drugs (Santos and Schulze, 2012); patients with any acute symptoms or major stress response (such as mental trauma or burns) in the past 2 weeks; and patients who had a cubital venous blood collected on an empty stomach in the early morning. Then, the blood samples were centrifuged for 2 min with 3500 r/min. The upper serum was frozen and stored at −80°C until being used to analyze.

Metabolic Analysis

Each serum sample taken from refrigerator with temperatures as low as 80°C was dissolved in 500 μl acetonitrile. Then, the serum samples were blended with the vortex device and centrifuged with 14000 r/min at 4°C for 5 min, and the supernatant of each sample was taken for LC-MS analysis, which was conducted using the triple-quadrupole tandem mass spectrometer (AB Sciex TripleTOF 5600, AB Sciex, United States) equipped with the high performance liquid chromatography (HPLC) column (2.1 × 150 mm, 3.5 μm) (Agilent Eclipse Plus C18, California, United States). The mobile phase A (positive ion: 0.1 formic acid and water; anion: water) and mobile phase B (positive ion: 0.1 formic acid and acetonitrile; anion: acetonitrile) were used for gradient elution (mobile phase A: 0–1.5 min: 80–80; 1.5~7 min: 80–5; 7–10 min: 5–80; 10–11 min: 80–80. Mobile phase B: 0–1.5 min: 20–20; 1.5–7 min: 20–95; 7–10 min: 95–20; 10–11 min: 20–20). The flow rate, sample load, and column temperature were 0.8 ml/min, 20 μl, and 45°C, respectively. The data acquisition software of the Analyst 1.5.1 software package was used to collect and organize the raw metabolic data, and we obtained the spectral peak index, sample name, and peak intensity area with retention time and precise mass-nucleus ratio. Meanwhile, the raw metabolic data matrix of 236 CRC patients was loaded into the MetaboAnalyst 4.0 online tool (https://www.metaboanalyst.ca/faces/home.xhtml) (Chong et al., 2018) to process with normalization, missing value processing, and scaling to acquire the processed metabolic data matrix including mass-charge ratio (m/z), retention time, and peak area.

Metabonomic Multivariate Statistical Analysis

To further distinguish metabonomic differences in CRC patients, based on the processed metabolic data matrix, we performed K-means unsupervised clustering algorithm to categorize 236 CRC into Kgroup 1 and Kgroup 2, and the procedures were repeated 1,000 times to determine the stabilization of grouping. The "survival R package" was utilized to perform Kaplan-Meier survival analysis for Kgroup 1 and Kgroup 2 to discern their prognostic differences. Principal component analysis (PCA) and partial least square discriminant analysis (PLSDA) were carried out with the MetaboAnalyst 4.0 online tool to ascertain the between-group variance and grouping rationalization. Then, we performed t test in the MetaboAnalyst 4.0 online tool for Kgroup 1 and Kgroup 2 to identify the differential metabolites, which were imported into The Human Metabolome Database (HMDB) (https://hmdb.ca/) to conduct structural identification and search for the differential lipid metabolites. Meanwhile, Kyoto Encyclopedia of Genes and Genomes (KEGG) functional pathway enrichment analysis for the structure-identified differential lipid metabolites was performed based on the MetaboAnalyst 4.0 online tool.

Multi-Lipid-Metabolite Prognostic Signature Analysis

Integrating the levels of differential lipid metabolites and survival data (survival status and survival time) of 236 CRC patients, we performed univariate Cox regression analysis with "survival R package" to acquire the survival-associated lipid metabolites ($p < 0.05$). Then, multivariate Cox regression analysis with "survival R package" and "survminer R package" was performed on survival-associated lipid metabolite matrix to obtain the multi-lipid metabolite prognostic signature, whose calculation method was listed along these lines: Lipid metabolite score (LMS) = $\Sigma_{x=1}^{n} Coef\,x * Level\,x$ (The x represented each metabolite for the establishment of the prognostic signature. Each coefficient, which was calculated by the "survival R package", was the fixed constant corresponding to each metabolite.). Univariate and multivariate independent prognostic analysis involved in clinical characteristics (age, gender, histological type, pathological stage, T staging, N staging, M staging) assessed the prognostic feasibility of the prognostic signature. Meanwhile, we plotted the composite nomogram mediated by independent prognostic factors with "rms R package" to methodically predict the CRC patients' overall survival. Moreover, clinical correlation analysis further revealed the correlation between the lipid metabolites of the prognostic signature and various clinical characteristics.

The Acquisition of Glycerophospholipid-Metabolism-Related Hub Genes by WGCNA

The raw CRC RNA expression matrices were obtained from the TCGA database and GEO datasets (GSE17536, GSE38832, and GSE103479). The TCGA Ensemble IDs of RNAs were transformed into the corresponding official symbol IDs retrieved from Ensemble database (https://asia.ensembl.org/index.html). The standard of RNA expression data was converted from FPKM to TPM. Meanwhile, the raw RNA probes of GSE17536, GSE38832, and GSE103479 were pinpointed their symbol IDs in the corresponding "GPL" file. The batch effect was eliminated with "sva R package" software. Differential expressed analysis was conducted for the TCGA gene expression matrix with filter criteria FDR < 0.05, |logFC| > 0.2. GSVA was used to determine the enrichment fractions of gene sets (Hänzelmann et al., 2013). WGCNA was utilized to depict a scale-free co-expression network, indicating that diverse genes with similar expression patterns might have the co-regulation functional pathway (Langfelder and Horvath, 2008). With reference to the above the results of the KEGG functional pathway enrichment analysis for the structure-identified

differential lipid metabolites, we concluded that glycerophospholipid metabolism was the pivotal metabolic pathway enriching lipid metabolites. We calculated the TCGA CRC patients' glycerophospholipid metabolism enrichment fraction of glycerophospholipid metabolism in the KEGG gene set (c2.cp.kegg.v7.4.symbols.gmt) retrieved from the MSigDB (http://www.gsea-msigdb.org/gsea/msigdb) with GSVA. Differential expressed RNA matrix was conducted with WGCNA to construct a co-expression network with the "WGCNA R package" to identify the GMRHGs. In order to enable the constructed co-expression network to be more consistent with the characteristics of the scale-free network, R^2 = 0.9 was set to determine the soft threshold. Meanwhile, the adjacency matrix was transformed into the topological overlap matrix (TOM) to reduce noise and spurious correlation. Based on the dynamic tree, we distinguished the dynamic shearing module with the minimum number of modular genes set to 30 and the shearing height set to 0.25. The relationships between the module eigengene (ME) gene sets and the clinical traits (age, gender and pathological stage) as well as glycerophospholipid metabolism were evaluated with gene significance (GS) and module membership (MM).

Multi-GMRHG Prognostic Signature

CRC patients retrieved from the TCGA and GEO (GSE17536, GSE103479 and GSE38832) were respectively categorized into training and test groups. To minimize the impacts of emergencies (such as acute cerebrovascular or cardiovascular disease, massive hemorrhage, and pulmonary embolism, and so forth) on predicting the survival outcome of CRC patients, we excluded the CRC patients with overall survival less than 30 days. Univariate Cox regression analysis and LASSO regression analysis were performed on the training group (TCGA CRC patients) to obtain prognosis-associated GMRHG. After the optimization of LASSO regression analysis with the "glmnet R package" software, we performed multivariate Cox regression analysis on the prognosis-associated GMRHG expression matrix to construct the multi-GMRHG prognostic signature. In the prognostic signature, the glycerophospholipid metabolism score (GMS) was calculated as follows: GMS = $\Sigma_{i=1}^{n} Coefi * Expi$ ("Expi" symbolized the expression of each GMRHG in the prognostic. "Coefi" denoted the constant corresponding to each GMRHG, and the constant derived from the calculation with "survival R package"). The test group gene expression matrix was substituted into the above formula to obtain the GMS of each sample, which was used to verify the prognostic significance of the prognostic signature.

Tumor Immune Infiltration and TME Analysis

Tumor immune analysis was utilized to explore relationship between the tumor immune infiltration and the prognostic signature. Through the ssGSEA for the gene expression matrix of TCGA patients in the high- and low-risk groups, single sample gene set enrichment analysis (ssGSEA) (Barbie et al., 2009), a single-sample function enrichment analysis for a particular gene set, defined an enrichment score indicating the absolute enrichment of the particular gene set in each sample. In this study, we conducted ssGSEA with the "GSVA R package" to calculate the relative abundances of tumor immunocytes and immune functions of each TCGA CRC sample (Angelova et al., 2015). Moreover, we performed the ESTIMATE algorithm to calculate the Immune score and Stromal score to quantify the immune and stromal components in tumor microenvironment (TME), and the sum of both was ESTIMATE score, which was utilized to evaluate the tumor infiltrations in TME. Tumor Immune Dysfunction and Exclusion (TIDE) algorithm was utilized to predict the anti-PD-1 and anti-CTLA4 immunotherapeutic response of CRC patients (Jiang et al., 2018). With the TIDE score escalating, the probability of immune escape escalated and the immunotherapeutic response worsened.

Somatic Mutation Landscape in the Prognostic Signature

The somatic mutation data retrieved from the TCGA "Masked Somatic Mutation" data type was summarized with the "maftools R package" to depict the genetic mutational landscape of CRC patients and of the characteristic genes of mismatch repair (MMR) system, namely, MLH1, MSH2, MSH6, and PMS2.

Gene Set Enrichment Analysis for the Prognostic Signature

GSEA, the connectivity analysis performed on the specific functional phenotypes for the whole gene chip expression profiling (Subramanian et al., 2005), was utilized to evaluate the enrichment of the gene expression matrices of high- and low-risk groups in the KEGG gene sets and hallmark gene set (NOM $p < 0.05$, FDR $q < 0.25$), thereby hierarchically identifying the specific biological function of CRC patients in high- and low-risk groups.

RESULTS

Metabonomics Pattern Recognition Analysis

Figure 1 shows the research idea of the study. **Table 1** shows the clinical statistics of CRC patients. Totally, we included 236 CRC patients with their clinical characteristics, including age, gender, histological type, pathological stage, T staging, N staging, and M staging. Moreover, we performed PCA and PLSDA to investigate the spatial distribution of metabolites in the subgroups (Kgroup 1 and Kgroup 2) completed with K-means unsupervised clustering algorithm (**Figure 2A** and **Supplementary Table S1**). K-M survival analysis (**Figure 2B**) indicated that the overall survival of CRC patients in Kgroup 2 was higher than that in Kgroup 1 ($p < 0.001$). **Figures 2C,D** show the results of PCA for Kgroup 1 and Kgroup 2, indicating that there were significant individual differences between two subgroups. Furthermore, in order to eliminate random deviations and deviations within the group, we

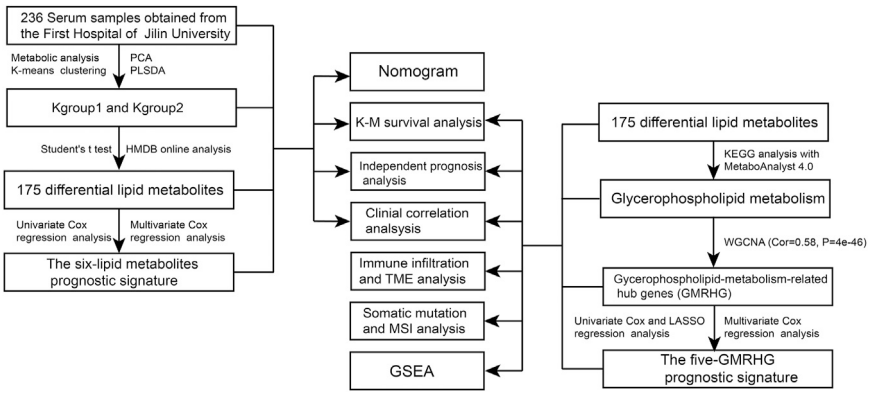


FIGURE 1 | The flow chart of the study.

TABLE 1 | The overview of JLUFH CRC patients' clinical characteristics.

Characteristics	Variates	Amounts (Percentage%)
Age	<=65	168 (71.19)
	>65	68 (28.81)
Gender	Female	106 (44.92)
	Male	130 (55.08)
Histological type	Adenocarcinoma	191 (80.93)
	Mucinous adenocarcinoma	45 (19.07)
T staging	T1-2	35 (14.83)
	T3-4	201 (85.17)
N staging	N0	124 (52.54)
	N1-2	112 (47.46)
M staging	M0	225 (95.34)
	M1	11 (4.66)
Pathological stage	Stage I–II	124 (52.54)
	Stage III–IV	112 (47.46)

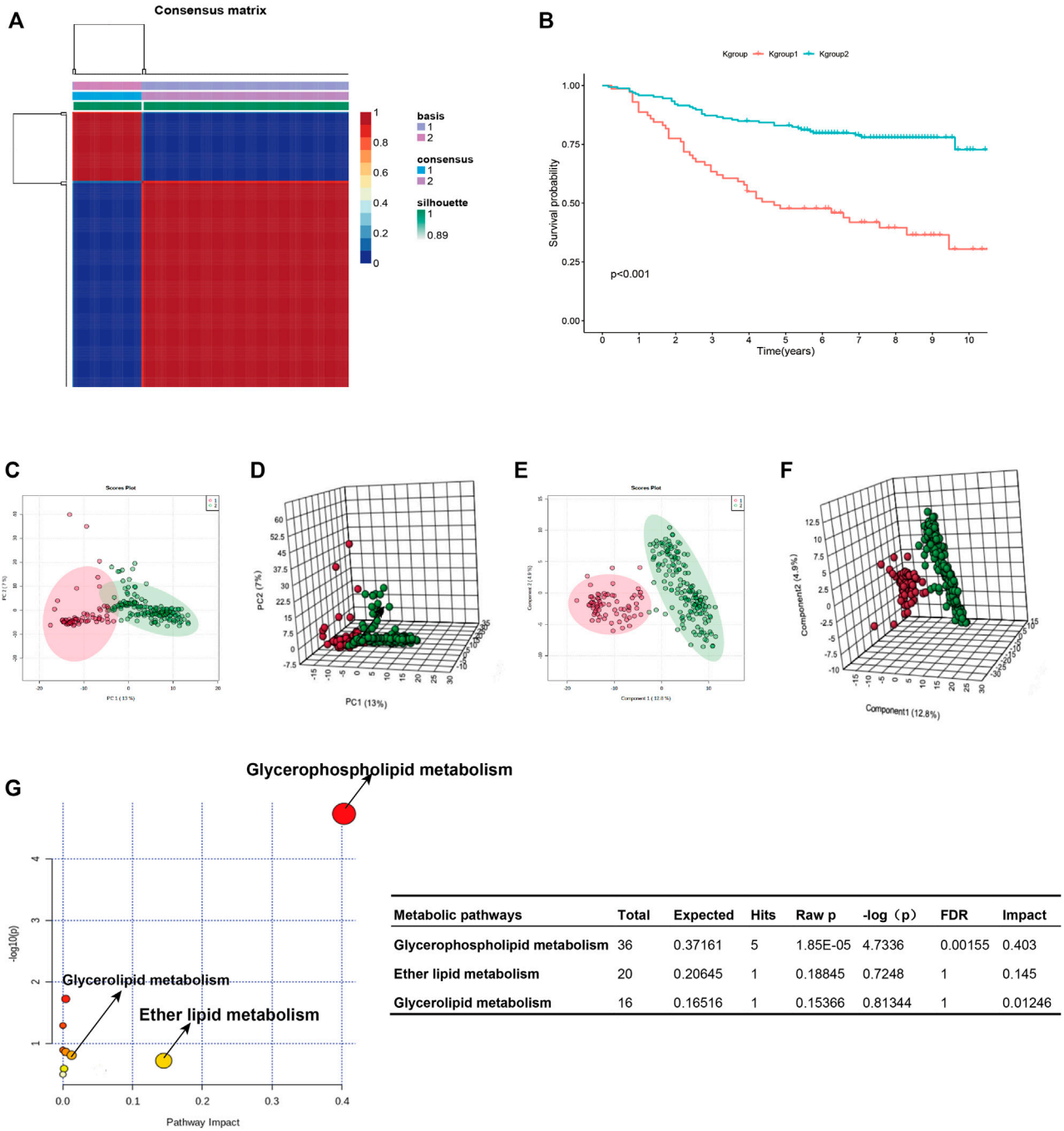
performed PLSDA for Kgroup 1 and Kgroup 2 to further validate the individual differences between two subgroups (**Figures 2E,F**).

The Biological Functions of Differential Metabolites

T test was employed to distinguish the differential 357 metabolites (FDR < 0.05), which were then conducted identifications of specific components on HMDB (https://hmdb.ca/). We confirmed 175 differential lipid metabolites (**Supplementary Table S2**), notably, most of which were fatty acid and glycophospholipin such as glycolipid products. KEGG pathway enrichment analysis on the MetaboAnalyst 4.0 online tool for the 175 differential lipid metabolites revealed that the differential lipid metabolites uppermost enriched in the glycerophospholipid metabolism (FDR = 0.0016, Impact = 0.403) (**Figure 2G**).

The Construction of the Six-Lipid Metabolite Prognostic Signature

Univariate Cox regression analysis conducted with the "survival R package" was utilized to investigate the survival-associated lipid metabolites ($p < 0.05$) and 14 survival-associated lipid metabolites were confirmed (**Table 2**). Then, we performed multivariate Cox regression analysis for the 14 survival-associated lipid metabolite to establish the prognostic signature based on six lipid metabolites (**Table 3**) including five high-risk lipid metabolites (HR > 1) (Cer(d18:0/14:0), Ganglioside GT3(d18:0/18:1(9Z), LysoPE(22:6(4Z,7Z,10Z,13Z,16Z, 19Z)/0:0), PA(20:3(5Z,8Z, 11Z)/24:1(15Z)), PS (20:4(5Z,8Z,11Z, 14Z)/14:1(9Z), and one low-risk lipid metabolites (HR < 1) (Substance P). LMS = $0.233^{*}\ \mathrm{Level}_{\mathrm{Cer(d18:0/14:0)}} + 0.303\ ^{*}\ \mathrm{Level}_{\mathrm{LysoPE[22:6(4Z,\ 7Z,\ 10Z,\ 13Z,16Z,\ 19Z)/0:0]}} + 0.298\ ^{*}\ \mathrm{Level}_{\mathrm{PS\ (20:4(5Z,\ 8Z,\ 11Z,\ 14Z)/14:1(9Z))}} + 0.968\ ^{*}\ \mathrm{Level}_{\mathrm{PA\ [20:3(5Z,\ 8Z,\ 11Z)/24:1(15Z)]}} + 1.079\ ^{*}\ \mathrm{Level}_{\mathrm{Ganglioside\ GT3[d18:0/18:1(9Z)]}} + (-0.179)\ ^{*}\ \mathrm{Level}_{\mathrm{Substance\ P}}$. Patients with an LMS greater than or equal to the median LMS (0.875) were categorized into the high-risk group, and those less than the median LMS (0.875) were classified into the low-risk group. The K-M survival curve (**Figure 3A**) for CRC patients in the high- and low-risk groups illustrated that the overall survival of the CRC patients in the low-risk group was higher than that in the high-risk group with significant statistical difference ($p = 0.000$). As shown in **Figure 3B**, the scatter plots showed that the overall survival of CRC patients with higher LMS was lower than that of those with lower LMS. The LMS curve (**Figure 3C**) showed the distributions of CRC patients with the LMS increasing. The heatmap (**Figure 3D**) showed the distributions and variations of six metabolite levels based on the prognostic signature. With the increase of LMS, the levels of five high-risk metabolites increased, and the level of the low-risk metabolites showed a decreasing trend. Furthermore, the AUCs of the ROC curves (**Figure 3E**) to predict the patients' one-, three-, and five-year overall survival was 0.769, 0.711, and 0.723, respectively. Combined with the LMS and clinical characteristics (age, gender, pathological stage, histological type, T staging, N staging, and M staging), univariate and multivariate independent prognostic analyses (**Figures 3F,G**) were conducted to examine the prognostic feasibility of the prognostic signature, and the LMS could be an independent prognostic factor to predict the patients' prognosis. With the integration of the prognostic signature and independent prognostic factors (histological type (adenocarcinoma and



Metabolic pathways	Total	Expected	Hits	Raw p	-log（p）	FDR	Impact
Glycerophospholipid metabolism	36	0.37161	5	1.85E-05	4.7336	0.00155	0.403
Ether lipid metabolism	20	0.20645	1	0.18845	0.7248	1	0.145
Glycerolipid metabolism	16	0.16516	1	0.15366	0.81344	1	0.01246

FIGURE 2 | K-group1 and K-group2 are independent with each other. **(A)**: K-means clustering heatmap of K-group1 and K-group2. **(B)**: The K-M survival curves indicated that the overall survival of JLUFH CRC patients in K-group 1 was lower than that in K-group2. **(C,D)**: Two-dimensional and three-dimensional PCA plots for K-group1 and K-group2. **(E,F)**: Two-dimensional and three-dimensional PLSDA plots for K-group1 and K-group2. **(G)**: The KEGG functional enrichment analysis for differential lipid metabolites.

mucinous adenocarcinoma), pathological stage, T staging), we constructed the composite nomogram (**Figure 3H**) to individually predict the CRC patients' overall survival. The sum of the corresponding scores of variates in the composite nomogram identified the less than one-, three-, and five-year overall survival probability. The one-, three-, and five-year calibration curves also verified the promising fitting and stability of the composite nomogram (**Figure 3I**). The one-, three-, and five-year AUCs of ROCs based on the nomogram was 0.815, 0.815, and 0.805, respectively, showing favorable accuracy in predicting patients' prognosis (**Figure 3J**). The one-, three-, and five-year AUCs of multi-index ROC curves

TABLE 2 | Univariate Cox regression analysis for the differential lipid metabolites.

Name	HMDB	HR	HR.95L	HR.95H	*p*-value
PA[16:0/14:1(9Z)]	HMDB0114834	1.341	1.075	1.673	0.019
Substance P	HMDB0001897	0.699	0.554	0.883	0.006
Tridecanoylglycine	HMDB0013317	1.251	1.046	1.497	0.03
PA[14:1(9Z)/22:6(4Z, 7Z, 10Z, 13Z, 16Z, 19Z)]	HMDB0114809	0.75	0.591	0.951	0.036
PGP(a-13:0/a-13:0)	HMDB0116512	1.352	1.116	1.637	0.005
PS(20:4(5Z, 8Z, 11Z, 4Z)/14:1(9Z))	HMDB0012430	1.589	1.332	1.894	0
PA(20:3(5Z, 8Z, 11Z)/24:1(15Z))	HMDB0115143	2.987	1.437	6.209	0.008
PS[18:0/20:4(8Z, 11Z, 14Z, 17Z)]	HMDB0010165	1.194	1.048	1.361	0.016
PS[DiMe(9,3)/MonoMe(11,3)]	HMDB0061584	0.726	0.579	0.909	0.012
Ganglioside GD2(d18:0/16:0)	HMDB0011839	0.744	0.592	0.935	0.023
Ganglioside GT3(d18:0/18:1(9Z))	HMDB0012057	1.309	1.144	1.497	0
Cer(d18:0/14:0)	HMDB0011759	1.32	1.087	1.603	0.011
LysoPC(18:0)	HMDB0010384	1.166	1.03	1.321	0.032
LysoPE[22:6(4Z,7Z,10Z,13Z,16Z,19Z)/0:0]	HMDB0011526	1.168	1.051	1.296	0.008

TABLE 3 | Multivariate Cox regression analysis for the survival-associated lipid metabolites.

Name	Coef	HR	HR.95L	HR.95H	*p* value
Cer(d18:0/14:0)	0.23250734	1.2617597	1.027315	1.5497072	0.026629
LysoPE[22:6(4Z,7Z,10Z,13Z,16Z,19Z)/0:0]	0.30315863	1.3541293	1.084041	1.6915098	0.0075644
PS[20:4(5Z,8Z,11Z,14Z)/14:1(9Z)]	0.2984225	1.3477311	1.0988127	1.653038	0.0041774
PA[20:3(5Z,8Z,11Z)/24:1(15Z)]	0.96788033	2.6323588	1.2307888	5.6299772	0.0125841
Ganglioside GT3[d18:0/18:1(9Z)]	1.07881366	2.9411882	1.7988826	4.8088675	1.70E-05
Substance P	−0.1788966	0.8361924	0.6503613	1.0751218	0.1629861

(**Figures 3K–M**) verified the enhanced prognostic accuracy of the composite nomogram.

Identification of GMRHGs With WGCNA

In order to ensure the tumor-specific gene selection, we used differential expressed gene matrix of TCGA CRC patients for WGCNA. Integration of 4432 DEG expression data (**Supplementary Figure S1**) and age, gender, and pathological stage was used for the input files of WGCNA. Hierarchical clustering (**Figure 4A**) and Pearson correlation analysis were carried out to construct WGCNA network. Soft threshold $\beta = 13$ (**Figure 4B**) assessed the optimal power value adjacency matrix transforming to TOM. Through the identification of dynamic tree cutting modules and the merge of similar modules (**Figure 4C**), 8 gene modules (MEbrown, MEturquoise, MEmidnightblue, MEyellow, MEmagenta, MEgreenyellow, MEpink, MEgrey) were identified (**Figure 4D**), among which the MEbrown module containing 342 DEGs was significantly correlated with glycerophospholipid metabolism (Cor = 0.58, p = 4e-46) (**Supplementary Table S3**). A total of 342 DEGs of MEbrown module were deemed as the GMRHG for further analysis.

The Construction of the Five-GMRHG Prognostic Signature

Supplementary Table S4 displayed the clinical statistics of 877 CRC patients retrieved from the TCGA and GEO datasets (GSE17536, GSE38832, and GSE103479). Univariate Cox regression analysis (**Figure 5A**), LASSO regression analysis (**Figures 5B,C**), and multivariate Cox regression analysis (**Figure 5D**) were performed on 342 GMRHGs expression data to construct the five-GMRHG (ACOX1, ATOH1, CPT2, PCSK5, and TINCR) prognostic signature. GMS = (−0.460) * Exp_{ACOX1} +(−0.184) * Exp_{ATOH1} + (−0.362) * Exp_{CPT2} + 0.248 * Exp_{PCSK5} + 0.463 * Exp_{TINCR}. CRC patients with a GMS greater than or equal to the median GMS (0.939) would be assigned into the high-risk group, and those less than the median GMS would be assigned into the low-risk group. K-M survival curve (**Figure 6A**) for the training and test group (**Figure 6E**) illustrated that there were significant differences of the overall survival between low-risk group and high-risk group (p = 4.779e−05 and p = 4.305e−05). Scatter plots for the training (**Figure 6B**) and test group (**Figure 6F**) indicated that the overall survival of patients with higher GMS was lower than that of patients with lower GMS. GMS curves for the training (**Figure 6C**) and test group (**Figure 6G**) showed the distributions of GMS, with the GMS increasing. The heatmaps for the training (**Figure 6D**) and test group (**Figure 6H**) showed the expression trends of five GMRHGs of the prognostic signature. With the increase of GMRHG, the expression of two high-risk GMRHGs (PCSK5 and TINCR) would be increased, while the expression of three low-risk GMRHGs (ATOH1 and ACOX1 and CPT2) would be decreased. The one-, three-, and five-year AUCs of time-dependent ROCs in the training (**Figure 6I**) and test group (**Figure 6J**) were 0.662, 0.716, 0.719 and 0.623, 0.660, 0.633, respectively. As shown in **Figures 6K, L**, the results of univariate and multivariate independent prognostic analysis indicated that GMS could be

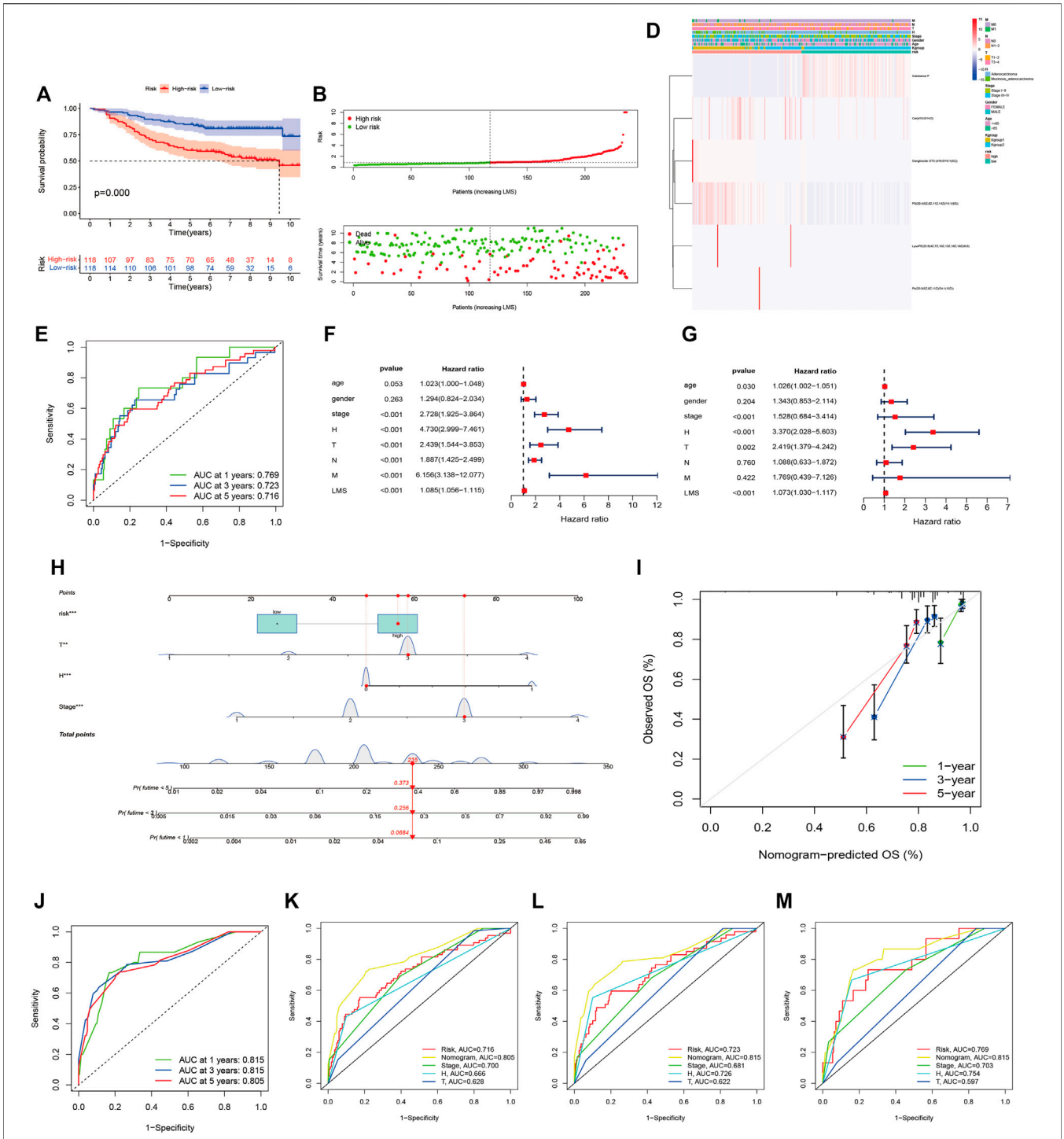


FIGURE 3 | The six-lipid-metabolite prognostic signature is the robust independent prognostic factor. **(A)**: The K-M survival curves based on the six-lipid metabolite prognostic signature. **(B)**: The scatter plot showing the correlation between the JLUFH CRC patients' survival status and LMS. **(C)**: The dot plot representing the distributions of JLUFH CRC patients' LMS in the high- and low-risk groups. **(D)**: The clustering heatmap exhibiting the level variation tendencies of the six lipid metabolites of the prognostic signature in various clinical characteristics, K-groups (K-group1 and K-group2) and risk grouping (High- and low-risk groups). **(E)**: The one-, three-, and five-year ROC curves of the six-lipid-metabolite prognostic signature. **(F,G)**: Univariate and multivariate independent prognostic analysis based on the six-lipid-metabolite prognostic signature, demonstrating that pathological stage, histological type and LMS could be deemed as the independent prognostic factors. **(H)**: The composite nomogram consisted of independent prognostic factors. **(I)**: The one-, three-, and five-year calibration curves of the composite nomogram. **(J):** The one-, three-, and five-year ROC curves of the composite nomogram. **(K–M)**: The one-, three-, and five-year multi-index ROC curves of the composite nomogram.

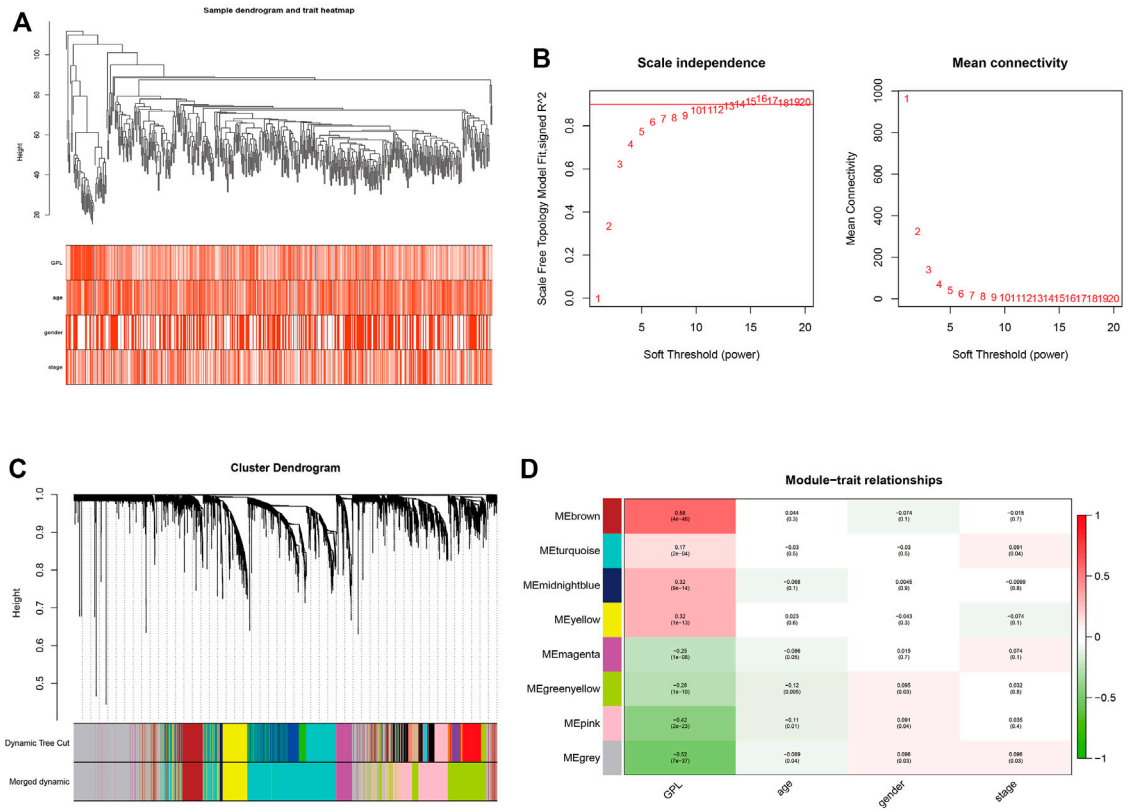


FIGURE 4 | The identification of GMRHGs with WGCNA. **(A)**: Hierarchical clustering dendrogram of TCGA CRC patients. **(B)**: The selection of soft threshold. **(C)**: The dynamic tree cut dendrogram for TCGA CRC patients. **(D)**: The correlation between gene module and clinical traits.

deemed as independent prognostic factors. The time-dependent ROC verified the preferable precision of the five-GMRHG prognostic signature to predict CRC patients' prognosis.

Clinical Correlation Analysis

We used the Chi-square test to compare the distributions of clinical characteristics (age, gender, histological type, pathological stage, T staging, N staging, M staging) based on JLUFH CRC patients and TCGA CRC patients (age, gender, pathological stage, T staging, N staging, M staging) in the high- and low-risk groups. For the JLUFH CRC patients, we found that CRC patients with advanced pathological stage (stage III and IV) and mucinous adenocarcinomas were mainly distributed in the high-risk group with significant statistical difference (p = 0.031 and p = 0.001) (**Figures 7A,B**). For the TCGA CRC patients, we found that TCGA CRC patients with advanced pathological stage (stage III and IV) and N staging (N1-2) were mostly distributed in the high-risk group (**Figures 7C,D**). Clinical correlation analysis with Wilcoxon ranking test was utilized to further explore the intra-group differences of six-lipid-metabolites level and five-GMRHG expression in clinical characteristics based on the JLUFH CRC patients and the TCGA CRC patients. "*", "**", and "***" respectively represented statistical difference, highly statistical difference, and significant statistical difference. It was found that the level of high-risk lipid metabolite, PS [20:4 (5Z, 8Z, 11Z, 14Z)/14:1 (9Z)], was higher in the JLUFH CRC patients in the advanced-pathological-stage group (**Figure 7E**), and the levels of five high-risk lipid metabolites (Cer(d18:0/14:0), Ganglioside GT3(d18:0/18:1(9Z), LysoPE[22: 6 (4Z, 7Z, 10Z, 13Z, 16Z, 19Z)/0: 0], PA[20:3 (5Z,8Z, 11Z)/24:1 (15Z)], PS[20:4 (5Z, 8Z, 11Z, 14Z)/14:1 (9Z)] in the mucinous-adenocarcinoma group were significantly higher than those in the adenocarcinoma group (**Figure 7F**). The expression of CPT2 was higher in the TCGA CRC patients no more than 65 years old compared with that in CRC patients greater than 65 years old (**Figure 7G**). Furthermore, the expressions of ATOH1 and CPT2 presented statistical differences in terms of pathological stage. The expression of ATOH1 and CPT2 was higher in stage I and II than that in stage III and IV (**Figure 7H**). In T staging, the expression of PCSK5 in T3-4 was higher than that in T1-2 (**Figure 7I**). Moreover, in N staging, with the increase of N staging, the expression of ATOH1 and CPT2 displayed downward tendencies (**Figure 7J**). For M staging, the expression of ATOH1 was higher in M0 than in M1 with statical differences (**Figure 7K**).

Characteristics of Tumor Immune Infiltration and TME Based on the Prognostic Signature

We calculated the relative abundances of 16 kinds of immunocytes and 13 kinds of immune functions with the

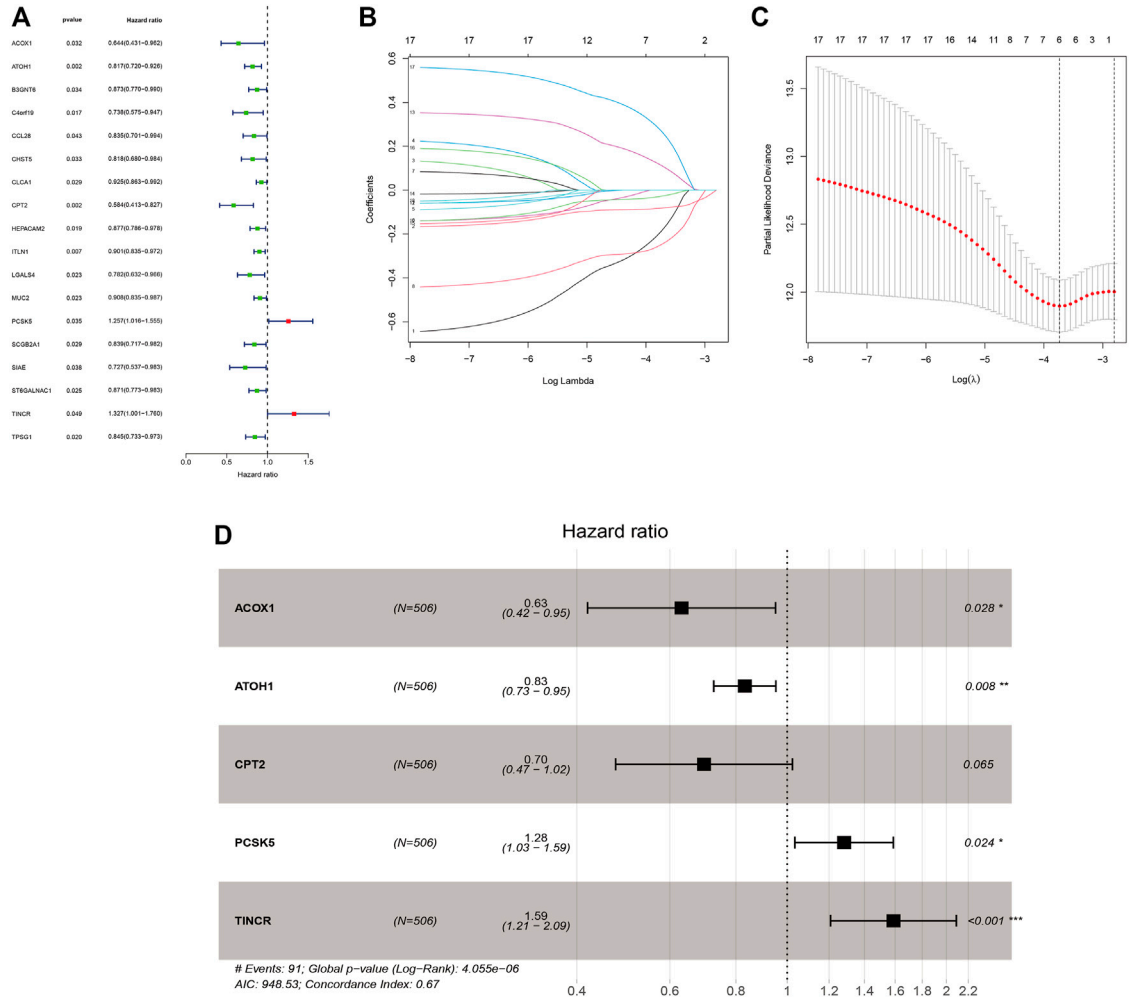


FIGURE 5 | The construction of the five-GMRHG prognostic signature. **(A)**: 18 prognosis-associated GMRHGs screened by univariate Cox regression analysis. **(B,C)**: The LASSO regression analysis for the prognosis-associated GMRHGs. **(D)**: Five prognosis-associated GMRHGs screened by multivariate Cox regression analysis.

"GSVA R package". The box plot (**Figures 8A,B**) iconically illustrated that the relative abundances of T_helper_cells, Tfh, Th1_cells, TIL, CCR, check-point, HLA, inflammation-promoting, T_cell_co-inhibition, T_cell_co-stimulation, Type_I_IFN_Response, and Type_II_IFN_Response in the high-risk group were higher than that in the low-risk group with statistical differences ($p < 0.05$). Meanwhile, we applied Spearman correlation analysis to further validate the linear relationship between the prognostic signature and tumor immune infiltration. As shown in **Supplementary Figures S2A–2L**, the GMS was respectively positively correlated with the relative abundances of T_helper_cells, Tfh, Th1_cells, TIL, CCR, check-point, HLA, inflammation-promoting, T_cell_co-inhibition, T_cell_co-stimulation, Type_I_IFN_Response, and Type_II_IFN_Response. In this study, we further uncovered the underlying correlations between the TME and tumor immune infiltrations. **Figure 8C** shows that Immune score, Stromal score, and ESTIMATE score were higher in the high-risk group than that in the low-risk group with statistical differences ($p < 0.05$). **Supplementary Figures S2M–2O** demonstrated that GMS was respectively linearly positively correlated with Stromal score, Immune score, and ESTIMATE score. The heatmap (**Figure 8D**) systematically depicted the distributional relationships between the Stromal score, Immune score, ESTIMATE score, and the relative abundances of tumor immunocytes and immune functions in the high- and low-risk groups. The distributions of the Stromal score, Immune score, and ESTIMATE score were collaborative with the distributions of the relative abundances of T_helper_cells, Tfh, Th1_cells, TIL, CCR, check-point, HLA, inflammation-promoting, T_cell_co-inhibition, T_cell_co-stimulation, Type_I_IFN_Response, and Type_II_IFN_Response in the high- and low-risk groups. Moreover, as shown in the **Figure 8E**, we found that the Stromal score, Immune score, and ESTIMATE score were all statistically positively correlated with the relative abundances of T_helper_cells, Tfh, Th1_cells,

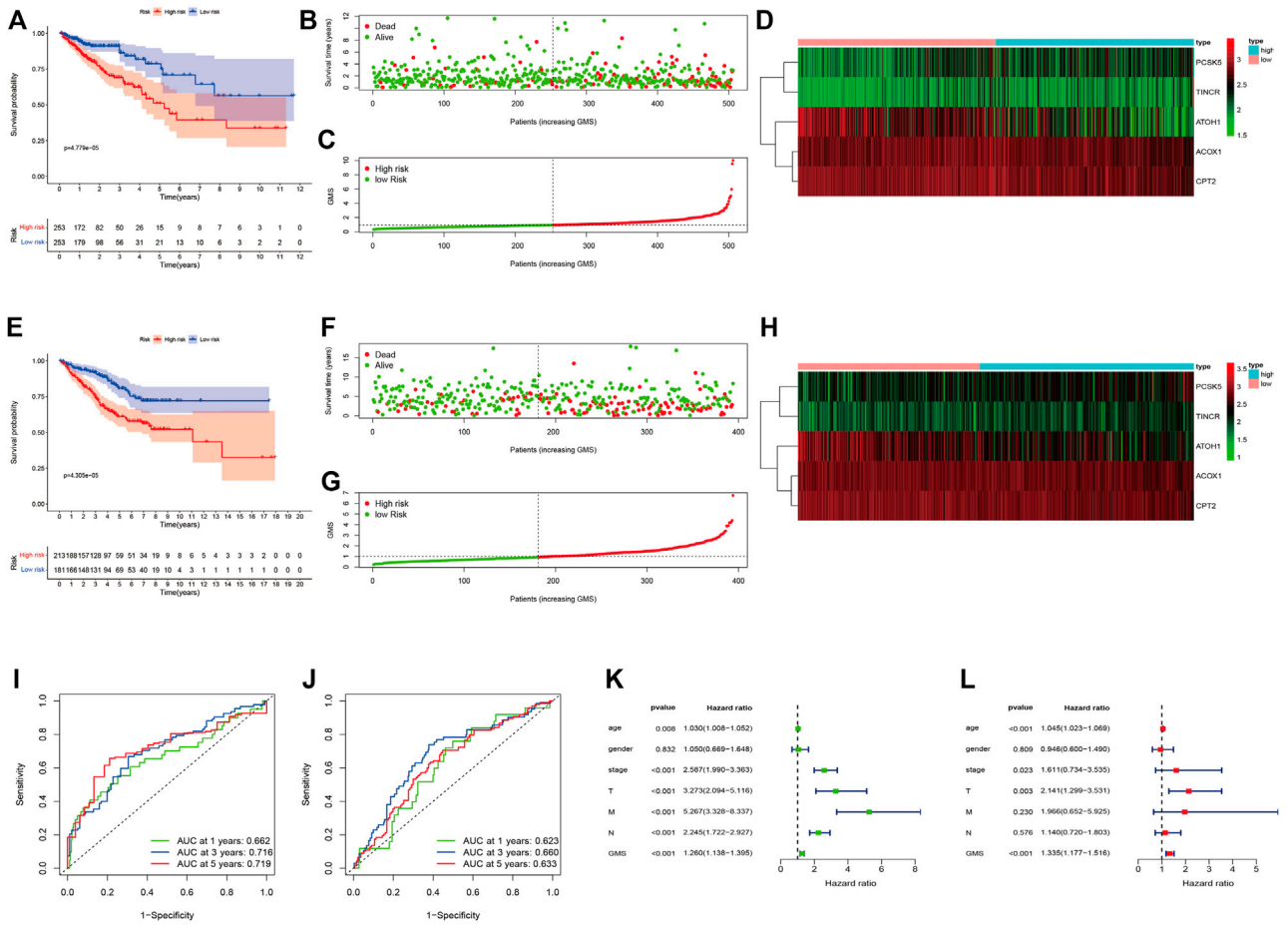


FIGURE 6 | CRC patients stratified according to the five-GMRHG prognostic signature. **(A–D)**: The survival plots for the TCGA CRC patients respectively illustrating the overall survival difference, the distribution of GMS, the survival status of CRC patients, and the expression variations of five GMRHGs in the high- and low-risk groups. **(E)–(H)**: The survival plots for the GEO CRC patients (GSE17536, GSE38832 and GSE103479) respectively illustrating the overall survival difference, the distribution of GMS, the survival status of CRC patients, and the expression variation of five GMRHGs in the high- and low-risk groups. **(I,J)**: The one-, three-, and five-year ROC curves of the five-GMRHG prognostic signature based on the training and test group. **(K,L)**: The univariate and multivariate independent prognostic analysis.

TIL, CCR, check-point, HLA, inflammation-promoting, T_cell_co-inhibition, T_cell_co-stimulation, Type_I_IFN_Response, and Type_II_IFN_Response in the high- and low-risk groups. The results of the comparisons between the expression of immune checkpoints in the high- and low-risk groups indicated that the expression of CTLA4, PDCD1, TIGIT, CD274, and HAVCR2 was higher in the high-risk group than in the low-risk group (**Figure 8F**). Spearman correlation analysis (**Supplementary Figures S2P–2T**) also validated the positive linear correlations between the expression of immune checkpoints and GMS. Furthermore, immunotherapeutic response analysis conducted with the TIDE algorithm was performed on the CRC patients in the high- and low-risk groups. As shown in the **Figure 8G**, the TIDE score in the low-risk group was lower than that in the high-risk group, indicating that samples in the low-risk group were more susceptive to anti-PD1 and anti-CTLA4 immunotherapy compared with samples in the high-risk group. CRC patients with lower TIDE scores were associated with favorable overall survival compared with CRC patients with higher TIDE scores ($p = 0.017$) (**Figure 8H**). Moreover, the results of stratified survival analysis for CRC patients in the high- and low-risk groups showed that CRC patients in the low-risk group with lower TIDE scores were related with better performance of overall survival (**Figure 8I**).

Somatic Mutation and MSI Overview

Figures 9A,B display the top 30 genes with highest mutation frequencies, among which TP53 showed the most significant statistical difference between the high- and low-risk groups (66 vs. 49%, $p < 0.001$) (**Figure 9C**). In order to further investigate the prognostic significance of the five-GMRHG prognostic signature in the TP53 mutational status, we conducted a hierarchical survival analysis for TP53-wild and TP53-mutation CRC patients, and found that the overall survival of TP53-mutation CRC patients in the high-risk group was lower than that in the low-risk group with a significant statistical difference ($p = 0.001$). Meanwhile, the overall survival of TP53-wild CRC patients exhibited the similar

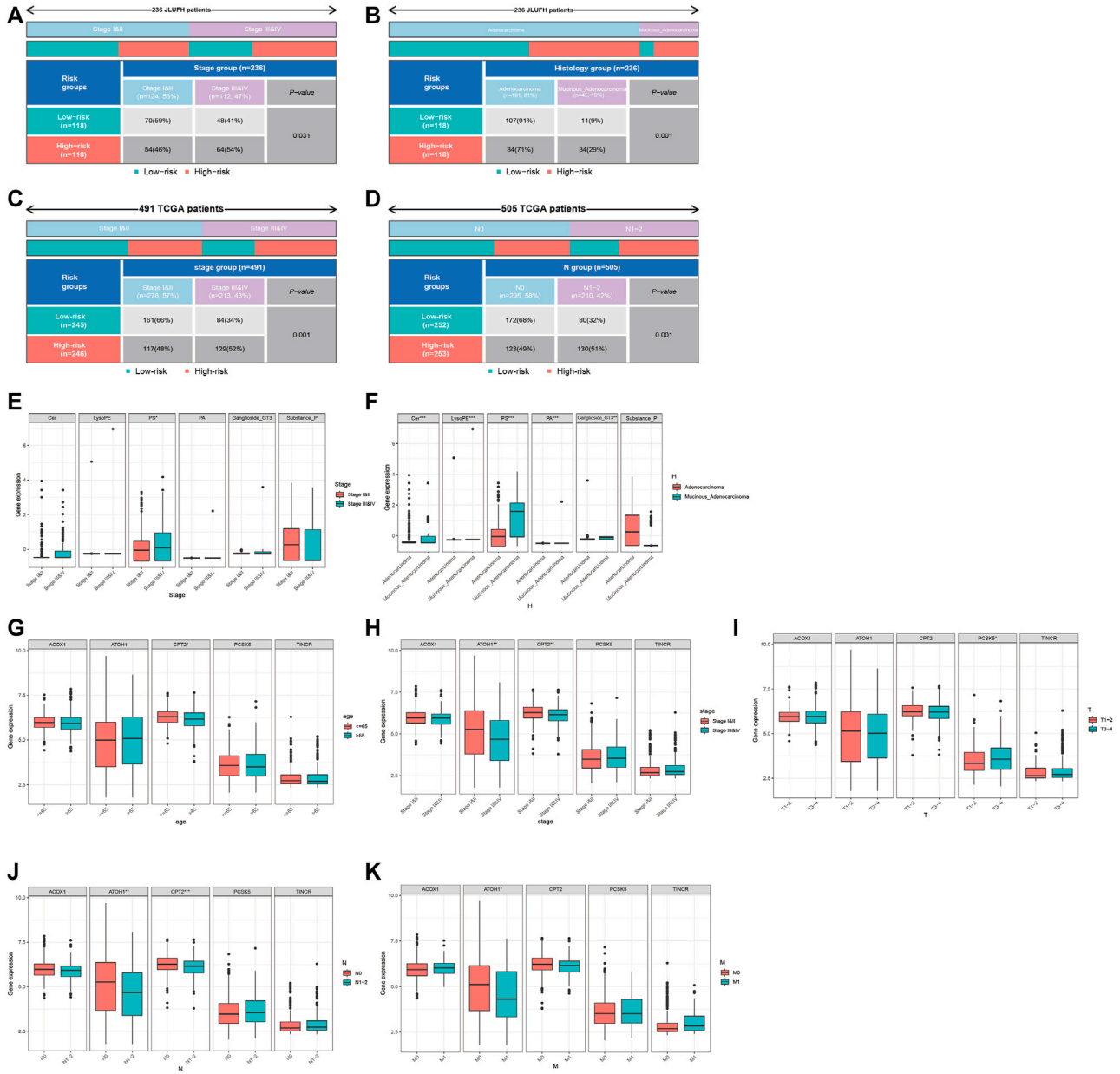


FIGURE 7 | Clinical correlation analysis for the JLUFH and TCGA CRC patients. **(A,B)**: The distributions of pathological stage and histological type of JLUFH CRC patients in the high- and low-risk groups. **(C,D)**: The distributions of pathological stage and N staging of JLUFH CRC patients in the high- and low-risk groups. **(E,F)**: The differences of the levels of six lipid metabolites in the JLUFH CRC patients with stage I and II, stage III and IV and adenocarcinoma, and mucinous adenocarcinoma. **(G–K)**: The differences of the levels of five GMRHGs in the TCGA CRC patients with various clinical characteristics (age, pathological stage, T staging, N staging, and M staging).

tendency to that of the overall survival of TP53-mutation CRC patients in the high- and low-risk groups (**Figure 9D**). Furthermore, we investigated the relationship between the MMR system and five-GMRHG prognostic signature. As shown in **Figures 9E,F**, MSH6 and MSH2 respectively demonstrated the highest mutation frequency in the high- and low-risk groups. Additionally, Pearson correlation analysis for the four characteristic genes of MMR system illustrated the underlying co-expression correlation among them, where MSH6-MSH2 displayed most fervent positive co-expression correlation (**Figure 9G**). The results of Wilcoxon ranking test for the expression differences of MSH6, MSH2, PMS2, and MLH1 indicated that the expression of MLH1 in the low-risk group was higher than that in the high-risk group (**Figure 9H**). The linear correlation conducted with Spearman correlation analysis also verified the negative relationship between the GMS and MLH1 expression ($R = -0.16$, $p = 2e\text{-}04$) (**Figure 9I**).

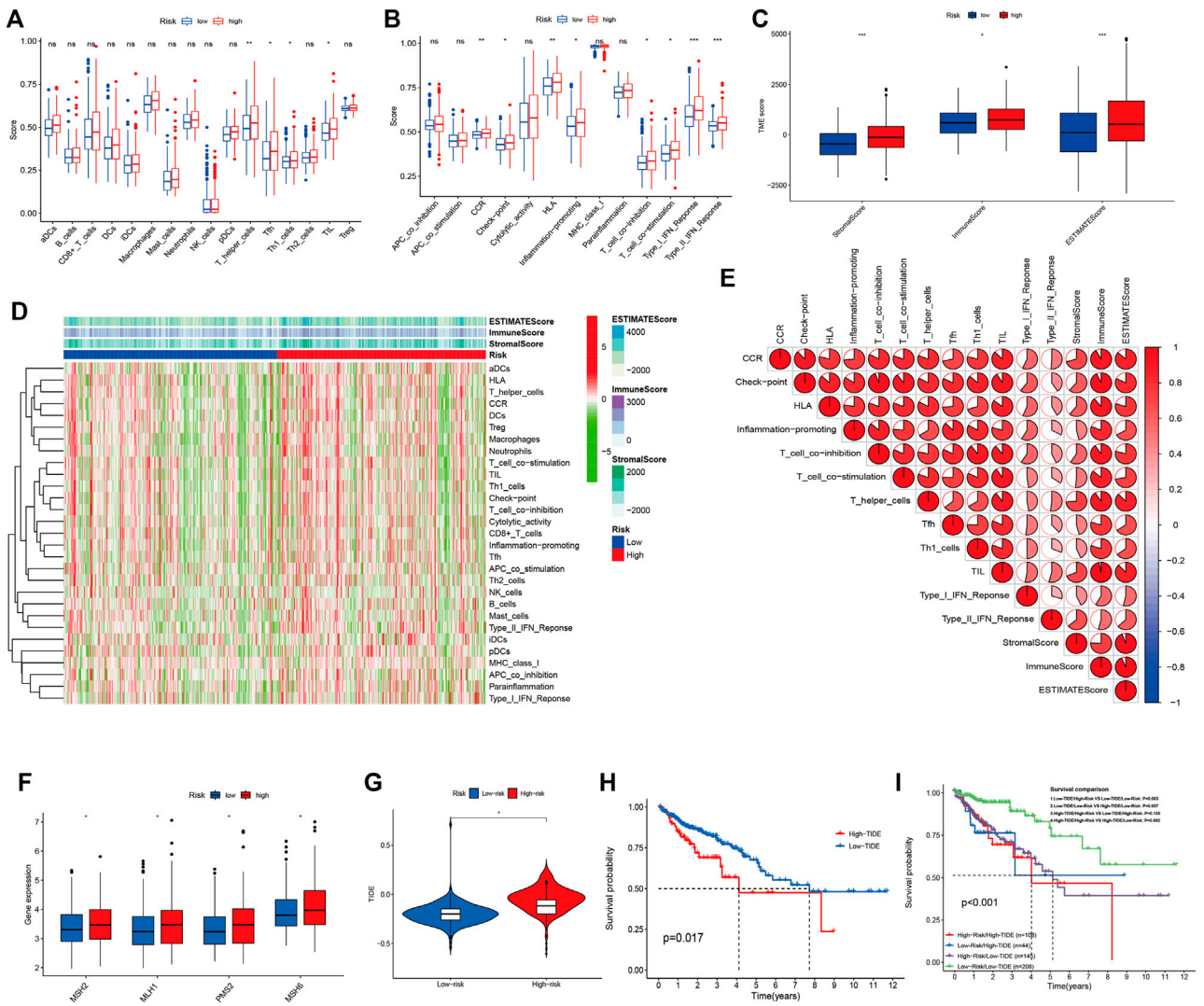


FIGURE 8 | The correlation between the immune infiltrations and TME. **(A,B)**: The differences of relative abundances of immune components in the high- and low-risk groups. **(C)**: The differences of Immune score, Stromal score and ESTIMATE score in the high- and low-risk groups. **(D)**: The distributional correlation between the immune components and TME scores. The red and green strips respectively represent the up-regulated and downregulated immune components. **(E)**: The co-expression relationship among the differential immune components (T_helper_cells, Tfh, Th1_cells, TIL, CCR, check-point, HLA, inflammation-promoting, T_cell_co-inhibition, T_cell_co-stimulation, type_I_IFN_Response, and type_II_IFN_Response) and TME scores. **(F)**: The expression differences of five immune checkpoints (CTLA4, PDCD1, TIGIT, CD274, and HAVCR2) in the high- and low-risk groups. **(G)**: The difference of TIDE score in the high- and low-risk groups. **(H)**: The K-M survival curves indicating the overall survival difference between the TCGA CRC patients with high-TIDE score and low-TIDE score. **(I)**: The K-M survival curves stratified by the five-GMRHG prognostic signature and TIDE score. The overall survival of TCGA CRC patients with Low-TIDE/High-Risk was lower than that with Low-TIDE/Low-Risk with statistical differences ($p = 0.003$). The overall survival of TCGA CRC patients with Low-TIDE/Low-Risk was lower than that with High-TIDE/Low-risk with statistical differences ($p = 0.007$). Moreover, there were no statistical differences of the overall survival between TCGA CRC patients with High-TIDE/High-Risk VS TCGA CRC patients with Low-TIDE/High-Risk ($p = 0.135$) and TCGA CRC patients with High-TIDE/High-Risk VS TCGA CRC patients with High-TIDE/Low-Risk ($p = 0.692$).

Gene Set Enrichment Analysis

Targeting to gene expression matrix in the high- and low-risk groups, we performed GSEA to hierarchically reveal the underlying biological function in the TCGA CRC patients. In the results of the GSEA for the HALLMARK functional pathways (**Figure 10A**), the gene expression matrix in the high-risk group was mainly enriched in the frequently investigated cancer-related pathways, such as epithelial-mesenchymal transition (NOM $p = 0.002$, FDR $q = 0.076$), angiogenesis (NOM $p = 0.035$, FDR $q = 0.215$), and KRAS signaling up (NOM $p = 0.044$, FDR $q = 0.180$), while the gene expression matrix in the low-risk group was mainly enriched in the metabolism-related pathways, namely, oxidative phosphorylation (NOM $p = 0.000$, FDR $q = 0.020$) and fatty acid metabolism (NOM $p = 0.002$, FDR $q = 0.013$). In the results of the GSEA for the KEGG functional pathways (**Figure 10B**), the gene expression matrix in the high-risk

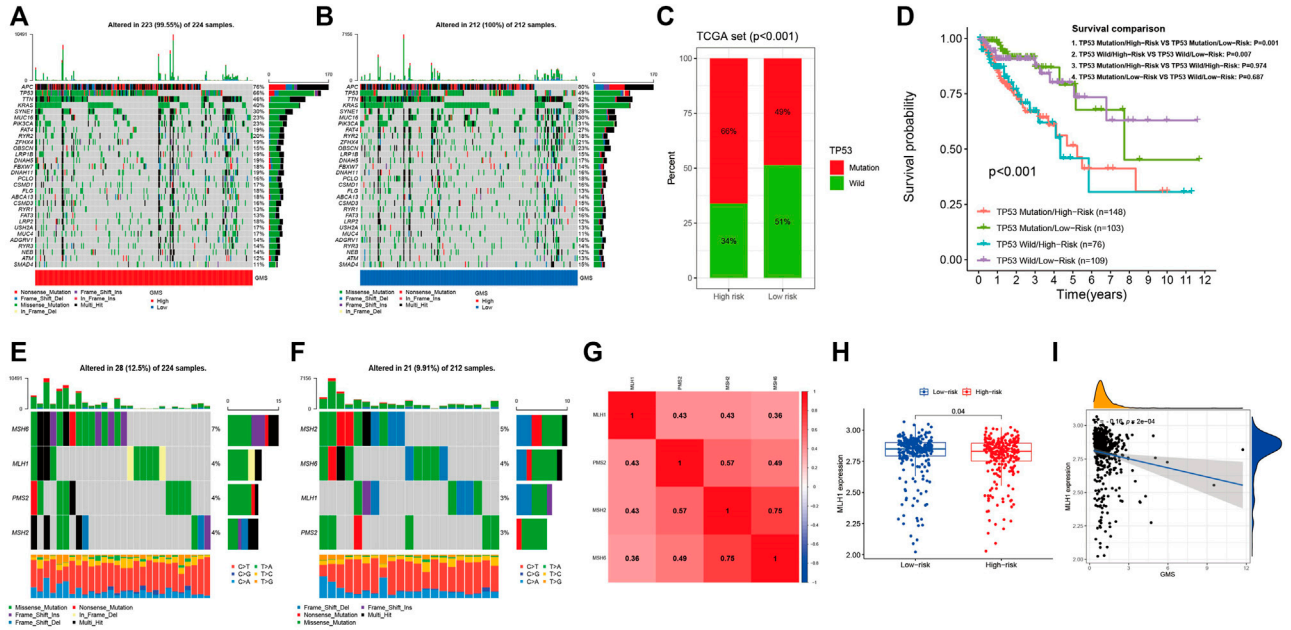


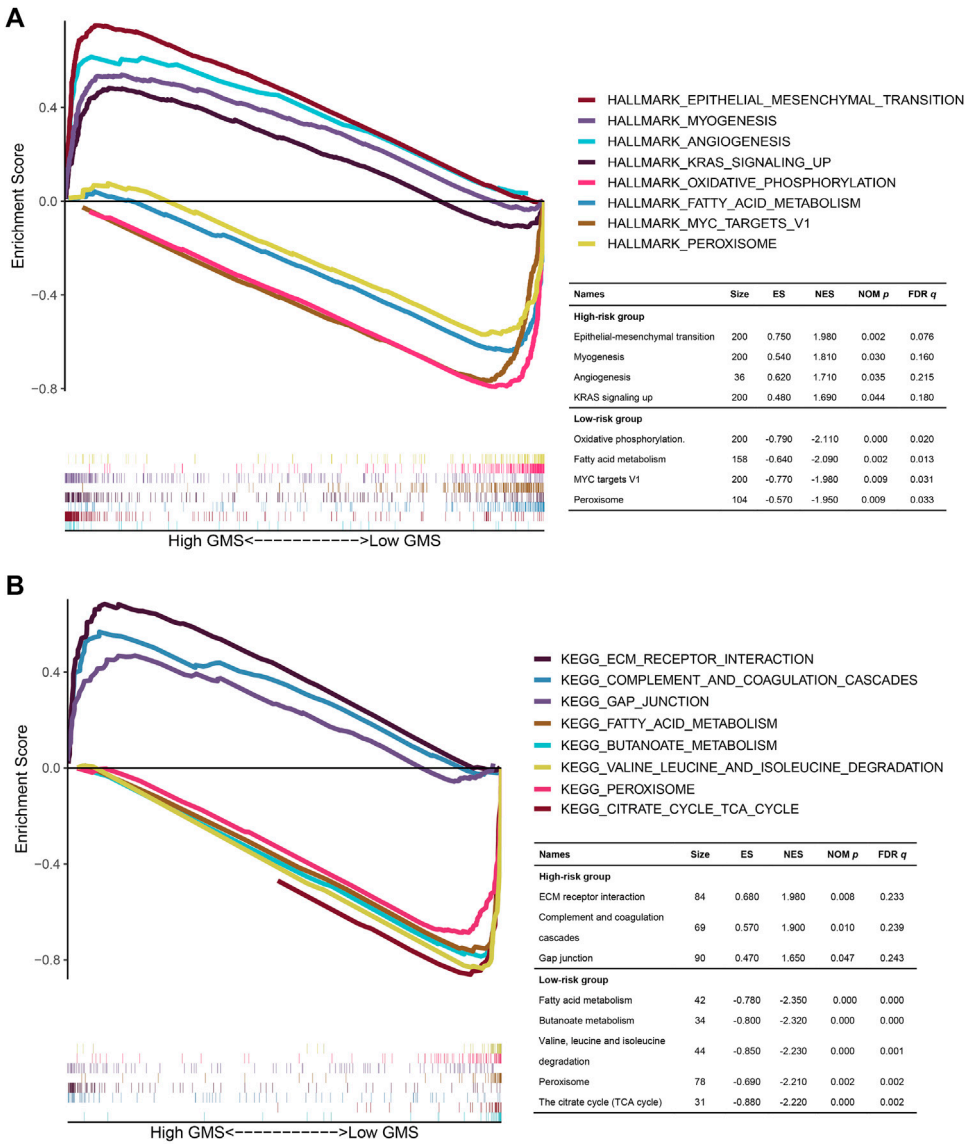
FIGURE 9 | The landscape of somatic mutation and four characteristic genes of MMR system. **(A,B)**: The waterfall demonstrating the top 30 genes with highest mutation frequencies. **(C)**: The bar plot indicating that the mutational frequencies of TP53 in the high-risk group were higher than that in the low-risk group. **(D)**: The K-M survival curves stratified by the five-GMRHG prognostic signature and TP53 mutational status (TP53 wild and TP53 mutation). The overall survival of TCGA CRC patients with TP53 Mutation/High-risk was lower than that with TP53 mutation/Low-risk with statistical differences (p = 0.001). The overall survival of TCGA CRC patients with TP53 Wild/High-risk was lower than that with TP53 Wild/Low-risk with statistical differences (p = 0.007). Moreover, there were no statistical differences of the overall survival between the TCGA CRC patients in the high- or low-risk group with different mutation status (TP53 Mutation/High-risk VS TP53 Wild/High-risk: p = 0.974; TP53 Mutation/Low-risk VS TP53 Wild/Low-risk: p = 0.687). **(E,F)**: The waterfall demonstrating the mutational frequencies of MSH6, MLH1, PMS2, and MSH2 in the high- and low-risk high group. **(G)**: The co-expression correlations among MSH6, MLH1, PMS2, and MSH2. **(H)**: The expression difference of MLH1 in the high- and low-risk groups. **(I)**: The positive liner correlation between the expression of MLH1 and the GMS.

group was mainly enriched in ECM receptor interaction (NOM p = 0.008, FDR q = 0.233), complement and coagulation cascades (NOM p = 0.010, FDR q = 0.239), and gap junction (NOM p = 0.047, FDR q = 0.243), while the gene expression matrix in the low-risk group was mainly enriched in 33 KEGG functional pathways, among which four of the top five pathways with the highest NES value adhering to the filtering standard (NOM p < 0.05, FDR q < 0.25) were mostly metabolism-related pathways (fatty acid metabolism (NOM p = 0.000, FDR q = 0.000), butanoate metabolism (NOM p = 0.000, FDR q = 0.000), valine leucine and isoleucine degradation (NOM p = 0.000, FDR q = 0.001), and citrate cycle (TCA cycle) (NOM p = 0.000, FDR q = 0.002). We also noted that in the analysis of HALLMARK and KEGG, the gene expression matrix in the low-risk group was both enriched in fatty acid metabolism and peroxisome, indicating that the CRC patients in the low-risk group might benefit from intervention against the fatty acid metabolism and peroxisome (HALLMARK: NOM p = 0.002, FDR q = 0.013; KEGG: NOM p = 0.002, FDR q = 0.002).

DISCUSSION

There are growing number of evidences indicated that an anomaly of lipid metabolism level contributes to the malignancy progression. As the result of the high proliferative activities, tumor cells are inclined to obtain the lipids with *de novo* synthesis, which gives rise to abnormal levels of lipase and signal transduction factors (Bao et al., 2016; Ru et al., 2017). Therefore, analyzing the differences in expression patterns of lipid metabolic mediators between normal samples and tumor samples would be conducive to determine the specific biological markers for the tumor diagnosis (Day et al., 2013).

In this study, 175 differential lipid metabolites were subjected to univariate and multivariate Cox regression analysis and we identified six lipid metabolites of the prognostic signature, including five high-risk lipid metabolites, namely, Cer(d18:0/14:0), LysoPE[22:6(4Z,7Z,10Z,13Z,16Z, 19Z)/0:0], PS [20:4(5Z,8Z,11Z, 14Z)/14:1(9Z)], PA[20:3(5Z,8Z, 11Z)/24:1(15Z)], and Ganglioside GT3[d18:0/18:1(9Z)], and one low-risk lipid metabolites (Substance P). Ceramide (Cer), one of the central active components of sphingolipid, is mainly composed of sphingoid long-chain base (LCB) and amide-linked to a fatty acyl chain (Ogretmen, 2018). Cer could be synthesized from serine and palmitate with *de novo* synthesis, and it could induce the apoptosis of cancer cells (Kramer et al., 2015) and protect cancer cells from apoptosis (Mesicek et al., 2010). Judith et al. found that the overexpression of CerS2 increased production of C(24:0) ceramide, which could protect cells from apoptosis induced by ionizing radiation (IR), and C(16:0) ceramide



Names	Size	ES	NES	NOM *p*	FDR *q*
High-risk group					
Epithelial-mesenchymal transition	200	0.750	1.980	0.002	0.076
Myogenesis	200	0.540	1.810	0.030	0.160
Angiogenesis	36	0.620	1.710	0.035	0.215
KRAS signaling up	200	0.480	1.690	0.044	0.180
Low-risk group					
Oxidative phosphorylation.	200	-0.790	-2.110	0.000	0.020
Fatty acid metabolism	158	-0.640	-2.090	0.002	0.013
MYC targets V1	200	-0.770	-1.980	0.009	0.031
Peroxisome	104	-0.570	-1.950	0.009	0.033

Names	Size	ES	NES	NOM *p*	FDR *q*
High-risk group					
ECM receptor interaction	84	0.680	1.980	0.008	0.233
Complement and coagulation cascades	69	0.570	1.900	0.010	0.239
Gap junction	90	0.470	1.650	0.047	0.243
Low-risk group					
Fatty acid metabolism	42	-0.780	-2.350	0.000	0.000
Butanoate metabolism	34	-0.800	-2.320	0.000	0.000
Valine, leucine and isoleucine degradation	44	-0.850	-2.230	0.000	0.001
Peroxisome	78	-0.690	-2.210	0.002	0.002
The citrate cycle (TCA cycle)	31	-0.880	-2.220	0.000	0.002

FIGURE 10 | The GSEA for the TCGA CRC patients in the high- and low-risk groups. **(A,B)**: The multi-GSEA plots depicting the hallmark and KEGG functional pathways enriched in the gene matrices of TCGA CRC patients in the high- and low-risk groups. The curves above the X-axis represented the functional pathways enriched in the gene matrices of TCGA CRC patients in the high-risk group. Correspondingly, the curves below the X-axis represented the functional pathways enriched in the gene matrices of TCGA CRC patients in the low-risk group.

produced by the overexpression of CERS5 and CERS6, which contributed to IR-mediated apoptosis in HeLa cells(31). In this study, we found that Cer(d18:0/14:0) was a high-risk factor to predict the prognosis of CRC, also indirectly proving the role of Cer(d18:0/14:0) in the progression of CRC. LysoPE is the hemolytic metabolic product of phosphatidylethanolamine (Park et al., 2013). It has been shown that LysoPE was involved in intercellular signal transduction through G protein coupled receptor (GPCR) (Park et al., 2013). The level of LysoPE in colon cancer tissue is significantly higher than that in normal colon epithelium (Hofmanová et al., 2021). This study further found that the level of LysoPE in mucinous adenocarcinoma was higher than that in adenocarcinoma at the serum level, which indirectly confirmed that LysoPE, as a high-risk factor, could predict the prognosis of CRC patients. Phosphatidylserine (PS), the only kind of phospholipid that could regulate the functional state of critical proteins in cell membrane, distributed on the vesicle membrane and participated in the process of related diseases (Skotland et al., 2019; Wodlej et al., 2019). Ran et al. emphasized that the exposure of PS in tumor cells was induced by oxidative stress and cytokines, and the exposed PS binding to immune cells regulated the immune infiltrations and biological process of tumor cells (Ran and Thorpe, 2002). Previous studies have shown that fatty acid chain elongation, one of the characteristics of malignant tumors, was represented by PS in prostate cancer (Butler et al., 2021). In CRC, Jiřina et al. utilized

LC-MS to determine phospholipid components, including phosphatidylserine as the biological markers to distinguish normal colorectal epithelial cells from CRC cells. The clinical correlation analysis also suggested that the level of PS was higher in the advanced pathological stage (stage III-IV) and the mucinous adenocarcinoma, which further demonstrated that PS was a high-risk factor affecting the prognosis of patients with colorectal cancer. Phosphatidic acid (PA) was found to be a high-risk factor in this study. Although it is a type of membrane phospholipid with a simple structure and it accounted for a minority of membrane lipid components, it exists in the cell membrane of almost all cells (Young et al., 2010; Kulig et al., 2019). Previous studies have shown that PA is mostly correlated with apoptosis regulatory signals that promote the proliferation of tumor cells. Zhao et al. proposed that the combination of PA and tBid would enhance lysosomal membrane permeabilization (LMP), thus triggering the apoptotic signaling (Zhao et al., 2012). Moreover, Taga et al. found that PA stimulated the mTORC1 signaling pathway to protect human neuroblastoma cells against apoptosis induced by oxygen stress (Taga et al., 2011). The latest study further confirmed that PA has been proven to be one of the key lipid signals mediating the activation of classic cancer signaling pathway-Hippo pathway (Han et al., 2018). Ganglioside GT3 is a kind of ganglioside mainly composed of glycosphingolipids and sialic acids linking to sugar chain (Labrada et al., 2018). It is a type of component concentrating on the lipid raft in the plasma membrane (Campos et al., 2021). The main difference among more than 60 known gangliosides depends on the locations and numbers of NANA residues (Fuse et al., 1982). Previous studies have shown that ganglioside GD2/3 (Yoshida et al., 2001) and GM3 (Hayashi et al., 2013) played important roles in tumor cell invasions and metastasis, but the role of ganglioside GT3 in malignancies including CRC has not been clearly elaborated. This study found that ganglioside GT3 was a high-risk factor for predicting the prognosis of CRC, and it might also provide inspiration for the role of ganglioside GT3 in CRC. SP is a member of tachykinin peptide family. It is located in the central nervous system and other several peripheral tissues, including the entire length of the gastrointestinal and colorectal tract. The biological effect of SP is mediated by three different G protein coupled receptors (GPCR), which were neurokinin (NK)-1, 2 and 3, respectively. SP had high affinity for NK-1 receptor (NK-1R) and low affinity for NK-2 and 3 receptors (Muñoz and Coveñas, 2020). Reubi et al. found that the levels of SP and NK1R in CRC tissue were higher than those in normal tissue (Lorestani et al., 2020). The SP content in proximal vein of CRC tissue was three to five times higher than that in distal vein of CRC tissue, suggesting that the distribution and regulation mechanism of SP in vascular bed are closely related to the mechanism of CRC metastasis (Reubi et al., 1996). The clinical correlation analysis of this study revealed that the level of SP in adenocarcinoma was higher than that in mucinous adenocarcinoma. SP as a low-risk factor to predict the prognosis of CRC patients might be related with this. Moreover, we utilized the independent prognostic factors to plot a nomogram to systematically evaluate the overall survival of CRC patients. The calibration curve and ROC curve indicated that the nomogram had favorable clinical application and accuracy in predicting the prognosis of CRC patients. Furthermore, through functional enrichment analysis, we found that 175 kinds of differential lipid metabolites were significantly enriched in glycerophospholipid metabolism with statistical significance (FDR = 0.0016, Impact = 0.403). Glycerophospholipids, the most abundant membrane lipids, are composed of glycerol backbone, two fatty acid chains, and a polar head group, which mainly produces phospholipid substances in human cells, such as phosphatidylserine (PS), phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylglycerol (PG), and phosphatidylinositol (PI). However, for now, there is no systematic exposition on the guiding role of glycerophospholipid metabolism in the prognosis of CRC patients. Herein, we have elaborately expounded the prognostic and molecular biological functions of glycerophospholipid metabolism from the transcriptomic level. Integrated with the prognostic information of 877 CRC patients, we developed the five-GMRHG (ACOX1, ATOH1, CPT2, PCSK5, and TINCR) prognostic signature to conduct stratified prognostic analysis for CRC patients. Of the five GMRHGs, PCSK5 and TINCR were high-risk GMRHGs and ACOX1, ATOH1, and CPT2 were low-risk GMRHGs. ACOX1, as a key enzyme involved in fatty acid oxidation (FAO) pathway, has been proved to be downregulated by p38MAPK/PPARα signaling pathway regulated by CD147 (a transmembrane protein regulated by lipid metabolism), thus inhibiting fatty acid-β oxidation (Li et al., 2015). Sun et al. have also found that ACOX1 overexpression weakened the enhancement of miR-15b-5p overexpression on the migration and invasion of CRC cells, and SIRT1/miR-15b5p/ACOX1 axis played an important role in the metastasis of CRC (Sun et al., 2017). ATOH1 is a member of the basic helix-loop-helix (BHLH) family of transcription factors (Zheng et al., 2011). Previous studies have shown that ATOH1 played an important role in the differentiation of intestinal secretory cells. Shroyer et al. found the loss of intestinal Paneth, goblet cells, and intestinal endocrine cells in adult rats after knocking down ATOH1 (Shroyer et al., 2007). Recent studies have shown that the overexpression of ATOH1 promoted functional intestinal stem cells to differentiate into secretory cells, and its expression was regulated by LKB1 (Gao et al., 2020). In the clinical correlation analysis, CPT2 exhibited higher expression levels in the early pathological stage and N staging, which could support the role of CPT2 as a low-risk GMRHG for CRC. PCSK5 is a type of proprotein convertase that can affect the metastatic property of CRC by participating in the production of adhesion molecules and growth factors (Bontemps et al., 2007). It has been proven that abnormally upregulated expression of PCSK5 indicates increased metastatic potential of CRC (Scamuffa et al., 2008). TINCR is a key lncRNA involved in the stability of mRNA which is responsible for the maintenance of epidermal tissue differentiation (Kretz et al., 2013). Yu et al. reported that TINCR expression level could be independently prognostic for CRC, and its abnormally upregulated expression was associated with more adverse outcomes. In the meantime, they also found that *in vitro* knockout of TINCR contributed to significantly decreased levels of cancer cell proliferation, while the

overexpression of TINCR was conducive to cancer progression with the involvement of miR-7-5p and PI3K/Akt/mTOR signaling (Yu et al., 2019). In the present study, CRC patients were grouped by the median GMS was proven to be significant in prognostic prediction with K-M survival analysis, independent prognostic analysis, and time-dependent ROC curves. Additionally, GSEA pathway analysis was performed and revealed much more highly activated pathways involved in epithelial-mesenchymal transition (EMT), angiogenesis, and KRAS signaling up in the high-risk group. It is known that EMT is active in the presence of most tumors, especially epithelial tumors. It is characterized by the gradual loss of epithelial properties while gaining mesenchymal properties as the tumor progresses (Grünert et al., 2003; Dongre and Weinberg, 2019), and it has been proven to have a close relationship with the distant metastasis of CRC (Lambert et al., 2017). Early in the 1970s, Folkman et al. reported that the occurrence and progression of tumors were accompanied by active angiogenesis (Folkman, 1971). Vascular endothelial growth factor (VEGF) is a typical gene involved in angiogenesis. Ferrara et al. revealed that suppression of VEGF expression could decrease vascular density and block tumor progression (Kim et al., 1993). Application of first-line anti-vascular agent Bevacizumab was conducive to largely prolonging the progression-free survival (PFS) of patients with metastatic CRC (Cunningham et al., 2013). Aashiq et al. found that PI3K signaling was crucial in tumor progression by mediating lipid metabolism and glycolysis. They reported that blockade of the PI3K signaling suppressed CRC cell migration and invasion following reductions in metabolic flux and tumor angiogenesis (Hussain et al., 2016). EMT and angiogenesis are tightly correlated in oncogenesis and metastasis (Sánchez-García, 2009), while the crosstalk between VEGF and Notch, an epidermal cell proliferation suppressor, under hypoxic stimuli can augment the EMT in cancer cells (Holderfield and Hughes, 2008). Here, EMT and angiogenesis were active in the high-risk group, indicative that the two biological functions might be much more significant in tumor progression of the high-risk patients.

Further, the characteristics depicting TME and somatic mutation map were hierarchically viewed for the CRC patients. It was revealed that aDCs, CD8+_T_cells, macrophages, T_helper_cells, Tfh, Th1_cells, TIL, CCR, checkpoint, HLA, inflammation-promoting, T_cell_co-inhibition, and T_cell_co-stimulation exhibited higher levels in the high-risk group versus the low-risk group with statistical differences. Meanwhile, it indirectly demonstrated that CRC patients in the high-risk had higher immune activities. Moreover, in the previous research on melanoma, it was found that the immune escape was realized based on T cells in the TME. Specifically, T cells target to recognize tumor antigen followed by upregulating PD-1 and expressing IFN-γ which can contribute to expressing PD-L1 by immune and tumor cells. In that setting, the expressed PD-L1 will bind to PD-1 to block T cell surveillance, ultimately leading to immune escape (Taube et al., 2012). Here, the high-risk group had an increase in the expression of IFN-γ, PD-1, and PD-L1 as analyzed by the TIDE algorithm, and it was more likely to experience immune escape. Similar processes could also occur in CRC. Somatic mutational analysis showed that mutations in APC, TP53, TTN, KRAS, and SYNE were highly frequent in both high- and low-risk groups. In most cases, biallelic inactivation of APC is the culprit for CRC occurrence and it plays a key role in tumor progression to adenomas by activating responsible pathways (Fennell et al., 2020). It is reported that CRC progression is largely associated with the mutations in KRAS and TP53 genes (Eklöf et al., 2013; Fennell et al., 2020). KRAS encodes K-Ras protein, which belongs to the Ras-Raf-MEK-ERK signaling pathway, thought to be responsible for CRC growth and proliferation (Eklöf et al., 2013). TP53 is a tumor suppressor that is abnormally mutated to CRC progression. Additionally, TP53 mutations can be independently prognostic for the outcome of CRC, predicting more adverse outcomes as well as more significant resistance to chemotherapeutics (Li et al., 2019). We here noted that the mutation frequency of TP53 in high-risk group was much higher versus that of the low-risk group, which was statistically significant. In the K-M survival analysis, CRC patients in the high-risk had a shorter survival under the same TP53 mutation frequency. This suggests that the GMS-based grouping method is superior to the conventional TP53-directed method for predicting the overall survival of CRC patients. MMR includes MLH1, MSH2, MSH6, and PMS2. Mutations or modifications (such as methylation) of their upstream genes responsible for protein synthesis can lead to defects of the MMR genes. In that circumstance, the replication errors in MS areas fail to get corrected, leading to error accumulation and the subsequent occurrence of MSI (Sahin et al., 2019). It has been reported that MSI affects approximately 15–20% of CRC patients and is of vital significance in CRC initiation and development (Richman, 2015; Sahin et al., 2019). In this study, the MMR gene mutation frequency was 19% in the high-risk group, compared to 15% in the low-risk group, which is consistent with the previous finding. Among the four MMR genes, MLH1 was proven to be abnormally silenced in CRC, and the silencing or mutation of MLH1 might be crucial in tumor progression (Ma et al., 2016; Han et al., 2020). Moreover, MLH1 downregulation is associated with poor outcome of CRC and can be independently prognostic for PFS (Han et al., 2020). Further differential expression and Spearman correlational analysis revealed that MLH1 exhibited higher expression in the low-risk group relative to the high-risk group. This supported the low expression of MLH1 in the high-risk group and its prognostic significance for poor outcome of CRC.

In summary, we systematically investigated the role of lipid metabolism in monitoring the prognosis of CRC patients with metabonomics and transcriptomic methods, and constructed a composite nomogram individually predicting the overall survival of CRC patients based on the independent prognostic factors (six-lipid-metabolite prognostic signature, pathological stage, histological type, and T staging). The five-GMRHG prognostic signature can not only be used to predict the prognosis of CRC patients, but also shed light on the CRC patients' current immune infiltration status, somatic mutational landscape, and potential biological functions, and provided inspiration for individualized treatment of CRC patients in the future.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the MetaboLights repository, accession number MTBLS3945.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the First Hospital of Jilin University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YZ and YS designed the study; BL carried out the data collection; YC and YX conducted the data analysis; YS and BL drafted the manuscript; YZ revised the manuscript. All authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.811957/full#supplementary-material

REFERENCES

Angelova, M., Charoentong, P., Hackl, H., Fischer, M. L., Snajder, R., Krogsdam, A. M., et al. (2015). Characterization of the Immunophenotypes and Antigenomes of Colorectal Cancers Reveals Distinct Tumor Escape Mechanisms and Novel Targets for Immunotherapy. *Genome Biol.* 16, 64. doi:10.1186/s13059-015-0620-6

Aran, V., Victorino, A. P., Thuler, L. C., and Ferreira, C. G. (2016). Colorectal Cancer: Epidemiology, Disease Mechanisms and Interventions to Reduce Onset and Mortality. *Clin. Colorectal Cancer* 15, 195–203. doi:10.1016/j.clcc.2016.02.008

Bao, J., Zhu, L., Zhu, Q., Su, J., Liu, M., and Huang, W. (2016). SREBP-1 Is an Independent Prognostic Marker and Promotes Invasion and Migration in Breast Cancer. *Oncol. Lett.* 12, 2409–2416. doi:10.3892/ol.2016.4988

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA Interference Reveals that Oncogenic KRAS-Driven Cancers Require TBK1. *Nature* 462, 108–112. doi:10.1038/nature08460

Bontemps, Y., Scamuffa, N., Calvo, F., and Khatib, A.-M. (2007). Potential Opportunity in the Development of New Therapeutic Agents Based on Endogenous and Exogenous Inhibitors of the Proprotein Convertases. *Med. Res. Rev.* 27, 631–648. doi:10.1002/med.20072

Butler, L. M., Mah, C. Y., Machiels, J., Vincent, A. D., Irani, S., Mutuku, S. M., et al. (2021). Lipidomic Profiling of Clinical Prostate Cancer Reveals Targetable Alterations in Membrane Lipid Composition. *Cancer Res.* 81, 4981–4993. doi:10.1158/0008-5472.CAN-20-3863

Campos, F. S. O., Piña-Rodrigues, F. M., Reis, A., Atella, G. C., Mermelstein, C. S., Allodi, S., et al. (2021). Lipid Rafts from Olfactory Ensheathing Cells: Molecular Composition and Possible Roles. *Cell Mol Neurobiol* 41, 525–536. doi:10.1007/s10571-020-00869-4

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., et al. (2018). MetaboAnalyst 4.0: towards More Transparent and Integrative Metabolomics Analysis. *Nucleic Acids Res.* 46, W486–W494. doi:10.1093/nar/gky310

Cunningham, D., Lang, I., Marcuello, E., Lorusso, V., Ocvirk, J., Shin, D. B., et al. (2013). Bevacizumab Plus Capecitabine versus Capecitabine Alone in Elderly Patients with Previously Untreated Metastatic Colorectal Cancer (AVEX): an Open-Label, Randomised Phase 3 Trial. *Lancet Oncol.* 14, 1077–1085. doi:10.1016/S1470-2045(13)70154-2

Day, S. D., Enos, R. T., McClellan, J. L., Steiner, J. L., Velázquez, K. T., and Murphy, E. A. (2013). Linking Inflammation to Tumorigenesis in a Mouse Model of High-Fat-Diet-Enhanced colon Cancer. *Cytokine* 64, 454–462. doi:10.1016/j.cyto.2013.04.031

De Schrijver, E., Brusselmans, K., Heyns, W., Verhoeven, G., and Swinnen, J. V. (2003). RNA Interference-Mediated Silencing of the Fatty Acid Synthase Gene Attenuates Growth and Induces Morphological Changes and Apoptosis of LNCaP Prostate Cancer Cells. *Cancer Res.* 63, 3799–3804.

DeLaney, K., Sauer, C., Vu, N., and Li, L. (2018). Recent Advances and New Perspectives in Capillary Electrophoresis-Mass Spectrometry for Single Cell "Omics". *Molecules* 24, 42. doi:10.3390/molecules24010042

Dongre, A., and Weinberg, R. A. (2019). New Insights into the Mechanisms of Epithelial-Mesenchymal Transition and Implications for Cancer. *Nat. Rev. Mol. Cel Biol* 20, 69–84. doi:10.1038/s41580-018-0080-4

Eklöf, V., Wikberg, M. L., Edin, S., Dahlin, A. M., Jonsson, B.-A., Öberg, Å., et al. (2013). The Prognostic Role of KRAS, BRAF, PIK3CA and PTEN in Colorectal Cancer. *Br. J. Cancer* 108, 2153–2163. doi:10.1038/bjc.2013.212

Fennell, L. J., Kane, A., Liu, C., McKeone, D., Fernando, W., Su, C., et al. (2020). APC Mutation Marks an Aggressive Subtype of BRAF Mutant Colorectal Cancers. *Cancers* 12, 1171. doi:10.3390/cancers12051171

Folkman, J. (1971). Tumor Angiogenesis: Therapeutic Implications. *N. Engl. J. Med.* 285, 1182–1186. doi:10.1056/NEJM197111182852108

Fuse, A., Handa, S., and Kuwata, T. (1982). Effect of Glycolipids and Glycophorin on the Activity of Human Interferon-β and -γ. *Antiviral Res.* 2, 161–166. doi:10.1016/0166-3542(82)90018-3

Gao, P., Zhou, C., Zhao, L., Zhang, G., and Zhang, Y. (2016). Tissue Amino Acid Profile Could Be Used to Differentiate Advanced Adenoma from Colorectal Cancer. *J. Pharm. Biomed. Anal.* 118, 349–355. doi:10.1016/j.jpba.2015.11.007

Gao, Y., Yan, Y., Tripathi, S., Pentinmikko, N., Amaral, A., Päivinen, P., et al. (2020). LKB1 Represses ATOH1 via PDK4 and Energy Metabolism and Regulates Intestinal Stem Cell Fate. *Gastroenterology* 158, 1389–1401. doi:10.1053/j.gastro.2019.12.033

Grünert, S., Jechlinger, M., and Beug, H. (2003). Diverse Cellular and Molecular Mechanisms Contribute to Epithelial Plasticity and Metastasis. *Nat. Rev. Mol. Cel Biol* 4, 657–665. doi:10.1038/nrm1175

Han, H., Qi, R., Zhou, J. J., Ta, A. P., Yang, B., Nakaoka, H. J., et al. (2018). Regulation of the Hippo Pathway by Phosphatidic Acid-Mediated Lipid-Protein Interaction. *Mol. Cel* 72, 328–340. doi:10.1016/j.molcel.2018.08.038

Han, Y., Peng, Y., Fu, Y., Cai, C., Guo, C., Liu, S., et al. (2020). MLH1 Deficiency Induces Cetuximab Resistance in Colon Cancer via Her-2/PI3K/AKT Signaling. *Adv. Sci.* 7, 2000112. doi:10.1002/advs.202000112

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* 14, 7. doi:10.1186/1471-2105-14-7

Hayashi, N., Chiba, H., Kuronuma, K., Go, S., Hasegawa, Y., Takahashi, M., et al. (2013). Detection of N-Glycolyated Gangliosides in Non-small-cell Lung Cancer Using GMR8 Monoclonal Antibody. *Cancer Sci.* 104, 43–47. doi:10.1111/cas.12027

Hofmanová, J., Slavík, J., Ciganek, M., Ovesná, P., Tylichová, Z., Karasová, M., et al. (2021). Complex Alterations of Fatty Acid Metabolism and Phospholipidome Uncovered in Isolated Colon Cancer Epithelial Cells. *Ijms* 22, 6650. doi:10.3390/ijms22136650

Holderfield, M. T., and Hughes, C. C. W. (2008). Crosstalk between Vascular Endothelial Growth Factor, Notch, and Transforming Growth Factor-β in Vascular Morphogenesis. *Circ. Res.* 102, 637–652. doi:10.1161/CIRCRESAHA.107.167171

Hung, Y.-H., Chan, Y.-S., Chang, Y.-S., Lee, K.-T., Hsu, H.-P., Yen, M.-C., et al. (2014). Fatty Acid Metabolic Enzyme Acyl-CoA Thioesterase 8 Promotes the Development of Hepatocellular Carcinoma. *Oncol. Rep.* 31, 2797–2803. doi:10.3892/or.2014.3155

Hussain, A., Qazi, A. K., Mupparapu, N., Guru, S. K., Kumar, A., Sharma, P. R., et al. (2016). Modulation of Glycolysis and Lipogenesis by Novel PI3K Selective

Molecule Represses Tumor Angiogenesis and Decreases Colorectal Cancer Growth. *Cancer Lett.* 374, 250–260. doi:10.1016/j.canlet.2016.02.030

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T Cell Dysfunction and Exclusion Predict Cancer Immunotherapy Response. *Nat. Med.* 24, 1550–1558. doi:10.1038/s41591-018-0136-1

Kim, K. J., Li, B., Winer, J., Armanini, M., Gillett, N., Phillips, H. S., et al. (1993). Inhibition of Vascular Endothelial Growth Factor-Induced Angiogenesis Suppresses Tumour Growth *In Vivo*. *Nature* 362, 841–844. doi:10.1038/362841a0

Kitahara, C. M., Berrington de González, A., Freedman, N. D., Huxley, R., Mok, Y., Jee, S. H., et al. (2011). Total Cholesterol and Cancer Risk in a Large Prospective Study in Korea. *Jco* 29, 1592–1598. doi:10.1200/JCO.2010.31.5200

Kramer, R., Bielawski, J., Kistner-Griffin, E., Othman, A., Alecu, I., Ernst, D., et al. (2015). Neurotoxic 1-deoxysphingolipids and Paclitaxel-induced Peripheral Neuropathy. *FASEB j.* 29, 4461–4472. doi:10.1096/fj.15-272567

Kretz, M., Siprashvili, Z., Chu, C., Webster, D. E., Zehnder, A., Qu, K., et al. (2013). Control of Somatic Tissue Differentiation by the Long Non-coding RNA TINCR. *Nature* 493, 231–235. doi:10.1038/nature11661

Kulig, W., Korolainen, H., Zatorska, M., Kwolek, U., Wydro, P., Kepczynski, M., et al. (2019). Complex Behavior of Phosphatidylcholine-Phosphatidic Acid Bilayers and Monolayers: Effect of Acyl Chain Unsaturation. *Langmuir* 35, 5944–5956. doi:10.1021/acs.langmuir.9b00381

Labrada, M., Dorvignit, D., Hevia, G., Rodríguez-Zhurbenko, N., Hernández, A. M., Vázquez, A. M., et al. (2018). GM3(Neu5Gc) Ganglioside: an Evolution Fixed Neoantigen for Cancer Immunotherapy. *Semin. Oncol.* 45, 41–51. doi:10.1053/j.seminoncol.2018.04.003

Lambert, A. W., Pattabiraman, D. R., and Weinberg, R. A. (2017). Emerging Biological Principles of Metastasis. *Cell* 168, 670–691. doi:10.1016/j.cell.2016.11.037

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Li, H., Zhang, J., Tong, J. H. M., Chan, A. W. H., Yu, J., Kang, W., et al. (2019). Targeting the Oncogenic P53 Mutants in Colorectal Cancer and Other Solid Tumors. *Ijms* 20, 5999. doi:10.3390/ijms20235999

Li, J., Huang, Q., Long, X., Zhang, J., Huang, X., Aa, J., et al. (2015). CD147 Reprograms Fatty Acid Metabolism in Hepatocellular Carcinoma Cells through Akt/mTOR/SREBP1c and P38/PPARα Pathways. *J. Hepatol.* 63, 1378–1389. doi:10.1016/j.jhep.2015.07.039

Liao, F., He, W.-z., Jiang, C., Yin, C.-x., Guo, G.-f., Chen, X., et al. (2015). A High LDL-C to HDL-C Ratio Predicts Poor Prognosis for Initially Metastatic Colorectal Cancer Patients with Elevations in LDL-C. *Ott* 8, 3135–3142. doi:10.2147/OTT.S90479

Lodi, A., Tiziani, S., Khanim, F. L., Günther, U. L., Viant, M. R., Morgan, G. J., et al. (2013). Proton NMR-Based Metabolite Analyses of Archived Serial Paired Serum and Urine Samples from Myeloma Patients at Different Stages of Disease Activity Identifies Acetylcarnitine as a Novel Marker of Active Disease. *PLoS One* 8, e56422. doi:10.1371/journal.pone.0056422

Lorestani, S., Ghahremanloo, A., Jangjoo, A., Abedi, M., and Hashemy, S. I. (2020). Evaluation of Serum Level of Substance P and Tissue Distribution of NK-1 Receptor in Colorectal Cancer. *Mol. Biol. Rep.* 47, 3469–3474. doi:10.1007/s11033-020-05432-4

Lupu, R., and Menendez, J. (2006). Pharmacological Inhibitors of Fatty Acid Synthase (FASN)-Catalyzed Endogenous Fatty Acid Biogenesis: A New Family of Anti-cancer Agents? *Cpb* 7, 483–494. doi:10.2174/138920106779116928

Ma, G., Ge, Y., Gu, D., Du, M., Chu, H., Chen, J., et al. (2016). Functional Annotation of Colorectal Cancer Susceptibility Loci identifiesMLH1rs1800734 Associated with MSI Patients. *Gut* 65, 1227–1228. doi:10.1136/gutjnl-2016-311543

Mesicek, J., Lee, H., Feldman, T., Jiang, X., Skobeleva, A., Berdyshev, E. V., et al. (2010). Ceramide Synthases 2, 5, and 6 Confer Distinct Roles in Radiation-Induced Apoptosis in HeLa Cells. *Cell Signal.* 22, 1300–1307. doi:10.1016/j.cellsig.2010.04.006

Muka, T., Kraja, B., Ruiter, R., de Keyser, C. E., Hofman, A., Stricker, B. H., et al. (2016). Dietary Polyunsaturated Fatty Acids Intake Modifies the Positive Association between Serum Total Cholesterol and Colorectal Cancer Risk: the Rotterdam Study. *J. Epidemiol. Community Health* 70, 881–887. doi:10.1136/jech-2015-206556

Muñoz, M., and Coveñas, R. (2020). Neurokinin Receptor Antagonism: a Patent Review (2014-present). *Expert Opin. Ther. Patents* 30, 527–539. doi:10.1080/13543776.2020.1769599

Ogretmen, B. (2018). Sphingolipid Metabolism in Cancer Signalling and Therapy. *Nat. Rev. Cancer* 18, 33–50. doi:10.1038/nrc.2017.96

Park, S.-J., Lee, K.-P., Kang, S., Chung, H.-Y., Bae, Y.-S., Okajima, F., et al. (2013). Lysophosphatidylethanolamine Utilizes LPA1 and CD97 in MDA-MB-231 Breast Cancer Cells. *Cell Signal.* 25, 2147–2154. doi:10.1016/j.cellsig.2013.07.001

Ramakrishna, M., Williams, L. H., Boyle, S. E., Bearfoot, J. L., Sridhar, A., Speed, T. P., et al. (2010). Identification of Candidate Growth Promoting Genes in Ovarian Cancer through Integrated Copy Number and Expression Analysis. *PLoS One* 5, e9983. doi:10.1371/journal.pone.0009983

Ran, S., and Thorpe, P. E. (2002). Phosphatidylserine Is a Marker of Tumor Vasculature and a Potential Target for Cancer Imaging and Therapy. *Int. J. Radiat. Oncology*Biology*Physics* 54, 1479–1484. doi:10.1016/s0360-3016(02)03928-7

Reubi, J., Mazzucchelli, L., Hennig, I., and Laissue, J. (1996). Local Up-Regulation of Neuropeptide Receptors in Host Blood Vessels Around Human Colorectal Cancers. *Gastroenterology* 110, 1719–1726. doi:10.1053/gast.1996.v110.pm8964396

Richman, S. (2015). Deficient Mismatch Repair: Read All about it (Review). *Int. J. Oncol.* 47, 1189–1202. doi:10.3892/ijo.2015.3119

Rodriguez-Broadbent, H., Law, P. J., Sud, A., Palin, K., Tuupanen, S., Gylfe, A., et al. (2017). Mendelian Randomisation Implicates Hyperlipidaemia as a Risk Factor for Colorectal Cancer. *Int. J. Cancer* 140, 2701–2708. doi:10.1002/ijc.30709

Ru, P., Hu, P., Geng, F., Mo, X., Cheng, C., Yoo, J. Y., et al. (2017). Feedback Loop Regulation of SCAP/SREBP-1 by miR-29 Modulates EGFR Signaling-Driven Glioblastoma Growth. *Cel Rep.* 18, 1076–1077. doi:10.1016/j.celrep.2017.01.016

Sahin, I. H., Akce, M., Alese, O., Shaib, W., Lesinski, G. B., El-Rayes, B., et al. (2019). Immune Checkpoint Inhibitors for the Treatment of MSI-H/MMR-D Colorectal Cancer and a Perspective on Resistance Mechanisms. *Br. J. Cancer* 121, 809–818. doi:10.1038/s41416-019-0599-y

Sánchez-García, I. (2009). The Crossroads of Oncogenesis and Metastasis. *N. Engl. J. Med.* 360, 297–299. doi:10.1056/NEJMcibr0808031

Santos, C. R., and Schulze, A. (2012). Lipid Metabolism in Cancer. *FEBS J.* 279, 2610–2623. doi:10.1111/j.1742-4658.2012.08644.x

Scamuffa, N., Siegfried, G., Bontemps, Y., Ma, L., Basak, A., Cherel, G., et al. (2008). Selective Inhibition of Proprotein Convertases Represses the Metastatic Potential of Human Colorectal Tumor Cells. *J. Clin. Invest.* 118, 352–363. doi:10.1172/jci32040

Shroyer, N. F., Helmrath, M. A., Wang, V. Y. C., Antalffy, B., Henning, S. J., and Zoghbi, H. Y. (2007). Intestine-specific Ablation of Mouse Atonal Homolog 1 (Math1) Reveals a Role in Cellular Homeostasis. *Gastroenterology* 132, 2478–2488. doi:10.1053/j.gastro.2007.03.047

Skotland, T., Hessvik, N. P., Sandvig, K., and Llorente, A. (2019). Exosomal Lipid Composition and the Role of Ether Lipids and Phosphoinositides in Exosome Biology. *J. Lipid Res.* 60, 9–18. doi:10.1194/jlr.R084343

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Sun, L.-N., Zhi, Z., Chen, L.-Y., Zhou, Q., Li, X.-M., Gan, W.-J., et al. (2017). SIRT1 Suppresses Colorectal Cancer Metastasis by Transcriptional Repression of miR-15b-5p. *Cancer Lett.* 409, 104–115. doi:10.1016/j.canlet.2017.09.001

Taga, M., Mouton-Liger, F., Paquet, C., and Hugon, J. (2011). Modulation of Oxidative Stress and Tau Phosphorylation by the mTOR Activator Phosphatidic Acid in SH-Sy5y Cells. *FEBS Lett.* 585, 1801–1806. doi:10.1016/j.febslet.2011.04.022

Taube, J. M., Anders, R. A., Young, G. D., Xu, H., Sharma, R., McMiller, T. L., et al. (2012). Colocalization of Inflammatory Response with B7-H1 Expression in Human Melanocytic Lesions Supports an Adaptive Resistance Mechanism of Immune Escape. *Sci. Transl. Med.* 4, 127ra37. doi:10.1126/scitranslmed.3003689

Tenori, L., Oakman, C., Morris, P. G., Gralka, E., Turner, N., Cappadona, S., et al. (2015). Serum Metabolomic Profiles Evaluated after Surgery May Identify Patients with Oestrogen Receptor Negative Early Breast Cancer at Increased

Risk of Disease Recurrence. Results from a Retrospective Study. *Mol. Oncol.* 9, 128–139. doi:10.1016/j.molonc.2014.07.012

Tian, J.-S., Peng, G.-J., Wu, Y.-F., Zhou, J.-J., Xiang, H., Gao, X.-X., et al. (2016). A GC-MS Urinary Quantitative Metabolomics Analysis in Depressed Patients Treated with TCM Formula of Xiaoyaosan. *J. Chromatogr. B* 1026, 227–235. doi:10.1016/j.jchromb.2015.12.026

Wodlej, C., Riedl, S., Rinner, B., Leber, R., Drechsler, C., Voelker, D. R., et al. (2019). Interaction of Two Antitumor Peptides with Membrane Lipids - Influence of Phosphatidylserine and Cholesterol on Specificity for Melanoma Cells. *PLoS One* 14, e0211187. doi:10.1371/journal.pone.0211187

Yin, P., Zhao, X., Li, Q., Wang, J., Li, J., and Xu, G. (2006). Metabonomics Study of Intestinal Fistulas Based on Ultraperformance Liquid Chromatography Coupled with Q-TOF Mass Spectrometry (UPLC/Q-TOF MS). *J. Proteome Res.* 5, 2135–2143. doi:10.1021/pr060256p

Yoshida, S., Fukumoto, S., Kawaguchi, H., Sato, S., Ueda, R., Furukawa, K., et al. (2001). Ganglioside G(D2) in Small Cell Lung Cancer Cell Lines: Enhancement of Cell Proliferation and Mediation of Apoptosis. *Cancer Res.* 61, 4244–4252.

Young, B. P., Shin, J. J. H., Orij, R., Chao, J. T., Chao, S. C., Guan, X. L., et al. (2010). Phosphatidic Acid Is a pH Biosensor that Links Membrane Biogenesis to Metabolism. *Science* 329, 1085–1088. doi:10.1126/science.1191026

Yu, S., Wang, D., Shao, Y., Zhang, T., Xie, H., Jiang, X., et al. (2019). SP1-induced lncRNA TINCR Overexpression Contributes to Colorectal Cancer Progression by Sponging miR-7-5p. *Aging* 11, 1389–1403. doi:10.18632/aging.101839

Zalloua, P., Kadar, H., Hariri, E., Abi Farraj, L., Brial, F., Hedjazi, L., et al. (2019). Untargeted Mass Spectrometry Lipidomics Identifies Correlation between Serum Sphingomyelins and Plasma Cholesterol. *Lipids Health Dis.* 18, 38. doi:10.1186/s12944-018-0948-5

Zhang, F., Zhang, Y., Zhao, W., Deng, K., Wang, Z., Yang, C., et al. (2017). Metabolomics for Biomarker Discovery in the Diagnosis, Prognosis, Survival and Recurrence of Colorectal Cancer: a Systematic Review. *Oncotarget* 8, 35460–35472. doi:10.18632/oncotarget.16727

Zhao, K., Zhou, H., Zhao, X., Wolff, D. W., Tu, Y., Liu, H., et al. (2012). Phosphatidic Acid Mediates the Targeting of tBid to Induce Lysosomal Membrane Permeabilization and Apoptosis. *J. Lipid Res.* 53, 2102–2114. doi:10.1194/jlr.M027557

Zheng, X., Tsuchiya, K., Okamoto, R., Iwasaki, M., Kano, Y., Sakamoto, N., et al. (2011). Suppression of Hath1 Gene Expression Directly Regulated by Hes1 via Notch Signaling Is Associated with Goblet Cell Depletion in Ulcerative Colitis. *Inflamm. Bowel Dis.* 17, 2251–2260. doi:10.1002/ibd.21611

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Sun, Liu, Chen, Xing and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 15 February 2022
doi: 10.3389/fcell.2022.817965

CD93 Correlates With Immune Infiltration and Impacts Patient Immunotherapy Efficacy: A Pan-Cancer Analysis

Zerui Zhang†, *Mengli Zheng*†, *Qiang Ding and Mei Liu**

Department of Gastroenterology, Institute of Liver and Gastrointestinal Diseases, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Background: The clinical implementation of immune-checkpoint inhibitors (ICIs) targeting CTLA4, PD-1, and PD-L1 has revolutionized the treatment of cancer. However, the majority of patients do not derive clinical benefit. Further development is needed to optimize the approach of ICI therapy. Immunotherapy combined with other forms of treatment is a rising strategy for boosting antitumor responses. CD93 was found to sensitize tumors to immune-checkpoint blocker therapy after the blockade of its pathway. However, its role in immune and ICB therapy across pan-cancer has remained unexplored.

Methods: In this study, we provide a comprehensive investigation of CD93 expression in a pan-cancer manner involving 33 cancer types. We evaluated the association of CD93 expression with prognosis, mismatch repair, tumor mutation burden, and microsatellite instability, immune checkpoints, tumor microenvironment, and immune using multiple online datasets, including The Cancer Genome Atlas, Cancer Cell Line Encyclopedia, Genotype Tissue-Expression, cBioPortal, Tumor Immune Estimation Resource database, and Tumor Immune Single-cell Hub.

Results: CD93 expression varied strongly among cancer types, and increased CD93 gene expression was associated with poor prognosis as well as higher immune factors in most cancer types. Additionally, the level of CD93 was significantly correlated with MMR, TMB, MSI, immune checkpoints, TME, and immune cell infiltration. Noticeably, our results mediated a strong positive contact between CD93 and CAFs, endothelial cells, myeloid dendritic cells, hematopoietic stem cells, mononuclear/macrophage subsets, and neutrophils while a negative correlation with Th1, MDSC, NK, and T-cell follicular helper in almost all cancers. Function analysis on CD93 revealed a link between itself and promoting cancers, inflammation, and angiogenesis.

Conclusion: CD93 can function as a prognostic marker in various malignant tumors and is integral in TME and immune infiltration. Inhibition of the CD93 pathway may be a novel and promising strategy for immunotherapy in human cancer. Further explorations of the mechanisms of CD93 in the immune system may help improve cancer therapy methods.

Keywords: CD93, pan-cancer, tumor immunity, TME, immunotherapy

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Lukman Olalekan Afolabi,
Shenzhen Institutes of Advanced Technology (CAS), China
Yue Zheng,
Tianjin Medical University, China

***Correspondence:**
Mei Liu
fliumei@126.com

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *18 November 2021*
Accepted: *06 January 2022*
Published: *15 February 2022*

Citation:
Zhang Z, Zheng M, Ding Q and Liu M (2022) CD93 Correlates With Immune Infiltration and Impacts Patient Immunotherapy Efficacy: A Pan-Cancer Analysis. Front. Cell Dev. Biol. 10:817965. doi: 10.3389/fcell.2022.817965

INTRODUCTION

Noncommunicable diseases, especially cancer, are responsible for most global deaths. Approximately 18.1 million new cancer cases and 9.6 million deaths occurred in 2018 (Bray et al., 2018). With the decline in the effectiveness of traditional cancer treatments, there is an urgent need to develop precise medicine to acquire better clinical outcomes. In recent years, the use of immune-checkpoint inhibitors (ICIs), such as targeting PD-1/PD-L1 or CTLA4, has markedly changed the field of cancer treatments (Passaro et al., 2020). However, there is still a significant portion of patients who showed minimal response to ICIs (Samstein et al., 2019). The tremendous differences between different patients can largely be explained by the complex tumor microenvironment (TME). The TME reportedly has a significant impact on clinical outcomes and response to therapy (Bagaev et al., 2021). It is of great significance for cancer treatments to focus on TME, explore new tumor immune targets, and develop drugs specifically targeting the TME.

CD93 may provide new opportunities for the immunotherapy of the TME. CD93 is a transmembrane protein consisting of an extracellular domain with a C-type lectin domain, five tandem EGF-like repeats, a serine-threonine-rich mucin-like domain, a transmembrane domain, and a short cytoplasmic domain prominently expressed in endothelial cells and some hematopoietic subsets (Langenkamp et al., 2015). The intracellular domain (ICD) of CD93, a regular transcription factor, is deemed to regulate gene expression with other transcription factors (Cokol et al., 2000). As one of the top 20 core genes for angiogenesis in human primary tumors, CD93 has been widely reported to play a critical role in inflammation and angiogenesis (Masiero et al., 2013; Lugano et al., 2018). Gao et al. (2021) demonstrated that a CD93-enriched subset in hematopoietic stem cells and multipotent progenitors with enhanced stem cell properties and CD93 can act as a regulatory factor in their development, revealing its critical role in hematopoiesis. More excitingly, a recent study found that blockade of the CD93 pathway could reduce hypoxia and promote drug delivery on account of vascular normalization, resulting in the enhanced antitumor. Resultingly, the antitumor efficacy of gemcitabine as well as 5-fluorouracil (5-FU) was greatly enhanced. Of particular interest is the fact that blocking of the CD93 pathway improves tumor vascular functions to promote T-cell infiltration and antitumor immunity, thereby sensitizing tumors to immune-checkpoint blocker (ICB) therapy. For instance, both $CD4^+$ and $CD8^+$ T-cell subsets, as well as the expression of PD-L1, were increased in CD93 mAb–treated tumors, whereas the population of myeloid-derived suppressor cells was reduced (Sun et al., 2021). This study found a safer therapeutic target for tumor vascular normalization and provided a novel remedy for oncotherapy. In addition, there are still many other inspiring studies on CD93. Huang et al. identified CD93 as an interleukin (IL)-17D receptor selectively expressed on group 3 innate lymphoid cells (ILC3s). The ILC3 function was regulated by the IL-17D-CD93 axis at the barrier surfaces, promoting homeostasis and protecting against inflammatory diseases (Huang et al., 2021). Moreover, CD93 was recognized as an important regulator of stemness of leukemia stem cells and a potential therapeutic target and the anti-emetic agent metoclopramide as a drug that blocks CD93 signaling in chronic myeloid leukemia (CML) (Riether et al., 2021). In the proliferating endothelium, the interaction between CD93 and the extracellular matrix regulated cell adhesion, migration, and vascular maturation by activating relevant signaling pathways (Barbera et al., 2021). Overall, all these findings support the central role of CD93 in carcinoma progression and treatment. It is worthwhile to look forward to exploring a brand-new therapeutic target and making greater progress in the research and clinical field via CD93. Thus, comprehensive CD93 analyses based on prognostic, TME, immunological effects, and therapeutic effects across all cancer types are acquired.

Therefore, we employed the most up-to-date data from numerous databases, including TCGA, Cancer Cell Line Encyclopedia (CCLE), Genotype Tissue-Expression (GTEx), cBioPortal, and tumor immune single-cell hub (TISCH), to systematically evaluate the correlation of CD93 with 33 types of cancer. Our results evaluated the connection between CD93 expression and prognosis, tumor mutation burden (TMB), microsatellite instability (MSI), immune checkpoints, TME, immune cell infiltration, and immune-related genes, providing supportive evidence of its vital role in the process of cancer.

MATERIALS AND METHODS

CD93 Expression Analysis

Transcriptome RNA-seq data of CD93 in 33 cancers, which include adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia (LAML), brain lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumors (TGCT), thyroid carcinoma (THCA), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS), and uveal melanoma (UVM), were downloaded from The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov) by UCSC Xena (https://xena.ucsc.edu/). To compare the CD93 expression between tumor and normal tissue, we additionally used the data of the Genotype Tissue Expression database (GTEx; https://commonfund.nih.gov/

GTEx). The Tumor Immune Single-cell Hub database was used to examine the biodiversity of the tumor cell community and find heterogeneity in the TME.

Survival Analysis and ROC Analysis

Four outcome parameters downloaded from TCGA, including overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI), were incorporated in our study. To compare the overall survival (OS) rate for patients based on CD93 expression, we used Kaplan–Meier survival analysis using the log-rank test ($p < 0.05$). The survival and survminer R packages were used to draw survival curves. Univariate Cox analysis was used to evaluate the prognostic values across the 33 studied types of cancer ($p < 0.05$). The survival results were displayed with HR, 95% CI, and log-rank p values. We used the rms R package to construct baseline nomograms in KICH, LIHC, and UVM patients for clinical application and plotted receiver operating characteristic curves (ROC curves) in several types of cancer.

Analysis of MMR, TMB, and MSI in Cancers

We surveyed the correlation between CD93 expression and several essential mismatch repair (MMR) genes, including the MutL homologous gene (MLH1) and the MutS homologous gene (MSH2, MSH6), and increased separation after meiosis (PMS2) and epithelial cell adhesion molecule (EPCAM) through TCGA. TMB is the total number of nonsynonymous mutations occurring in each coding region in the tumor genome (Chalmers et al., 2017). The higher the TMB is, the better clinical benefits obtained from immunotherapy (Meléndez et al., 2018; Marabelle et al., 2020). MSI, the condition of genetic hypermutability caused by the inactivation of mismatch repair genes, is another biomarker for immunotherapy (Ward et al., 2001; Boussios et al., 2019). The data were downloaded from UCSC Xena, and the correlation between CD93 expression and TMB and MSI was analyzed via Spearman's rank correlation coefficient.

TME and Immune Infiltration

The relationship of CD93 expression to TME was studied through evaluating the ratio of stromal and immune cells in tumors by estimate R packages, and the quantified results were displayed by tumor purity, stromal score, immune score, and estimate score. Specifically, stromal score and immune score stand for the stromal cells and the immunocyte infiltration level in tumor tissues, respectively. Estimate score was the sum of the preceding two, which infers the tumor purity. The estimated statistical significance and Spearman's correlation coefficient were generated through correlation analysis. The Tumor Immune Estimation Resource (TIMER) database (https://cistrome.shinyapps.io/timer/) and the TIMER2.0 database (http://timer.cistrome.org/), which provided an integration of immune cells, including B cells, $CD4^+$ T cells, $CD8^+$ T cells, macrophages, neutrophils, and dendritic cells, for RNA sequencing samples from TCGA, were used to evaluate the abundance of tumor-infiltrating immune cells (TIICs) across diverse cancer types (Li et al., 2017). We next explored the relationship between CD93 and tumor purity as well as the abundance of TIICs by it. Finally, we explored the association between CD93 expression and immune-related genes, which includes HLA, immunostimulatory genes, immunosuppressive genes, chemokine, and chemokine receptor proteins, in 33 types of cancer.

Enrichment Analysis

To explore the biological functions of CD93 across the pan-cancer, gene set enrichment analysis (GSEA) was performed between the low-CD93 and high-CD93 groups. Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets and HALLMARK gene sets were downloaded from the official GSEA website (https://www.gsea-msigdb.org/gsea/downloads.jsp). The enriched gene sets were selected based on $p < 0.05$, net enrichment score (NES) of greater than 1, and a false discovery rate (FDR) < 0.25.

Statistical Analysis

All the data of gene expression were normalized by $\log_2$ transformation. The correlation analysis between the two variables used Spearman's or Pearson's test; $p < 0.05$ was considered significant. Student's t-test and Kruskal–Wallis test were used for comparisons between two groups and for comparisons among >2 groups, respectively. All statistical analyses were processed by the R software (Version 4.0.2).

RESULTS

Expression Analysis of CD93

To explore the basic expression of CD93, we first analyzed the mRNA expression of CD93 in 31 types of normal tissues through the Genotype-Tissue Expression (GTEx) portal (Kruskal–Wallis test, $p < 0.001$) (**Figure 1A**). The data showed that the expression of CD93 was the highest in lung and breast tissues, while it was the lowest in the bone marrow. However, it was low in most normal cell lines based on the data of the Cancer Cell Line Encyclopedia (CCLE) database (Kruskal–Wallis test, $p < 0.001$) (**Figure 1B**). Next, we evaluated its expression in TCGA pan-cancer and found it highly expressed in KIRC but low in LGG and LIHC (**Figure 1C**). Notably, there were obvious differences in CD93 expression between some tumor tissues and corresponding adjacent noncancerous tissues. Increased expression was found in CHOL, GBM, KIRC, LGG, LIHC, STAD, and THCA while decreased expression in BLCA, BRCA, KICH, KIRP, LUAD, LUSC, and UCEC. To fully analyze the differential expression of CD93 between normal and tumor samples as compelling a manner as possible, we integrated the data of the TCGA and GTEx databases and found that they were still differentially expressed in many cancers, most of which were consistent with the previous analysis (**Figure 1D**). More precisely, CD93 expression was significantly increased in ACC CHOL, COAD, ESCA, GBM, KIRC, LAML, LGG, LIHC, OV, PAAD, PRAD, SKCM, STAD, TGCT, and THCA, whereas it was downregulated in BLCA,

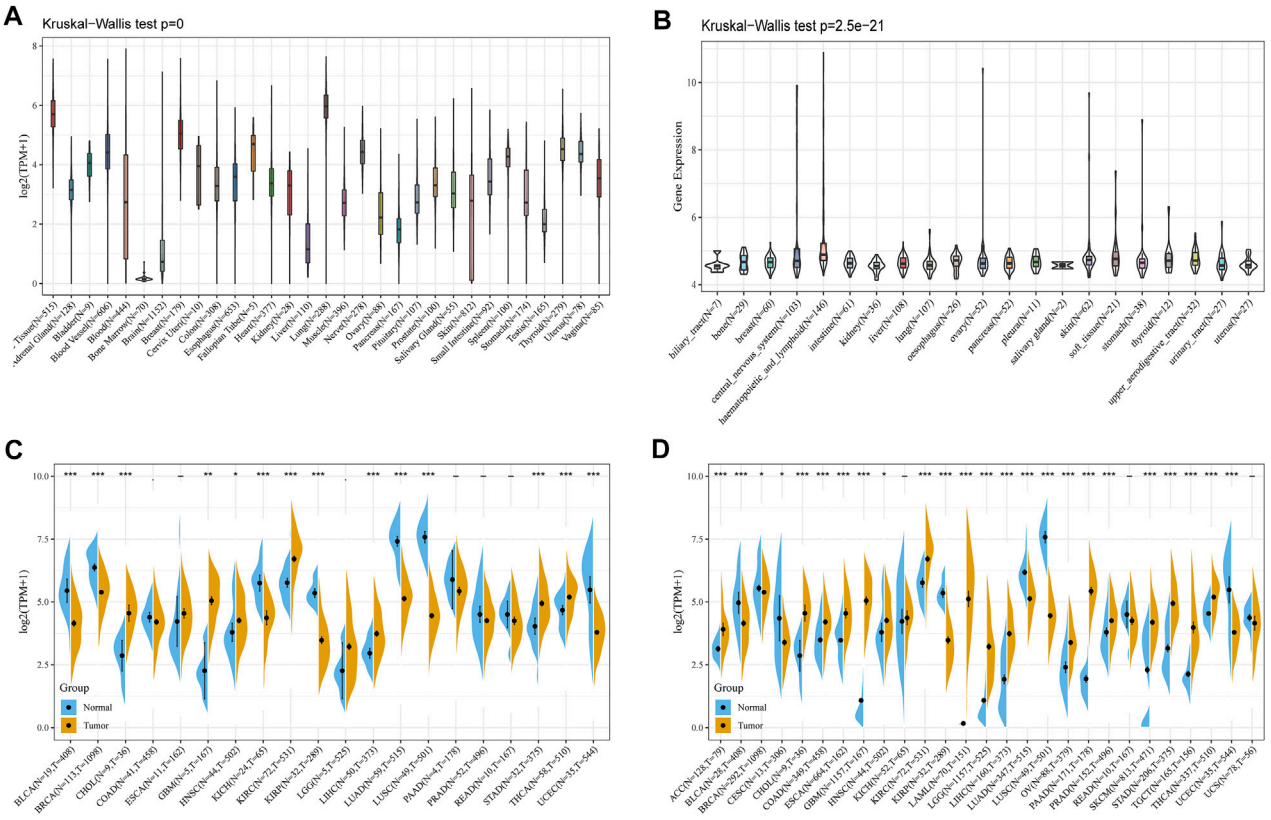


FIGURE 1 | Expression analysis of CD93. **(A)** CD93 expression in various normal tissues. **(B)** CD93 expression in different tumor cell lines. **(C)** CD93 expression in the TCGA database. **(D)** Comparison of CD93 expressions between tumors and normal samples in the TCGA and the GTEx databases. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

KIRP, LUAD, LUSC, and UCEC. Now we have established that CD93 is expressed differently in different cancers.

Prognostic and Diagnostic Value Analysis of CD93 in Pan-Cancer

Now that CD93 is known to be differentially expressed in certain cancers, we explored whether its expression is associated with the prognosis of different cancers. We then performed a survival association analysis for each cancer. The outcome parameters assessed were OS, DSS, PFI, and DFI. The analyses of OS by Cox regression indicated that CD93 acted as a general risk factor in BLCA ($p = 0.0035$), KIRP ($p = 0.0087$), LGG ($p = 0.000017$), LUSC ($p = 0.041$), OV ($p = 0.0049$), STAD ($p = 0.019$), and UVM ($p = 0.0087$) (**Figure 2A**). Interestingly, we obtained similar result and conclusion when we used the Kaplan–Meier survival analysis, revealing that CD93 was a high-risk factor in BLCA ($p = 0.013$), KIRP ($p = 0.002$), LGG ($p < 0.001$), LUSC ($p = 0.023$), OV ($p = 0.013$), STAD ($p = 0.002$), and UVM ($p = 0.004$) (**Figures 2B–H**). However, in this way, CD93 was a protective factor in KIRC ($p < 0.001$) (**Figure 2I**). Next, a correlation analysis was applied between CD93 expression and DSS to avoid the bias from people who did not die of cancer. Cox regression analysis showed that low CD93 expression levels were associated with poorer DSS in KIRP ($p = 0.00078$), LGG ($p < 0.0001$), OV ($p = 0.028$), and UVM ($p = 0.0043$), while they were associated with better DSS in KIRC ($p < 0.0001$) (**Figure 3A**). Similarly, the Kaplan–Meier analysis revealed worse prognostic impacts in KIRP ($p < 0.001$), LGG ($p < 0.001$), OV ($p = 0.035$), and UVM ($p = 0.001$) (**Figures 3B–E**), while better in KIRC ($p < 0.001$) (**Figure 3F**). The following results are about the expression of CD93 and PFI. Forest plots demonstrated that high levels of CD93 expression were closely related to adverse outcomes in patients with ACC ($p = 0.027$), KIRP ($p = 0.018$), LGG ($p < 0.0001$), and UVM ($p < 0.0001$), whereas that meant a contrary result in KIRC ($p = 0.0084$) and THCA ($p = 0.033$) (**Figure 4A**). In accordance with the Cox regression model, KM survival analysis showed that elevated CD93 expression was significantly related to a poorer PFI in KIRP ($p < 0.001$), LGG ($p < 0.001$), and UVM ($p = 0.001$) (**Figures 4B–D**), while it was significantly related to a better PFI in KIRC ($p = 0.002$) and THCA ($p = 0.003$) (**Figures 4E,F**). There was no obvious correlation between CD93 and DFI but UCEC (detailed information in **Supplementary Figure S1**).

To better predict the prognosis of KIRC, LIHC, and UVM patients in a clinic, a prognostic nomogram was developed by integrating gender, tumor stage, and CD93 expression into a univariate Cox regression model. A score based on the

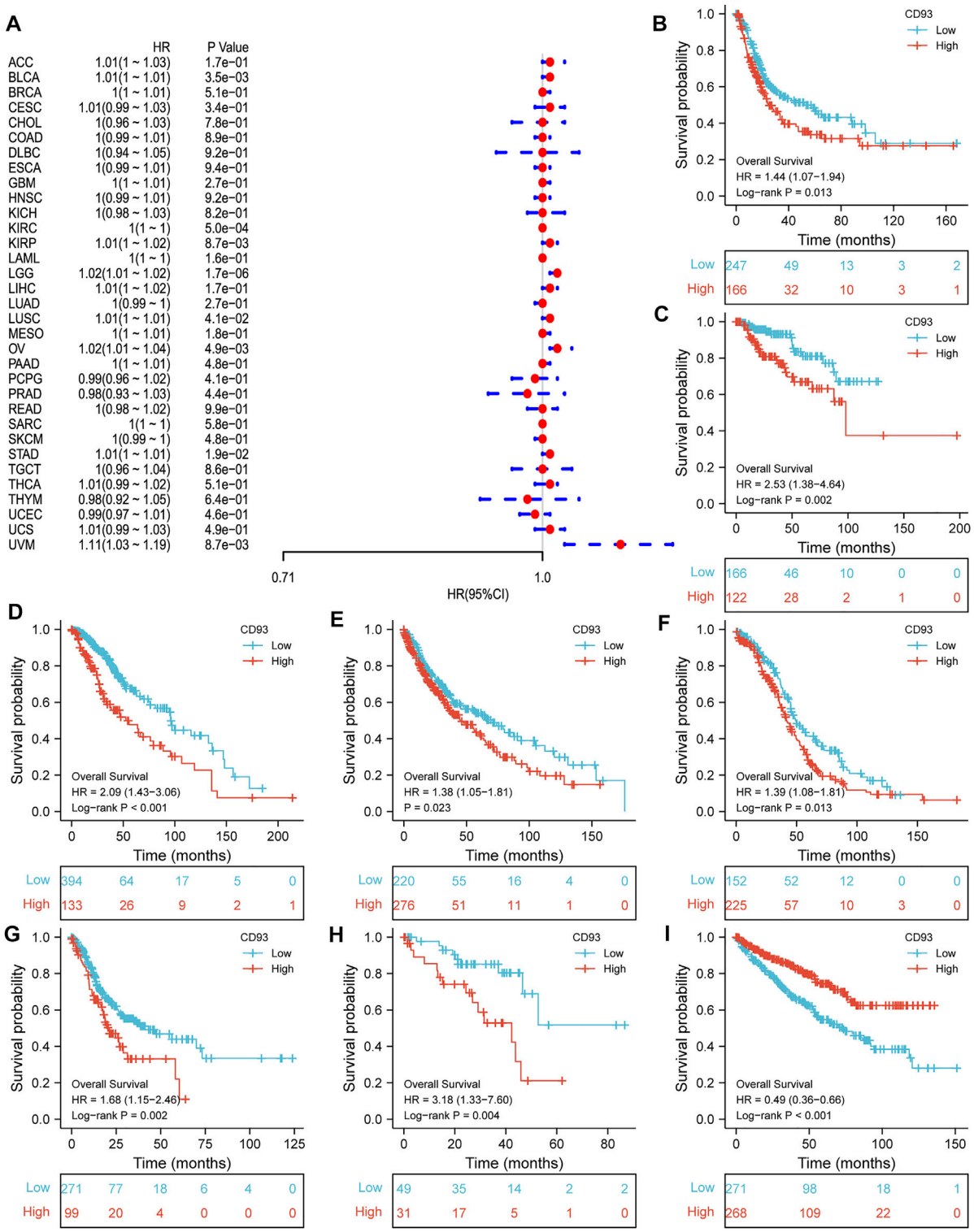


FIGURE 2 | Association between CD93 and OS. **(A)** Forest plot of OS associations in 33 types of tumors. **(B–I)** Kaplan–Meier analysis of the association between CD93 and OS in BLCA, KIRP, LGG, LUSC, OV, STAD, UVM, and KIRC.

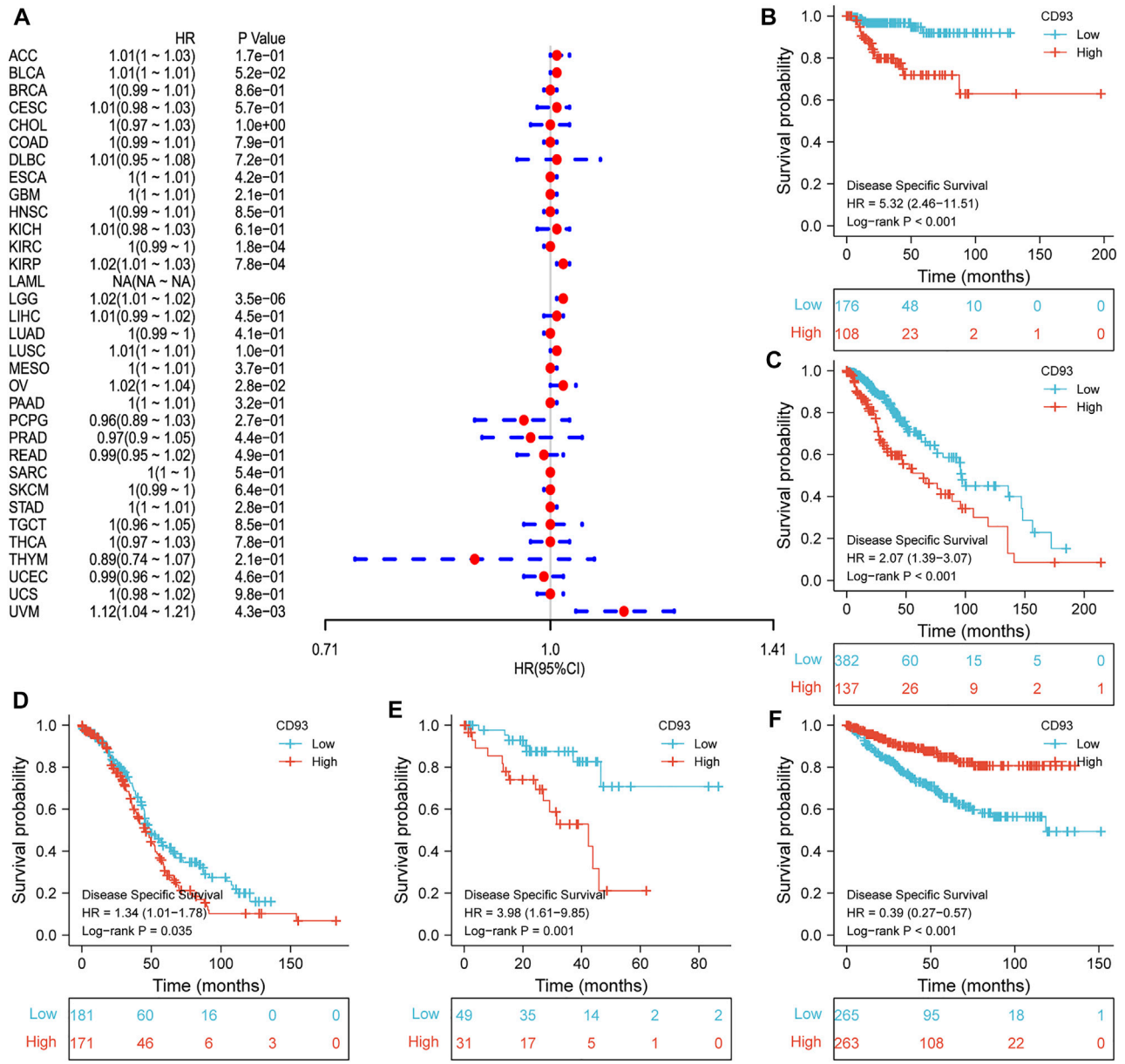


FIGURE 3 | Association between CD93 and DSS. **(A)** Forest plot of DSS associations in 33 types of tumors. **(B–F)** Kaplan–Meier analysis of the association between the CD93 expression and DSS in KIRP, LGG, OV, UVM, and KIRC.

nomogram was calculated to predict the 1-, 3-, 5-, and 7-year OS for individual patients. The calibration plot showed that the nomogram performed well in predicting the OS of patients according to an ideal model. Additionally, AUC plotted for different durations of OS showed that CD93 expression was good at predicting the OS in KICH, LIHC, and UVM (**Supplementary Figures S2–S4**). We also demonstrated that CD93 can be a potential diagnostic biomarker in serval types of cancer via ROC curves (**Supplementary Figure S5**).

All of these data implied that the expression of CD93 was significantly associated with patient prognosis in multiple cancer types, especially in patients with KIRP, LGG, UVM, and KIRC.

Correlations Between CD93 Expression and MMR, TMB, and MSI in Cancers

Microsatellites (MSs), which are simple repetitive sequences of nucleotide bases, can generate errors during DNA replication, and MMR genes can recognize and fix this process. Tumors with defects in the MMR system are vulnerable to mutations in microsatellites, leading to high levels of MSI, leading in turn to the accumulation of mutation loads in cancer-related genes

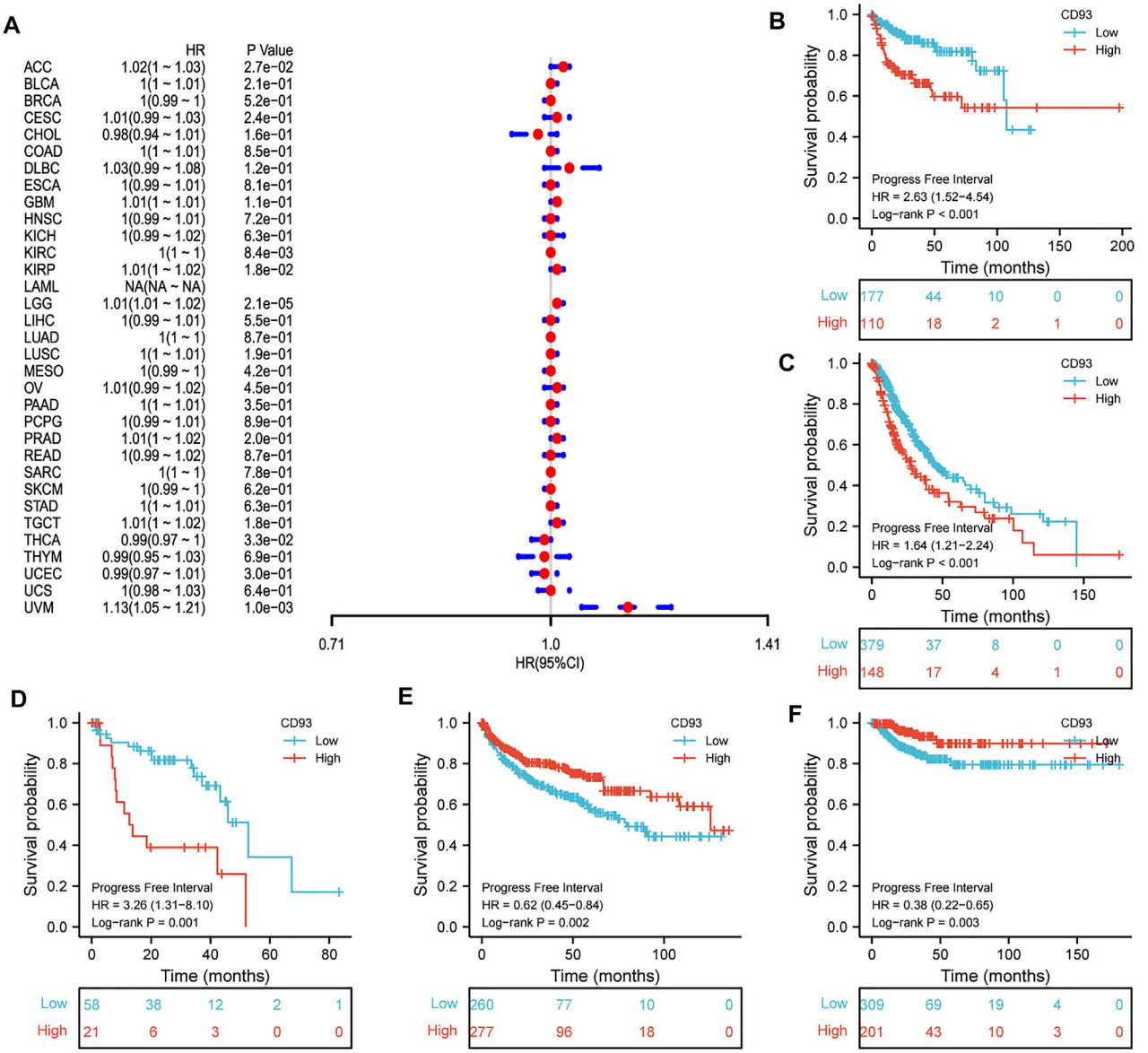


FIGURE 4 | Association between CD93 expression and PFI. **(A)** Forest plot of PFI associations in 33 types of tumors. **(B–F)** Kaplan–Meier analysis of the association between the CD93 expression and PFI in KIRP, LGG, UVM, KIRC, and THCA.

and the aggravation of TMB (Yarchoan et al., 2017; Joshi and Badgwell, 2021). We surveyed the correlation between CD93 expression and several essential MMR genes. CD93 expression was closely related to MMR genes in almost all 33 cancers but ACC, CESC, CHOL, ESCA, GBM, MESO, PCPG, SARC, and UCS (**Figure 5A**). Given that TMB status is gradually regarded as one of the promising pan-cancer biomarkers for forecasting ICB therapeutic effect, which has already been approved by the FDA (Chan et al., 2019), we checked the correlation between CD93 expression and TMB across cancers. As shown in **Figure 5B**, CD93 expression was positively related to TMB in LGG and THYM while negatively related to TMB in BRCA, CESC, HNSC, KIRP, LIHC, LUAD, LUSC, PAAD, PRAD, TAD, THCA, and UVMP. As another biomarker associated with ICI response, the association of MSI with CD93 expression was evaluated in our study (**Figure 5C**). Based on our analysis, the expression of CD93 had a positive relation with MSI in TGCT and COAD, while it indicated a negative relation in LUSC, PAAD, SKCM, STAD, THCA, BRCA, DLBC, and HNSC.

CD93 and TME

As we all know, TME plays a critical role in regulating malignant progression and modulating reactions to therapies, so strategies to therapeutically target TME have emerged as a promising approach for cancer treatment in recent years (Bejarano et al., 2021). To assess the association of CD93 expression and TME, we

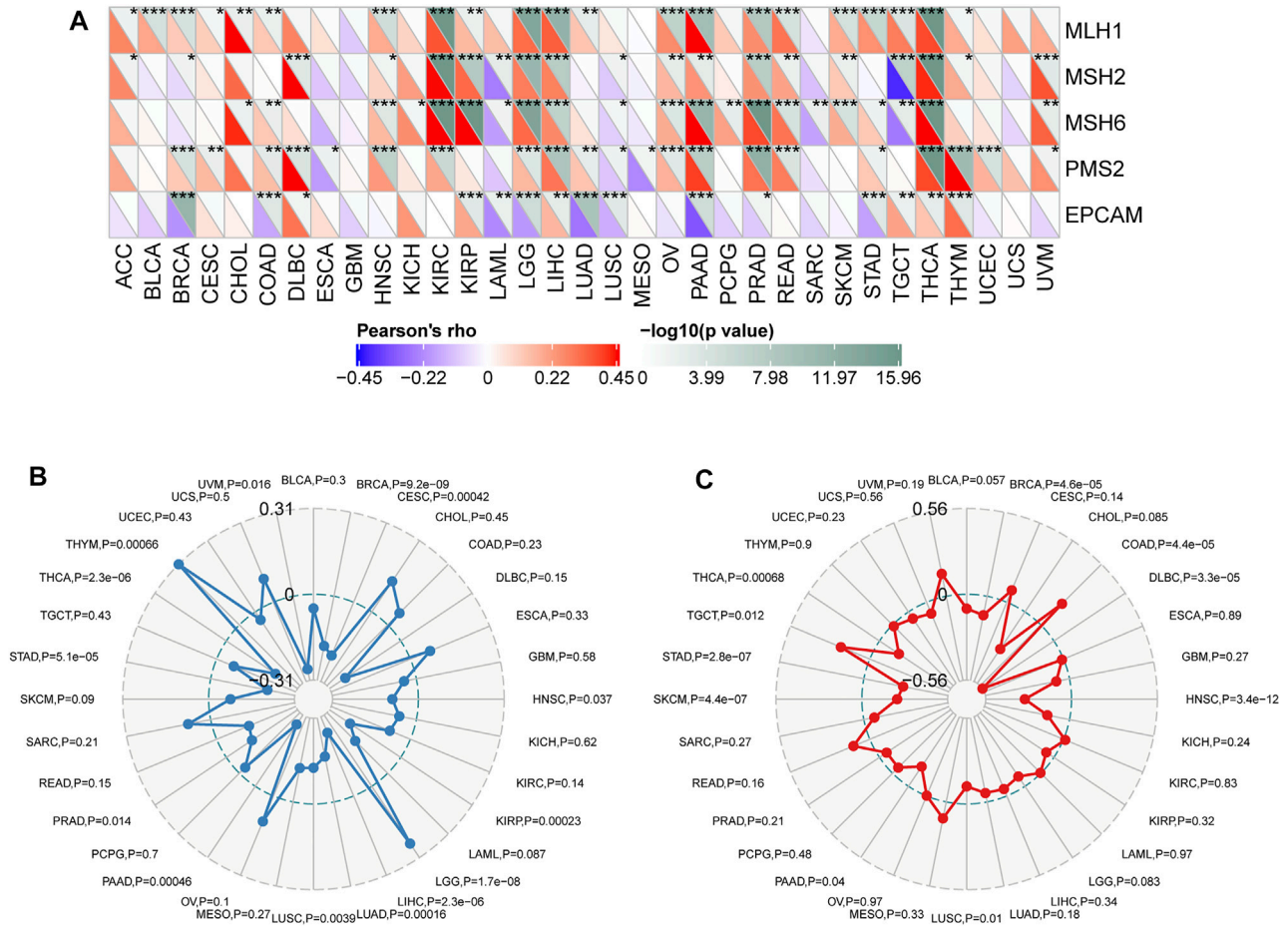


FIGURE 5 | Correlations between CD93 expression and MMR, TMB, and MSI in cancers. **(A)** Correlations between CD93 expression and MMR genes (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). **(B)** Correlations between CD93 expression and TMB across cancers. **(C)** Correlations between CD93 and MSI across cancers. The value in black reveals the range, and the curves in blue and red reveal the correlation coefficients.

calculated the stromal, immune scores, ESTIMATE scores, and tumor purity in 33 types of cancer (**Figure 6A**). As shown in this study, the expression of CD93 is positively connected with stromal scores, immune scores, as well as ESTIMATE scores in the overwhelming majority of cancers but TGCT, THCA, THYM, and UCEC. However, our analysis indicated a negative contact between CD93 and tumor purity in most cancers. Specifically, CD93 was significantly associated with StromalScores and EstimateScores in COAD, BLCA, HNSC, KICH, LUSC, MESO, PAAD, READ, and USC and with immune scores in BLCA, COAD, KICH, LAML, LUSC, PAAD, READ, and UVM. At the same time, CD93 expression was closely related to tumor purity in ACC, BLCA, COAD, PAAD, and READ. We then visualized the top three cancers in detail among these analyses (**Figures 6B–J**). Additionally, we found varying degrees of heterogeneity between diverse landscapes of the TME by the Tumor Immune Single-cell Hub database. As shown in **Supplementary Figure S6**, we found that CD93 was extremely enriched in endothelial cells and mononuclear/macrophage subsets among all of these cancers. The expression of CD93 is concentrated malignantly in STAD, DC in STAD, and SKCM.

The Relevance of CD93 Expression and Immune Infiltrates in Pan-Cancer

As one of the crucial parts of TME, immune cells make a great contribution to the homeostasis and evolution of TME (Vitale et al., 2019). To explore the correlation between CD93 expression and immune infiltration in human pan-cancer, we mainly focused on the association between CD93 and infiltrating immune cells by the TIMER2.0 database in various cancers (**Figure 7**). Notably, our results demonstrated that it had a strong positive contact between CD93 and CAFs, endothelial cells, myeloid dendritic cells, hematopoietic stem cells, mononuclear/macrophage subsets, and neutrophils while it had a negative correlation with Th1, MDSC, NK, and T follicular helper cells in almost all cancers we analyzed. Noticeably, CD93 was essentially enriched in CAFs, endothelial cells, and hematopoietic stem cells among these

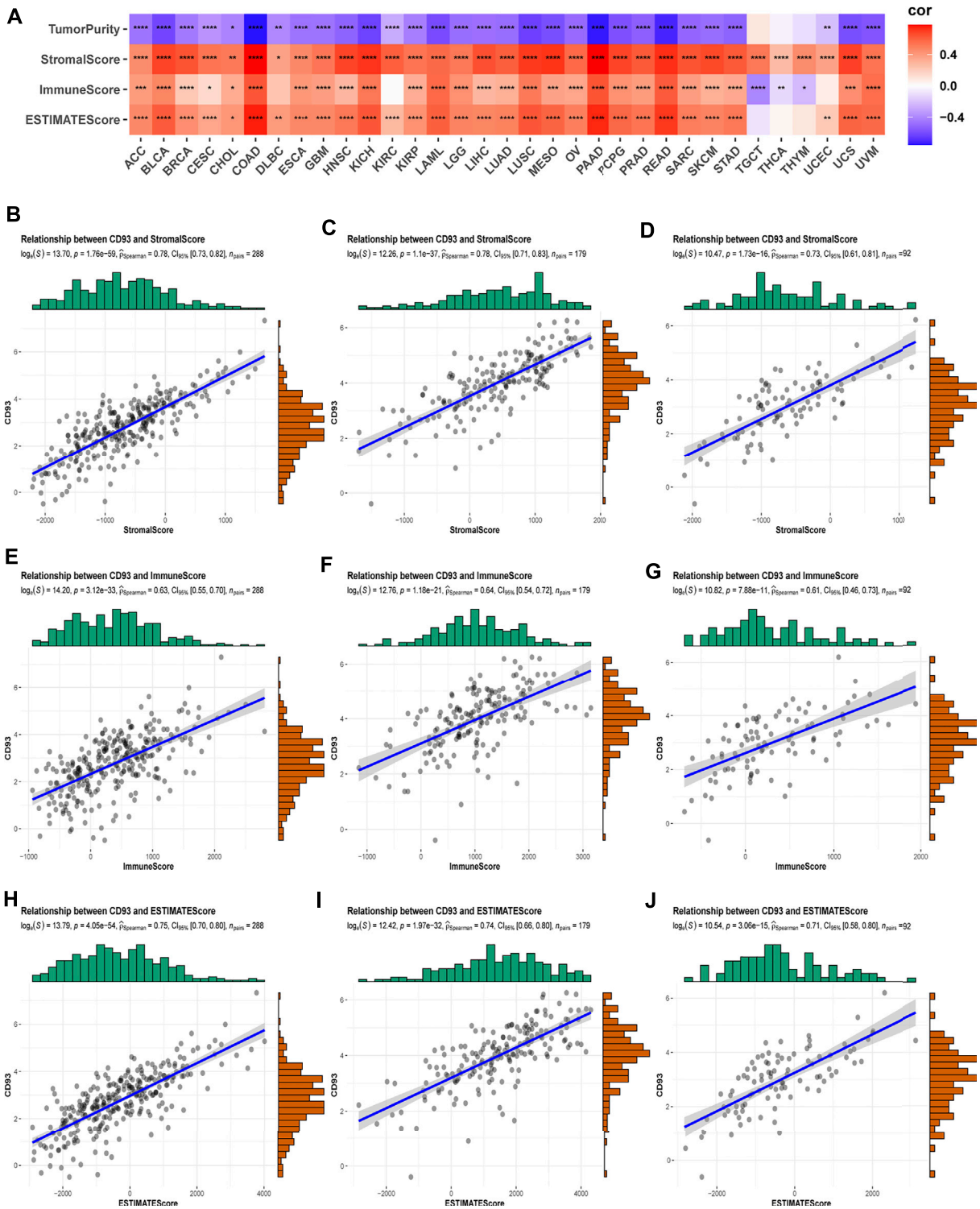


FIGURE 6 | Correlation of CD93 expression with ImmuneScore and StromalScore in various cancers. **(A)** The Spearman's correlation analysis of CD93 expression with ImmuneScore, StromalScores, and EstimateScores in 33 cancers (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$). Correlation of CD93 expression with three scores in COAD **(B–D)**, PAAD **(E–G)**, and READ **(H–J)**.

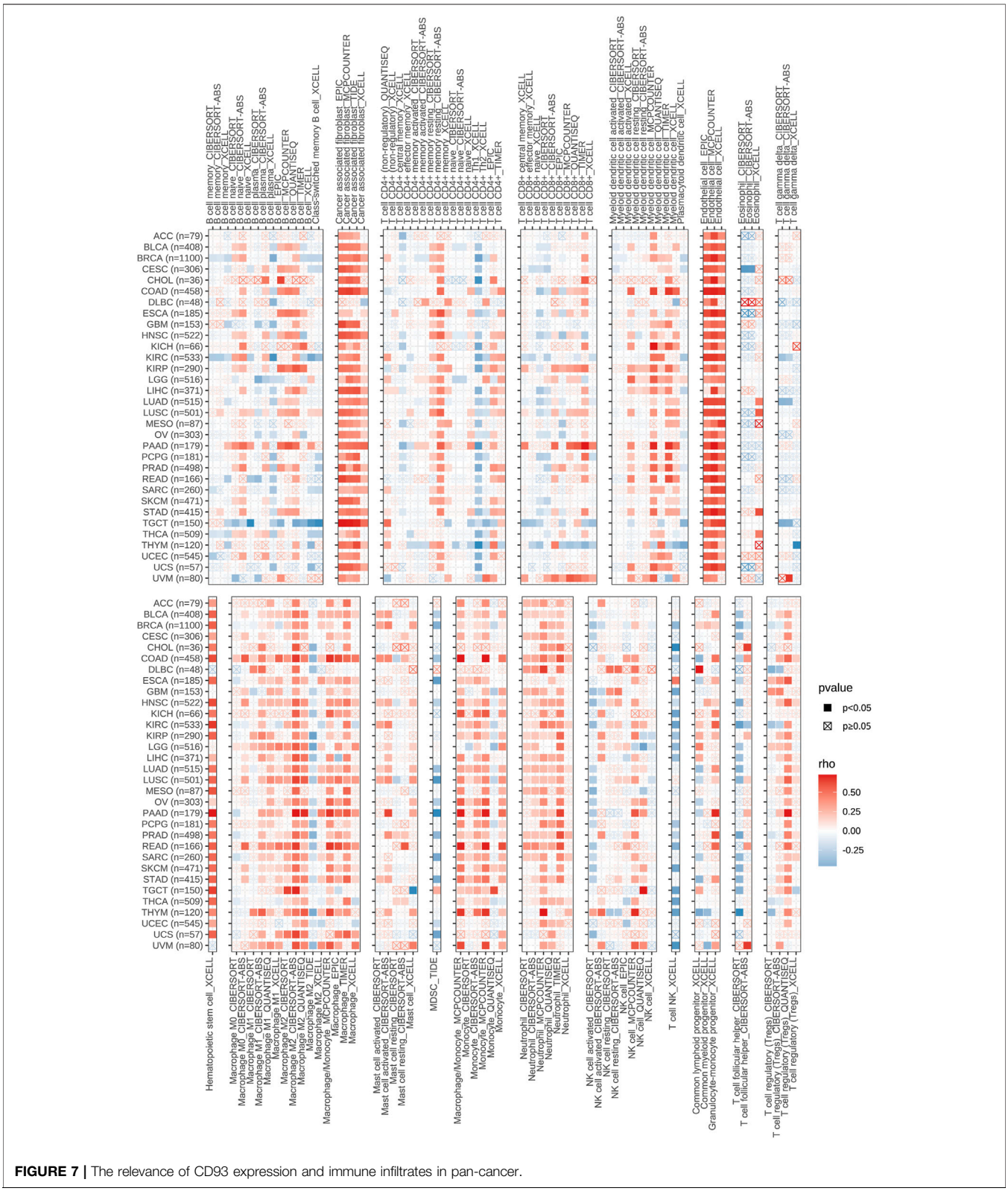


FIGURE 7 | The relevance of CD93 expression and immune infiltrates in pan-cancer.

cancers. In addition, our data showed that increased CD93 expression was related to increased B cell infiltration in PAAD, KIRP, and ESCA, while it was related to decreased B cell infiltration in TGCT. CD93 expression was positively correlated with $CD8^+$ T cells in UVM, PAAD, and KIRP while negatively in THYM. Concurrently, CD93 is enriched in

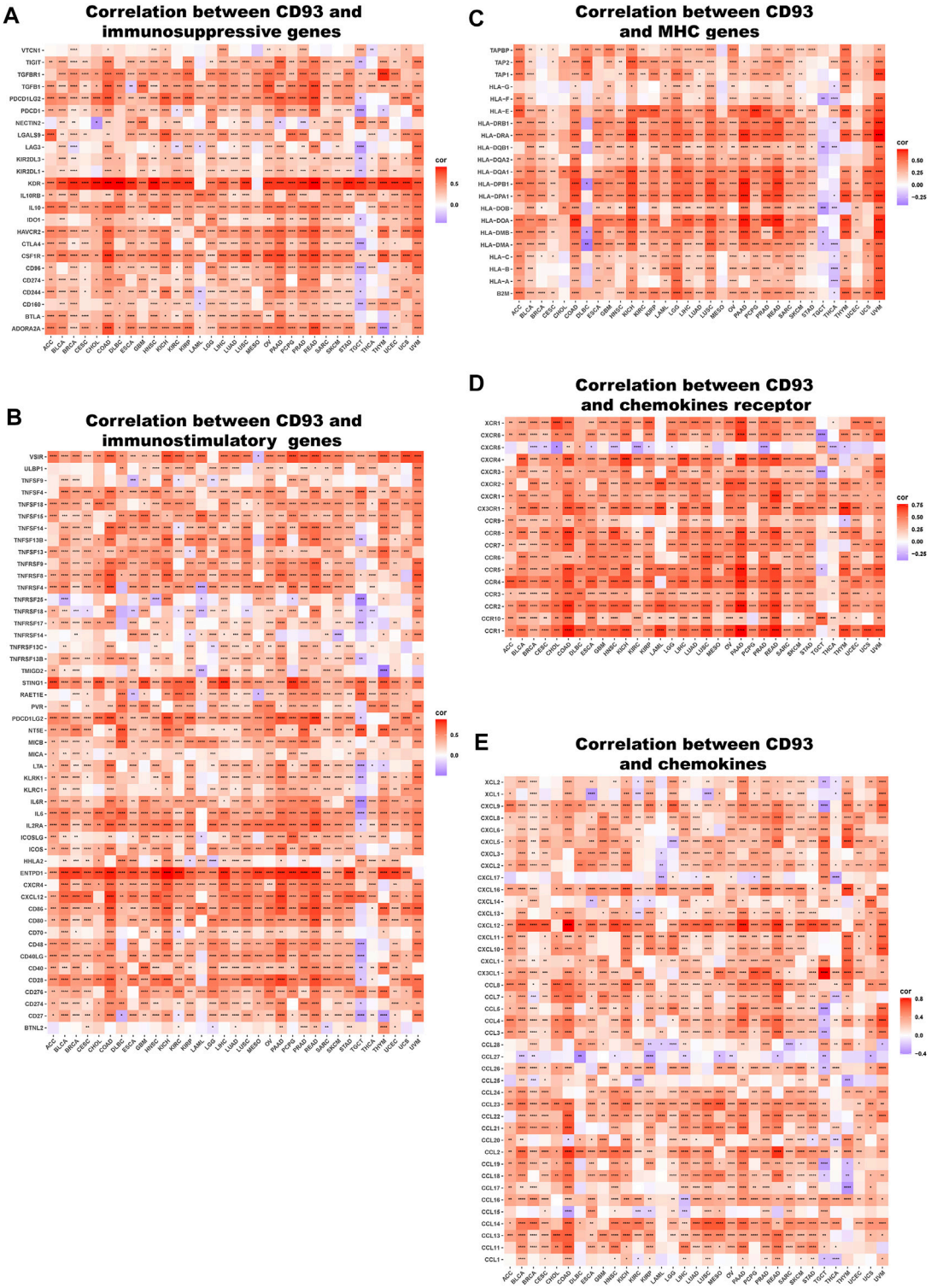


FIGURE 8 | Relationship between CD93 and **(A)** immunosuppressive factors, **(B)** immunostimulatory factors, **(C)** MHC genes, **(D)** chemokine receptors, **(E)** and chemokines. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

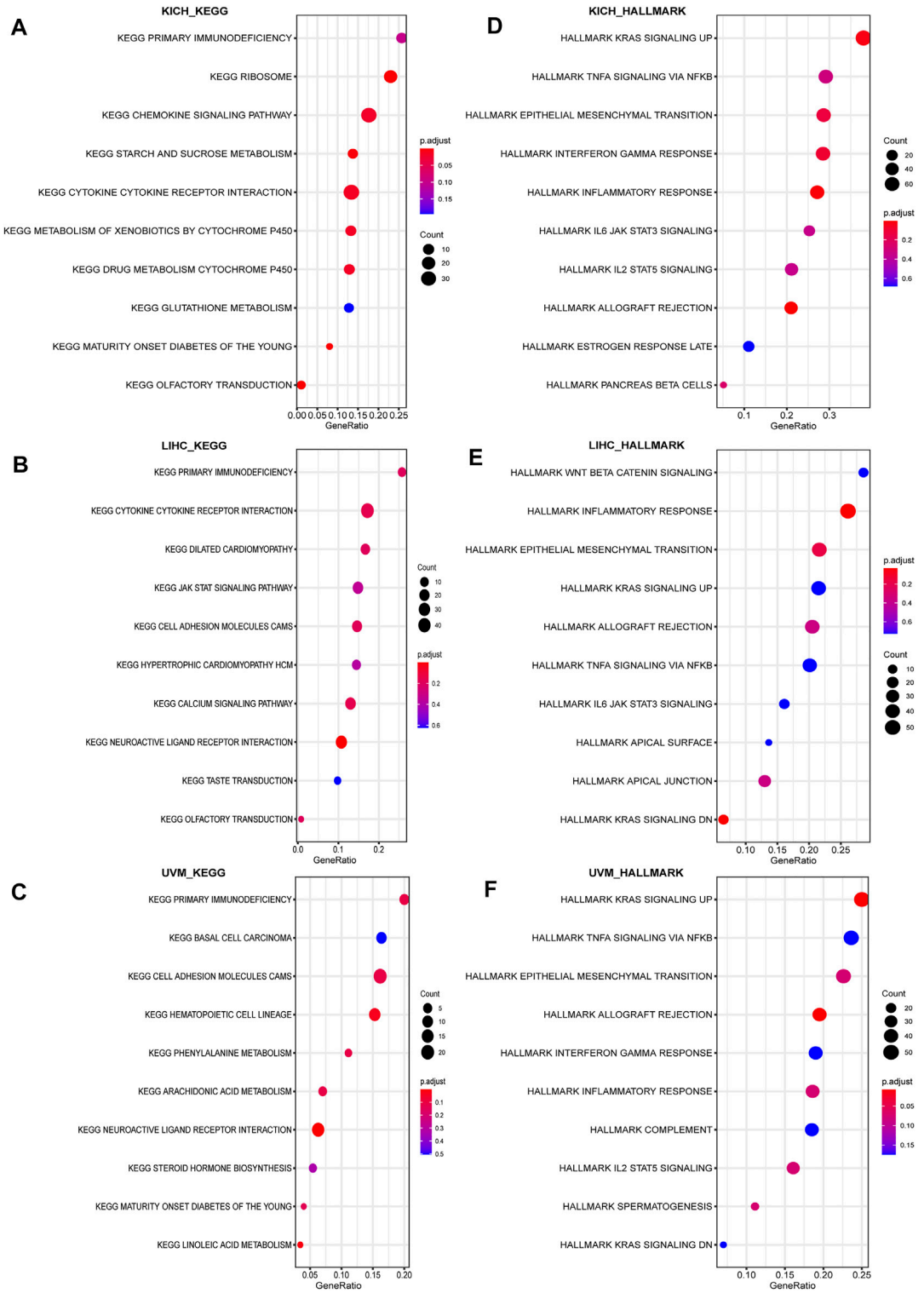


FIGURE 9 | Result of GSEA. Different KEGG and HALLMARK pathways regulated in KICH **(A–B)**, LIHC **(B–D)**, and UVM **(E–F)**.

neutrophils in THYM and DLBC, and mast cells in READ and PAAD. Compared with M0-M1-like macrophages, there was a stronger relationship between CD93 expression and M2-like macrophages, especially in UVM, UCS, TGCT, READ, PAAD, and COAD. We also analyzed the correlations between CD93 and B cells, CD8 + T cells, CD4 + T cells, macrophages, neutrophils,

and dendritic cells in 32 types of cancer via the TIMER database. As shown in **Supplementary Table S1**, CD93 expression was mainly associated with six infiltrating immune cells in ACC, CHOL, COAD, HNSC, KIRP, LGG, LIHC, LUAD, LUSC, PAAD, PRAD, READ, SKCM, THCA, and UCEC. Among these cancers, CD93 was associated with six infiltrating immune cells of all cancers.

Relationship Between CD93 and Immune Factors

We next explored the association between CD93 expression and immune-related genes, which includes HLA, immunostimulatory genes, immunosuppressive genes, chemokines, and chemokine receptor proteins, in 33 types of cancer. Our data showed that CD93 expression was positively correlated with almost all immune-related genes in nearly all cancers. As previously mentioned, elevated CD93 expression was related to increased immunosuppressive factors in all types of cancer except TGCT (**Figure 8A**). Noticeably, across the cancers we analyzed, the correlation was strongest between CD93 and KDR, and TGFBR1, TGFB1, PDCD1LG2, IL10, HAVCR2, CSF1R, and ADORA2A came second. Similar to immunosuppressive factors, CD93 expression was positively associated with immunostimulatory genes in all cancers we analyzed but TGCT (**Figure 8B**). In all analyzed cancers, ENTPD had the strongest correlation with CD93. Unlike other cancers, there was a more negative association between CD93 and immunostimulatory genes in TGCT. As shown by the heatmap presented in **Figure 8C**, there was a strong correlation between CD93 expression and almost all HLA genes in most cancers except for CHOL, DLBC, TGCT, and UCEC. Some of HLA genes, such as HLA-DMA, HLA-DOB, HLA-DQB1, and HLA-F, showed a negative correlation with CD93 in DLBC, TGCT, and THCA. In addition, we also analyzed the association between CD93 and chemokines as well as chemokine receptors (**Figures 8D,E**). The results showed that CD93 expression was positively correlated with chemokines and chemokine receptors, especially CXCL12 and CCR4.

Functional Analysis of CD93

Lastly, GSEA was used to study the biological role of CD93 in different tumor tissues. Our data indicated that the top three positively enriched KEGG, from a pan-cancer perspective, in the elevated expression of CD93 were cytokine–cytokine/immunodeficiency. The HALLMARK enrichment term showed that CD93 expression was positively associated with the process of promoting cancers and inflammation, including interferon gamma response, inflammatory response, TNF signaling via NF-κB, complements, angiogenesis, EMT, IL-2-STAT5 signaling, kras signaling up, myc targets V2, and hedgehog signaling, in almost all cancers we analyzed. Here we showed the results in KICH (**Figures 9A,B**), LIHC (**Figures 9C,D**), and UVM (**Figures 9E,F**). To further validate the link between CD93 expression and other cancer-related biological functions in pan-cancer, we analyzed additional two biological roles of CD93 in various human tumors including autophagy, EMT, TGFB1, and Wnt pathway (**Supplementary Figure S7**). Interestingly, there was a positive correlation between CD93 and associated genes.

DISCUSSION

As a marker of tumor vasculature, the Group XIV family member endosialin was shown to regulate tumor growth and CD93 is homologous in structure to it (Nanda et al., 2006; Maia et al., 2011; Greenlee-Wacker et al., 2012). This suggests that CD93 may play key roles during cancer. However, despite the recent attention CD93 has received, its role in pan-cancer has remained unexplored prior to this investigation. In this study, we demonstrate that the differential expressions of CD93 in cancers help with prognosis estimation, the exploration and research of its mechanism, prediction of therapeutic effects, and the development of new treatments. It is conceivable to expect that the exploration of CD93 may therefore lead to major breakthroughs in the treatment of cancer, bringing great hope to those patients who have troubles with cancer diseases.

We first evaluated the expression levels and prognostic significance of CD93 in 33 types of cancers to explore its role across cancers and whether it can serve as a prognostic biomarker, which is a meaningful study to facilitate the translation of basic science to clinical studies. The results showed that elevated CD93 expression was found in CHOL, LGG, GBM, LIHC, and KIRC while low expression in KIRP, LUAD, and LUSC. Notably, the overexpression of CD93 was correlated with a worse prognosis in multiple cancers, such as LGG, OV, KIRP, and LUSC but KIRC. The expression of CD93 in tumors seems to make varying contributions among different types of cancer. Supporting that, CD93 overexpression was found in tumor vasculatures, and it influenced the survival of the host in PDAC, PNET, melanoma, and colon cancer (Langenkamp et al., 2015; Olsen et al., 2015; Sun et al., 2021). High CD93 levels were also found correlated with poor survival and increased proliferation capacity in leukemia (Pessoa Rodrigues and Akhtar, 2021). These findings clearly demonstrated the supporting role of CD93 in the prognosis of various cancers.

As a vital factor associated with the stability and integrity of genome, MMR has drawn considerable attention (Fishel, 2015; Russo et al., 2019). In addition, TMB and MSI are novel sensitive predictors of immunotherapy (Boland and Goel, 2010; Yarchoan et al., 2017; Lee et al., 2019). In our study, CD93 expression was enormously related to MMR genes in almost all cancers we analyzed. Additionally, CD93 also correlated with TMB and MSI in certain types of cancer. Then, the associations of CD93 expression with multiple checkpoint markers were compared across different cancer types. As expected, our data showed that CD93 was highly positively associated with almost all checkpoint genes in most cancers, which significantly indicated that CD93 may play an essential role in tumor immunity and may be the next promising marker in immunotherapy. Sun et al. observed that PD-L1 expression elevated in tumor tissues once anti-CD93 treatment and the growth of tumor were suppressed via the combination of CD93 and PD1 mAbs. Their results proved that blockade of the CD93 pathway can sensitize tumors to ICB therapy and facilitate cancer immunotherapy (Sun et al., 2021). This further demonstrated the importance of CD93 in immunotherapy.

Tumor cells exist in TME, which plays a central role in tumor initiation, development, metastasis, and clinical treatment and influences angiogenesis, tumor development, and immune escape

(Lyssiotis and Kimmelman, 2017; Han et al., 2021; Hiam-Galvez et al., 2021). Our data showed that the expression of CD93 was positively connected with StromalScores, ImmuneScores, as well as ESTIMATE scores in human pan-cancer. Sun et al. (2021) also found that CD93 is selectively upregulated in tumor vasculature due to VEGF exposure in the TME, and its antibody can suppress tumor growth in mice, which hinted at the important role of CD93 in TME

However, despite its significant role in TME, the role of CD93 in immune infiltration needs to be further elucidated. Therefore, we further investigated on the role of CD93 in immune-cell infiltration levels in 33 types of cancer. Previous studies have reported that CD93 is expressed on a variety of cells such as neutrophils and endothelial cells (Greenlee-Wacker et al., 2011; Huang et al., 2021). In line with these findings, we showed that CD93 was enriched in CAFs, endothelial cells, and hematopoietic stem cells across different cancers. Supporting that, ECs are the primary cell type expressing CD93 revealed by published single-cell RNA sequencing datasets of multiple human tumor tissues (Sun et al., 2021). Single-cell transcriptomics revealed that CD93 can enrich HSC/MPP subsets with enhanced stem cell properties and regulate their development (Gao et al., 2021). In accordance with previous results on the role of CD93 in regulating immune infiltrates, we observed that CD93 was enriched in mononuclear/macrophage subsets, neutrophils, and B cells in some types of cancer. Consistently, CD93 was reported to be also expressed on several immune cell types, including monocytes, immature B cells, and platelets (Greenlee-Wacker et al., 2012). Based on the positive association of CD93 and M2 macrophages, which leads to an immunosuppressive phenotype, we can suppose that CD93 may exert an immune suppressive role in TME. More recently, numbers of works support the importance of CD93 expression in immune response during the development of cancers via alterations in immune cell phenotype, function, and infiltration (Galvagni et al., 2017; Lugano et al., 2018; Huang et al., 2021; Riether et al., 2021). For instance, CD93 was recently found to improve tumor vascular functions to promote T-cell infiltration and antitumor immunity after blockade of its pathway (Sun et al., 2021). In addition, we also found that CD93 expression was positively related to almost all immune-related genes, which are known to promote the growth of tumor cells and tumor angiogenesis as well as contribute to tumor metastasis (Li et al., 2021; Schaafsma et al., 2021) in nearly all cancers, which further demonstrated the importance of CD93 in the process of immune regulation.

To specifically address the function of CD93 in human pan-cancer, we performed enrichment analysis on CD93 and found a link between it and promoting cancers and inflammation. Concurrently, we have further validated the link between CD93 expression and autophagy, EMT, TGFB1, and the Wnt pathway. The finding that high expression of CD93 was mainly associated with inflammation and angiogenesis may imply that not only does CD93 exerts a crucial role in cancer, but it may be also important in inflammatory diseases and general homeostasis. Consistent with our data, CD93 was reported to be involved in apoptosis and inflammation and had a suggested role in angiogenesis, and thus involved in the development and dissemination of cancer (Olsen et al., 2015). CD93 also played roles in apoptosis, phagocytosis, inflammation, and cell adhesion (Nepomuceno et al., 1997; McGreal et al., 2002; Norsworthy et al., 2004). Interestingly, a key step during vascular maturation included the deposition of Multimerin-2, which acted not only as a substrate for pericyte adhesion but also as a central modifier of the expression pattern of important molecules regulating the crosstalk between ECs and pericytes and vascular stability (Fejza et al., 2021). However, as we all know, multimerin-2 is a known CD93 ligand (Galvagni et al., 2017; Khan et al., 2017). Consistently, IGFBP7, another known ligand for CD93, has been shown to promote EC angiogenesis via CD93 (Hooper et al., 2009; Komiya et al., 2014; Sun et al., 2021). These results suggest that CD93 is strongly associated with tumor invasion and metastasis.

These promising findings suggest that applying CD93 to the clinic is a feasible goal. Despite the provided new insights into CD93, more research will be required to understand how CD93 could modulate immune response. This would improve the efficiency of newer immunotherapy drugs, notably in combination with immunomodulatory therapies, which showed promising antitumor effects in some types of cancer in recent years. Nonetheless, our study identified a role for CD93 in regulating TME, influencing clinical prognosis and immunoregulation effect, particularly immune infiltration, and expanding our understanding of CD93 biology and tissue–immune system interaction, which may lead to new treatments for inflammatory diseases and cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ML designed and supervised the study. ZZ analyzed the data, and MZ wrote the original draft. QD edited the draft. Both the authors have read and approved the final manuscript.

FUNDING

The study was supported in part by the National Natural Science Foundation of China (No. 82002609).

ACKNOWLEDGMENTS

The authors express gratitude to the public database TCGA, GTEx, and CCLE.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.817965/full#supplementary-material

REFERENCES

Bagaev, A., Kotlov, N., Nomie, K., Svekolkin, V., Gafurov, A., Isaeva, O., et al. (2021). Conserved Pan-Cancer Microenvironment Subtypes Predict Response to Immunotherapy. *Cancer Cell* 39, 845–865. e7. doi:10.1016/j.ccell.2021.04.014

Barbera, S., Lugano, R., Pedalina, A., Mongiat, M., Santucci, A., Tosi, G. M., et al. (2021). The C-type Lectin CD93 Controls Endothelial Cell Migration via Activation of the Rho Family of Small GTPases. *Matrix Biol.* 99, 1–17. doi:10.1016/j.matbio.2021.05.006

Bejarano, L., Jordão, M. J. C., and Joyce, J. A. (2021). Therapeutic Targeting of the Tumor Microenvironment. *Cancer Discov.* 11, 933–959. doi:10.1158/2159-8290.CD-20-1808

Boland, C. R., and Goel, A. (2010). Microsatellite Instability in Colorectal Cancer. *Gastroenterology* 138, 2073–2087. e3. doi:10.1053/j.gastro.2009.12.064

Boussios, S., Ozturk, M., Moschetta, M., Karathanasi, A., Zakynthinakis-Kyriakou, N., Katsanos, K., et al. (2019). The Developing story of Predictive Biomarkers in Colorectal Cancer. *Jpm* 9, 12. doi:10.3390/jpm9010012

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J. Clinicians* 68, 394–424. doi:10.3322/caac.21492

C. Greenlee-Wacker, M., D. Galvan, M., and S. Bohlson, S. (2012). CD93: Recent Advances and Implications in Disease. *Cdt* 13, 411–420. doi:10.2174/138945012799424651

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., et al. (2017). Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational burden. *Genome Med.* 9, 34. doi:10.1186/s13073-017-0424-2

Chan, T. A., Yarchoan, M., Jaffee, E., Swanton, C., Quezada, S. A., Stenzinger, A., et al. (2019). Development of Tumor Mutation burden as an Immunotherapy Biomarker: Utility for the Oncology Clinic. *Ann. Oncol.* 30, 44–56. doi:10.1093/annonc/mdy495

Cokol, M., Nair, R., and Rost, B. (2000). Finding Nuclear Localization Signals. *EMBO Rep.* 1, 411–415. doi:10.1093/embo-reports/kvd092

Fejza, A., Poletto, E., Carobolante, G., Camicia, L., Andreuzzi, E., Capuano, A., et al. (2021). Multimerin-2 Orchestrates the Cross-Talk between Endothelial Cells and Pericytes: A Mechanism to Maintain Vascular Stability. *Matrix Biol. Plus* 11, 100068. doi:10.1016/j.mbplus.2021.100068

Fishel, R. (2015). Mismatch Repair. *J. Biol. Chem.* 290, 26395–26403. doi:10.1074/jbc.R115.660142

Galvagni, F., Nardi, F., Spiga, O., Trezza, A., Tarticchio, G., Pellicani, R., et al. (2017). Dissecting the CD93-Multimerin 2 Interaction Involved in Cell Adhesion and Migration of the Activated Endothelium. *Matrix Biol.* 64, 112–127. doi:10.1016/j.matbio.2017.08.003

Gao, S., Shi, Q., Zhang, Y., Liang, G., Kang, Z., Huang, B., et al. (2021). Identification of HSC/MPP Expansion Units in Fetal Liver by Single-Cell Spatiotemporal Transcriptomics. *Cell Res* 32, 38–53. doi:10.1038/s41422-021-00540-7

Greenlee-Wacker, M. C., Briseño, C., Galvan, M., Moriel, G., Velázquez, P., and Bohlson, S. S. (2011). Membrane-associated CD93 Regulates Leukocyte Migration and C1q-Hemolytic Activity during Murine Peritonitis. *J.I.* 187, 3353–3361. doi:10.4049/jimmunol.1100803

Han, Z., Dong, Y., Lu, J., Yang, F., Zheng, Y., and Yang, H. (2021). Role of Hypoxia in Inhibiting Dendritic Cells by VEGF Signaling in Tumor Microenvironments: Mechanism and Application. *Am. J. Cancer Res.* 11, 3777–3793.

Hiam-Galvez, K. J., Allen, B. M., and Spitzer, M. H. (2021). Systemic Immunity in Cancer. *Nat. Rev. Cancer* 21, 345–359. doi:10.1038/s41568-021-00347-z

Hooper, A. T., Shmelkov, S. V., Gupta, S., Milde, T., Bambino, K., Gillen, K., et al. (2009). Angiomodulin Is a Specific Marker of Vasculature and Regulates Vascular Endothelial Growth Factor-A-dependent Neoangiogenesis. *Circ. Res.* 105, 201–208. doi:10.1161/CIRCRESAHA.109.196790

Huang, J., Lee, H.-y., Zhao, X., Han, J., Su, Y., Sun, Q., et al. (2021). Interleukin-17D Regulates Group 3 Innate Lymphoid Cell Function through its Receptor CD93. *Immunity* 54, 673–686. e4. doi:10.1016/j.immuni.2021.03.018

Joshi, S. S., and Badgwell, B. D. (2021). Current Treatment and Recent Progress in Gastric Cancer. *CA A. Cancer J. Clin.* 71, 264–279. doi:10.3322/caac.21657

Khan, K. A., Naylor, A. J., Khan, A., Noy, P. J., Mambretti, M., Lodhia, P., et al. (2017). Multimerin-2 Is a Ligand for Group 14 Family C-type Lectins CLEC14A, CD93 and CD248 Spanning the Endothelial Pericyte Interface. *Oncogene* 36, 6097–6108. doi:10.1038/onc.2017.214

Komiya, E., Sato, H., Watanabe, N., Ise, M., Higashi, S., Miyagi, Y., et al. (2014). Angiomodulin, a Marker of Cancer Vasculature, Is Upregulated by Vascular Endothelial Growth Factor and Increases Vascular Permeability as a Ligand of Integrin α V β 3. *Cancer Med.* 3, 537–549. doi:10.1002/cam4.216

Langenkamp, E., Zhang, L., Lugano, R., Huang, H., Elhassan, T. E. A., Georganaki, M., et al. (2015). Elevated Expression of the C-type Lectin CD93 in the Glioblastoma Vasculature Regulates Cytoskeletal Rearrangements that Enhance Vessel Function and Reduce Host Survival. *Cancer Res.* 75, 4504–4516. doi:10.1158/0008-5472.CAN-14-3636

Lee, D.-W., Han, S.-W., Bae, J. M., Jang, H., Han, H., Kim, H., et al. (2019). Tumor Mutation burden and Prognosis in Patients with Colorectal Cancer Treated with Adjuvant Fluoropyrimidine and Oxaliplatin. *Clin. Cancer Res.* 25, 6141–6147. doi:10.1158/1078-0432.CCR-19-1105

Li, L., Yao, W., Yan, S., Dong, X., Lv, Z., Jing, Q., et al. (2021). Pan-cancer Analysis of Prognostic and Immune Infiltrates for CXCs. *Cancers* 13, 4153. doi:10.3390/cancers13164153

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* 77, e108–e110. doi:10.1158/0008-5472.CAN-17-0307

Lugano, R., Vemuri, K., Yu, D., Bergqvist, M., Smits, A., Essand, M., et al. (2018). CD93 Promotes β1 Integrin Activation and Fibronectin Fibrillogenesis during Tumor Angiogenesis. *J. Clin. Invest.* 128, 3280–3297. doi:10.1172/JCI97459

Lyssiotis, C. A., and Kimmelman, A. C. (2017). Metabolic Interactions in the Tumor Microenvironment. *Trends Cel Biol.* 27, 863–875. doi:10.1016/j.tcb.2017.06.003

Maia, M., DeVriese, A., Janssens, T., Moons, M., Lories, R. J., Tavernier, J., et al. (2011). CD248 Facilitates Tumor Growth via its Cytoplasmic Domain. *BMC Cancer* 11, 162. doi:10.1186/1471-2407-11-162

Marabelle, A., Fakih, M., Lopez, J., Shah, M., Shapira-Frommer, R., Nakagawa, K., et al. (2020). Association of Tumour Mutational burden with Outcomes in Patients with Advanced Solid Tumours Treated with Pembrolizumab: Prospective Biomarker Analysis of the Multicohort, Open-Label, Phase 2 KEYNOTE-158 Study. *Lancet Oncol.* 21, 1353–1365. doi:10.1016/S1470-2045(20)30445-9

Masiero, M., Simões, F. C., Han, H. D., Snell, C., Peterkin, T., Bridges, E., et al. (2013). A Core Human Primary Tumor Angiogenesis Signature Identifies the Endothelial Orphan Receptor ELTD1 as a Key Regulator of Angiogenesis. *Cancer Cell* 24, 229–241. doi:10.1016/j.ccr.2013.06.004

McGreal, E. P., Ikewaki, N., Akatsu, H., Morgan, B. P., Gasque, P., and Human, C. (2002). Human C1qRp Is Identical with CD93 and the mNI-11 Antigen but Does Not Bind C1q. *J. Immunol.* 168, 5222–5232. doi:10.4049/jimmunol.168.10.5222

Meléndez, B., Van Campenhout, C., Rorive, S., Remmelink, M., Salmon, I., and D'Haene, N. (2018). Methods of Measurement for Tumor Mutational burden in Tumor Tissue. *Transl. Lung Cancer Reslung Cancer Res.* 7, 661–667. doi:10.21037/tlcr.2018.08.02

Nanda, A., Karim, B., Peng, Z., Liu, G., Qiu, W., Gan, C., et al. (2006). Tumor Endothelial Marker 1 (Tem1) Functions in the Growth and Progression of Abdominal Tumors. *Proc. Natl. Acad. Sci.* 103, 3351–3356. doi:10.1073/pnas.0511306103

Nepomuceno, R. R., Henschen-Edman, A. H., Burgess, W. H., and Tenner, A. J. (1997). cDNA Cloning and Primary Structure Analysis of C1qRP, the Human C1q/MBL/SPA Receptor that Mediates Enhanced Phagocytosis *In Vitro*. *Immunity* 6, 119–129. doi:10.1016/s1074-7613(00)80419-7

Norsworthy, P. J., Fossati-Jimack, L., Cortes-Hernandez, J., Taylor, P. R., Bygrave, A. E., Thompson, R. D., et al. (2004). Murine CD93 (C1qRp) Contributes to the Removal of Apoptotic Cells *In Vivo* but Is Not Required for C1q-Mediated Enhancement of Phagocytosis. *J. Immunol.* 172, 3406–3414. doi:10.4049/jimmunol.172.6.3406

Olsen, R. S., Lindh, M., Vorkapic, E., Andersson, R. E., Zar, N., Löfgren, S., et al. (2015). CD93 Gene Polymorphism Is Associated with Disseminated Colorectal Cancer. *Int. J. Colorectal Dis.* 30, 883–890. doi:10.1007/s00384-015-2247-1

Passaro, A., Stenzinger, A., and Peters, S. (2020). Tumor Mutational burden as a Pan-Cancer Biomarker for Immunotherapy: The Limits and Potential for Convergence. *Cancer Cell* 38, 624–625. doi:10.1016/j.ccell.2020.10.019

Pessoa Rodrigues, C., and Akhtar, A. (2021). Differential H4K16ac Levels Ensure a Balance between Quiescence and Activation in Hematopoietic Stem Cells. *Sci. Adv.* 7. doi:10.1126/sciadv.abi5987

Riether, C., Radpour, R., Kallen, N. M., Bürgin, D. T., Bachmann, C., Schürch, C. M., et al. (2021). Metoclopramide Treatment Blocks CD93-Signaling-Mediated Self-Renewal of Chronic Myeloid Leukemia Stem Cells. *Cel Rep.* 34, 108663. doi:10.1016/j.celrep.2020.108663

Russo, M., Crisafulli, G., Sogari, A., Reilly, N. M., Arena, S., Lamba, S., et al. (2019). Adaptive Mutability of Colorectal Cancers in Response to Targeted Therapies. *Science* 366, 1473–1480. doi:10.1126/science.aav4474

Samstein, R. M., Lee, C.-H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., et al. (2019). Tumor Mutational Load Predicts Survival after Immunotherapy across Multiple Cancer Types. *Nat. Genet.* 51, 202–206. doi:10.1038/s41588-018-0312-8

Schaafsma, E., Fugle, C. M., Wang, X., and Cheng, C. (2021). Pan-cancer Association of HLA Gene Expression with Cancer Prognosis and Immunotherapy Efficacy. *Br. J. Cancer* 125, 422–432. doi:10.1038/s41416-021-01400-2

Sun, Y., Chen, W., Torphy, R. J., Yao, S., Zhu, G., Lin, R., et al. (2021). Blockade of the CD93 Pathway Normalizes Tumor Vasculature to Facilitate Drug Delivery and Immunotherapy. *Sci. Transl. Med.* 13. doi:10.1126/scitranslmed.abc8922

Vitale, I., Manic, G., Coussens, L. M., Kroemer, G., and Galluzzi, L. (2019). Macrophages and Metabolism in the Tumor Microenvironment. *Cel Metab.* 30, 36–50. doi:10.1016/j.cmet.2019.06.001

Ward, R., Meagher, A., Tomlinson, I., O'Connor, T., Norrie, M., Wu, R., et al. (2001). Microsatellite Instability and the Clinicopathological Features of Sporadic Colorectal Cancer. *Gut* 48, 821–829. doi:10.1136/gut.48.6.821

Yarchoan, M., Hopkins, A., and Jaffee, E. M. (2017). Tumor Mutational burden and Response Rate to PD-1 Inhibition. *N. Engl. J. Med.* 377, 2500–2501. doi:10.1056/NEJMc1713444

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhang, Zheng, Ding and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 17 February 2022
doi: 10.3389/fcell.2022.790214

Gene Expression Profile Reveals a Prognostic Signature of Non–MSI-H/pMMR Colorectal Cancer

Zaoqu Liu[1,2,3†], *Hui Xu*[1,2,3†], *Xiaoyong Ge*[1†], *Siyuan Weng*[1,2,3], *Qin Dang*[4] *and Xinwei Han*[1,2,3*]

[1]Department of Interventional Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2]Interventional Institute of Zhengzhou University, Zhengzhou, China, [3]Interventional Treatment and Clinical Research Center of Henan Province, Zhengzhou, China, [4]Department of Colorectal Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Xin Hu,
Tianjin Medical University Cancer Institute and Hospital, China
Mingjian Lu,
Institute of Guangzhou Medical University, China
Weihao Kong,
First Affiliated Hospital of Anhui Medical University, China
Peng Luo,
The University of Hong Kong, Hong Kong SAR, China

***Correspondence:**
Xinwei Han
fcchanxw@zzu.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 06 October 2021
Accepted: 18 January 2022
Published: 17 February 2022

Citation:
Liu Z, Xu H, Ge X, Weng S, Dang Q and Han X (2022) Gene Expression Profile Reveals a Prognostic Signature of Non–MSI-H/pMMR Colorectal Cancer. Front. Cell Dev. Biol. 10:790214. doi: 10.3389/fcell.2022.790214

Studies have demonstrated that non–MSI-H/pMMR colorectal cancer (CRC) has a worse prognosis and relapse rate than microsatellite instability-high (MSI-H)/mismatch repair deficient (dMMR) CRC. Hence, searching for a novel tool to advance the prognostic management of non–MSI-H/pMMR CRC is vital. In this study, using three independent public cohorts and a clinical in-house cohort, we developed and validated a microsatellite stable–associated signature (MSSAS). The initial signature establishment was performed in GSE39582 (n = 454). This was followed by independent validation of this signature in The Cancer Genome Atlas–CRC (n = 312), GSE39084 (n = 54), and in-house cohort (n = 146). As a result, MSSAS was proven to be an independent risk factor for overall survival and relapse-free survival in non–MSI-H/pMMR CRC. Receiver operating characteristic analysis showed that MSSAS had a stable and accurate performance in all cohorts for 1, 3, and 5 years, respectively. Further analysis suggested that MSSAS performed better than age, gender, and the T, N, M, and AJCC stages, adjuvant chemotherapy, tumor mutation burden, neoantigen, and *TP*53, *KRAS*, *BRAF*, and *PIK3CA* mutations. The clinical validation was executed to further ensure the robustness and clinical feasibility of this signature. In conclusion, MSSAS might be a robust and promising biomarker for advancing clinical management of non–MSI-H/pMMR CRC.

Keywords: microsatellite-stability, colorectal cancer, gene signature, recurrence, prognosis

INTRODUCTION

Despite tremendous advances in the treatment of cancer, colorectal cancer (CRC) remains the second leading cause of cancer-related deaths (Carethers and Doubeni, 2020). Based on mutational status, CRC can be divided into mismatch repair deficient (dMMR)/microsatellite instability-high (MSI-H) and mismatch repair proficient (pMMR)/non–MSI-H tumors (Ganesh et al., 2019; Endo et al., 2020). Chemotherapy alone (5-fluorouracil) or in combination with radiation is the standard treatment for patients with advanced CRC (Maiuthed et al., 2018; Dekker et al., 2019). Although non–MSI-H/pMMR CRC has better efficacy for 5-fluorouracil–based chemotherapy, non–MSI-H/pMMR CRC patients have demonstrated a worse prognosis and higher relapse rate (Lizardo et al., 2020). In recent years, immunotherapy has attracted tremendous attention due to its significant effect on solid tumors, including CRC (Liu et al., 2021a). Accumulating evidence has revealed that immunotherapy improves the prognosis of MSI-H/pMMR CRC patients but has no effect on non–MSI-H/pMMR CRC (Biller and Schrag, 2021). Besides, non–MSI-H/pMMR CRC patients are

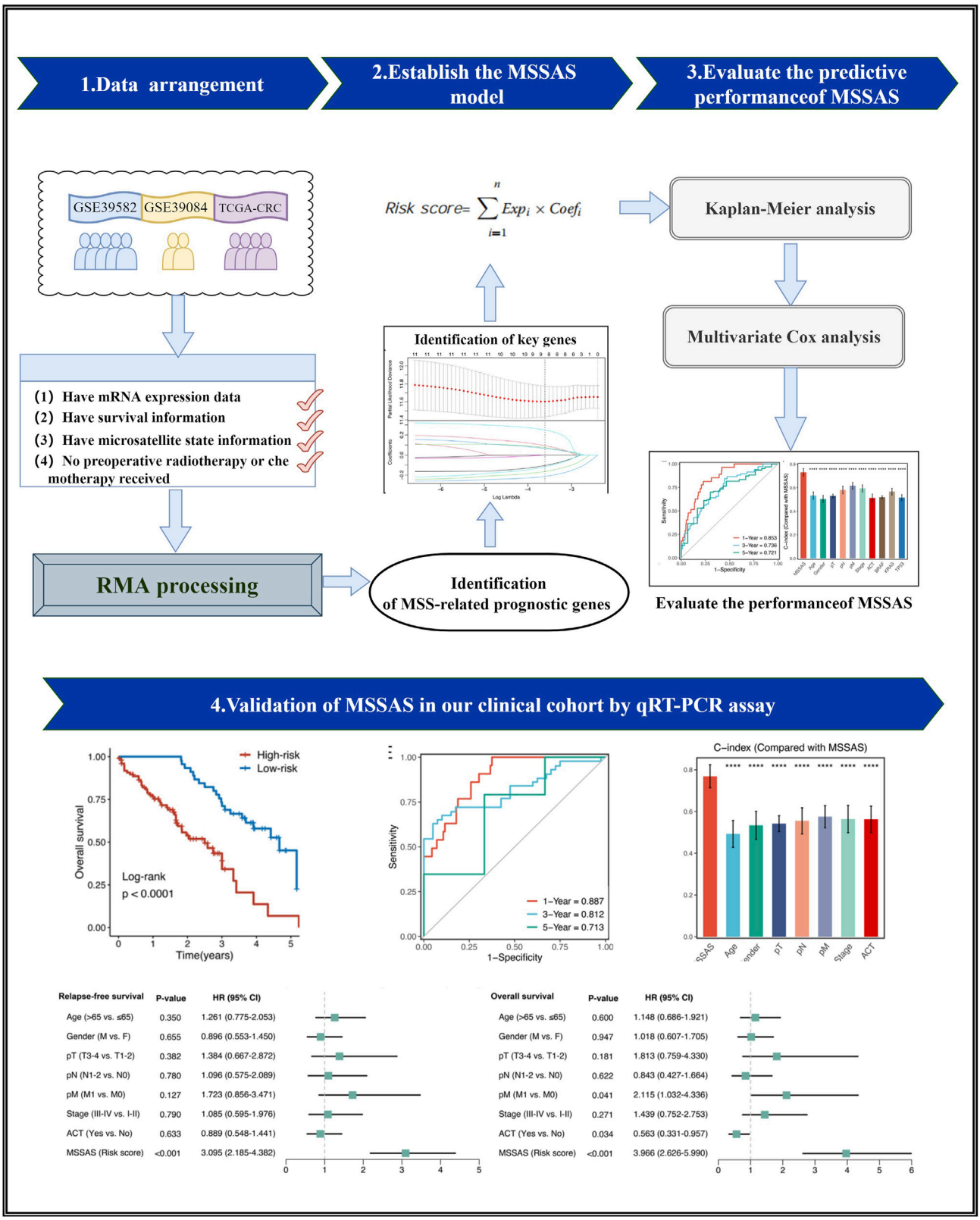


FIGURE 1 | The flowchart of this study.

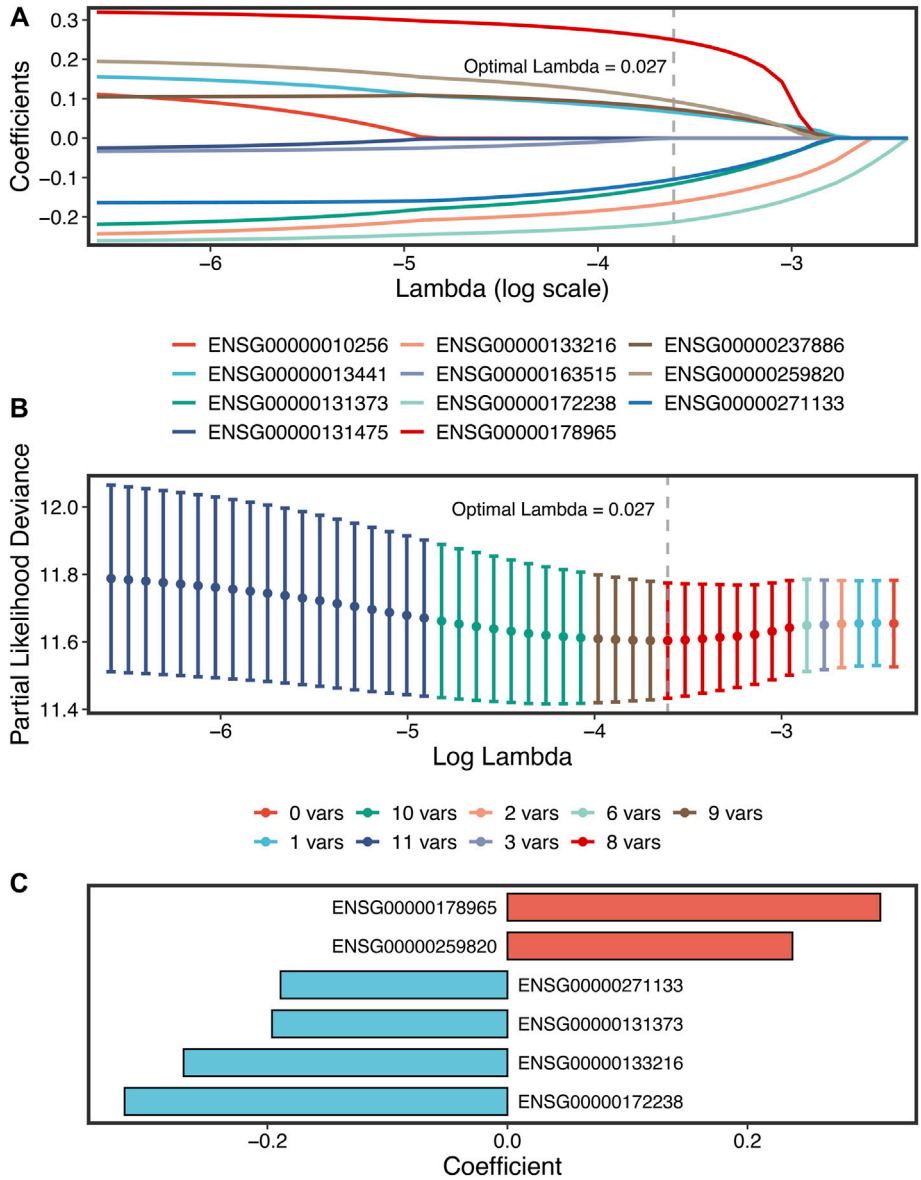


FIGURE 2 | The development of the MSSAS signature based on the LASSO stepwise Cox regression algorithms. **(A)** Ten-fold cross-validations to tune the parameter selection in the LASSO model. **(B)** LASSO coefficient profiles of the candidate genes for MSSAS construction. The vertical lines are drawn at the optimal values by minimum criteria. **(C)** Coefficients of six genes finally obtained in stepwise Cox regression.

likely to develop resistance after immunotherapy (Kather et al., 2018). Considering the dismal prognosis, recurrence, and the unsatisfactory immunotherapeutic outcomes of non–MSI-H/pMMR CRC, assessing the risk of non–MSI-H/pMMR CRC for further clinical intervention is warranted.

We hypothesized that comprehensively identifying core prognosis–related genes to construct a predictive model would enhance the accuracy of prognosis evaluation for non–MSI-H/pMMR CRC patients. Commonly used genetic testing techniques such as quantitative real-time polymerase chain reaction (qRT-PCR) are currently defective in detecting a large number of genes (Liu et al., 2021b), while the development of bioinformatics has made this possible. Presently, it is very easy to use large-scale genes for subsequent analyses. Furthermore, with the tremendous progress of machine learning, such as the least absolute shrinkage and selection operator (LASSO) algorithm, the most important elements in gene expression profiles can be determined and models with powerful extrapolation capabilities fitted (Liu et al., 2021c).

In our study, three independent public cohorts (n = 820) were used to develop and validate a microsatellite stable–associated signature (MSSAS). A total of 146 non–MSI-H CRC tissues were collected for external verification to assess the robustness and accuracy of MSSAS. Finally, a novel six-genes signature was

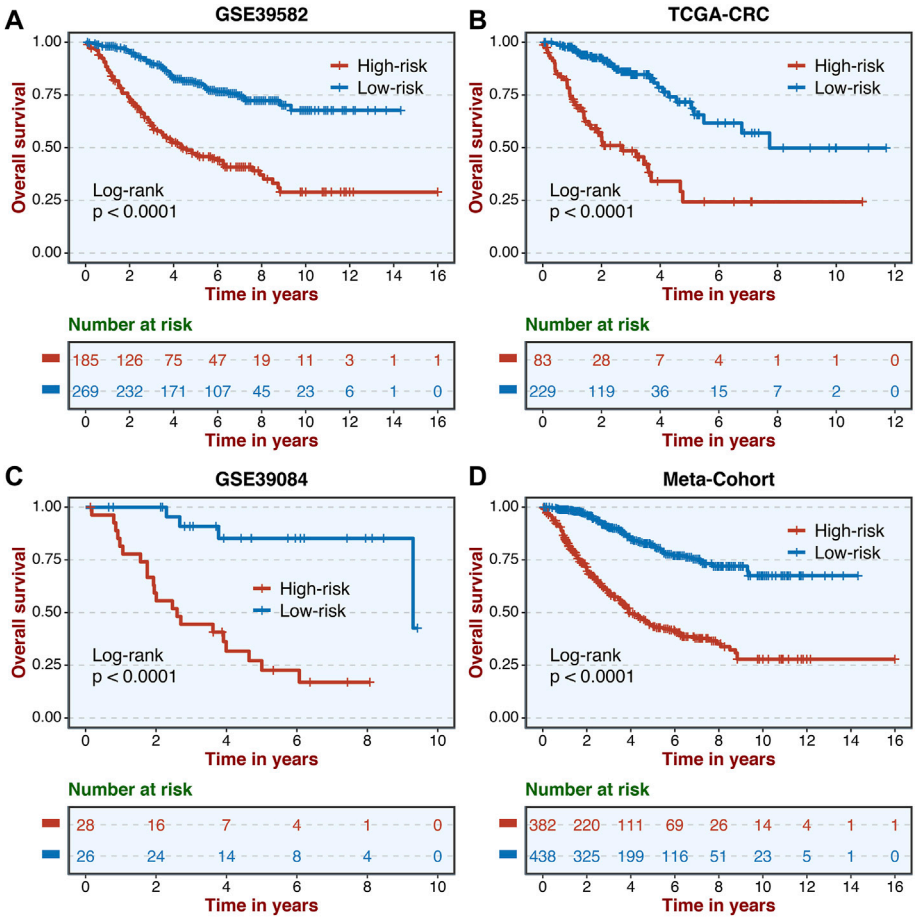


FIGURE 3 | Kaplan–Meier analysis of OS according to MSSAS. Kaplan–Meier curves of OS according to the MSSAS in GSE39582 **(A)**, TCGA–CRC **(B)**, GSE39084 **(C)**, and Meta-Cohort **(D)**.

reported. This signature not only offers robust and high reliability in identifying patients at high recurrence risk and worse overall survival but also can be readily applied into clinical practice due to the inexpensiveness and simplicity of the PCR-based assays. Taken together, MSSAS provides an excellent platform for assessing prognosis and recurrence risk and might be a promising tool to facilitate the clinical management of non–MSI-H/pMMR CRC patients.

MATERIALS AND METHODS

Data Collection and Processing

The overall flowchart is shown in **Figure 1**. Three independent CRC cohorts were retrieved from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) and The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov), including GSE39582, GSE92921, and TCGA-CRC. The microarray data were normalized by the robust multi-array average algorithm, and the RNA-seq raw count data were performed by transcripts per kilobase million and log-2 transformation. A combat algorithm was used to remove the batch effects in Meta-Cohort (derived from three microarray data sets). In these three cohorts, patients who met the following criteria were retained: those with 1) mRNA expression data, 2) survival information, 3) microsatellite state information, and 4) no preoperative radiotherapy or chemotherapy received. Ultimately, TCGA-CRC had 54 MSI-H CRC and 312 non–MSI-H CRC; GSE39582 had 77 dMMR CRC and 459 proficient mismatch repair (pMMR) CRC; and GSE39084 had 16 MSI-H CRC and 54 non–MSI-H CRC. The detailed baseline is summarized in **Supplementary Table S1**.

Identification of Microsatellite Stable–Related Prognostic Genes

To identify the microsatellite stable (MSS)–related prognostic genes, we developed a pipeline as follows: 1) across the three cohorts, the *limma* package was applied to perform differential analysis between the non–MSI-H/PMMR CRC and MSI-H/dMMR CRC groups with an FDR < 0.01; the overlap of the three cohorts was defined as MSS-related genes. Subsequently, univariate Cox regression analysis was performed on these genes.

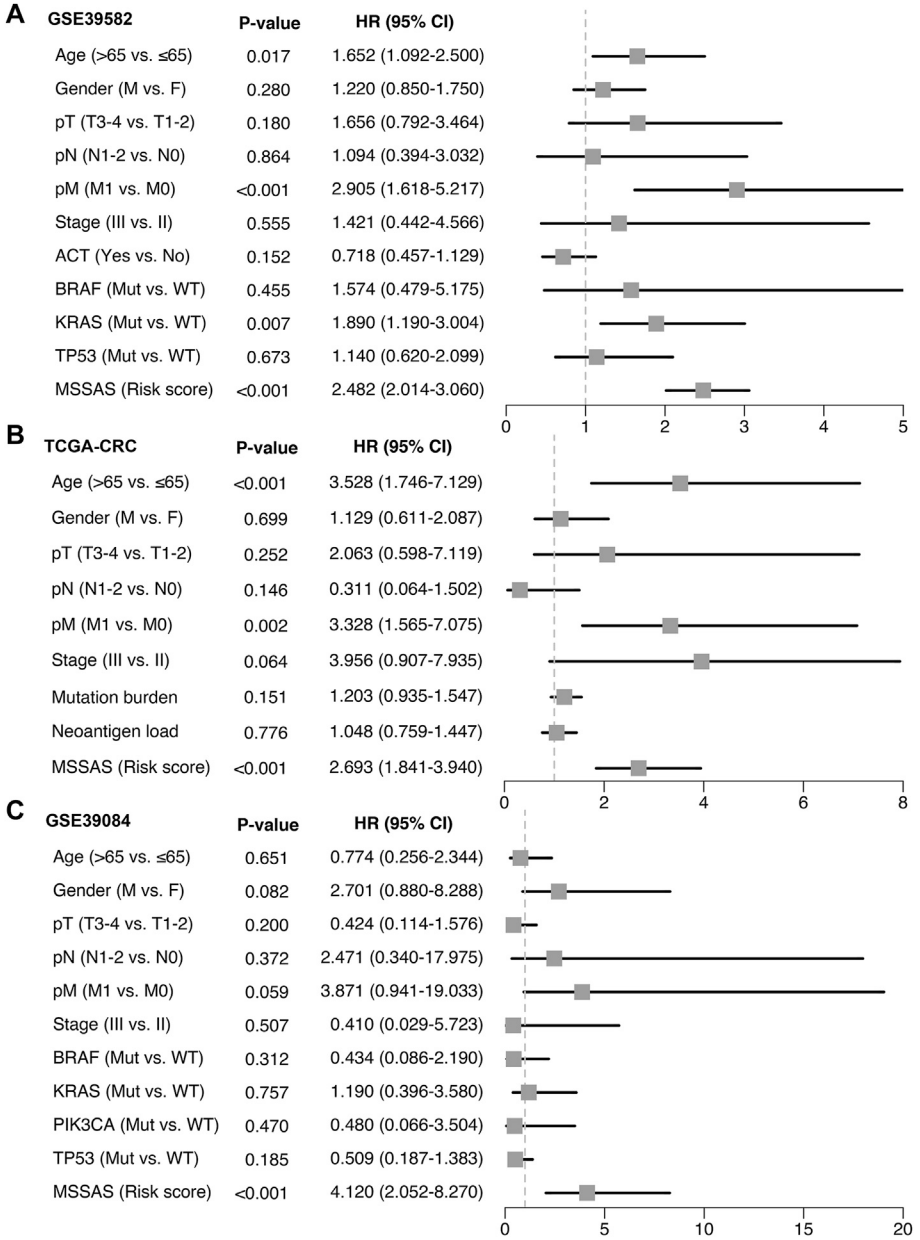


FIGURE 4 | Multivariable Cox regression analysis of OS. Multivariable Cox regression analysis of MSSAS for OS in GSE39582 **(A)**, TCGA–CRC **(B)**, and GSE39084 **(C)**.

Ultimately, genes with a hazards ratio (HR) consistently >1 or <1 and $p < 0.05$ across the three cohorts were considered as stable prognostic genes.

Establishment of Microsatellite Stable–Associate Signature

The LASSO is a prevalent machine learning algorithm, which was employed for variable selection to identify key genes in this study. By 10-fold cross-validation, the optimal lambda was

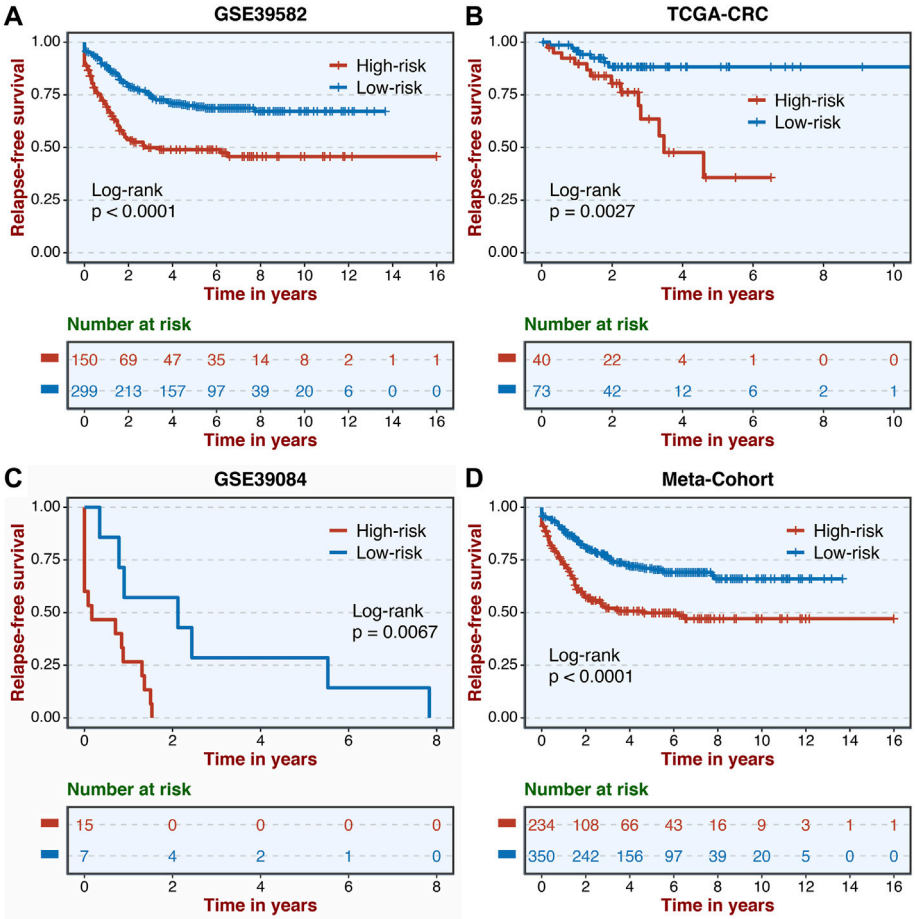


FIGURE 5 | Kaplan–Meier analysis of RFS according to MSSAS. Kaplan–Meier curves of RFS according to the MSSAS in GSE39582 **(A)**, TCGA–CRC **(B)**, GSE39084 **(C)**, and Meta-Cohort **(D)**.

generated when the partial likelihood deviance reached the minimum value. To make our signature more practical, we further adopted the stepwise Cox algorithm and generated a risk score formula which includes the regression coefficients of weighted genes.

Human Colorectal Cancer Specimens and Quantitative Real-Time Polymerase Chain Reaction Analysis

A total of 146 frozen resected CRC tissues with non–MSI-H were collected from The First Affiliated Hospital of Zhengzhou University. The clinical staging of the specimens was based on NCCN (2019) guidelines; see **Supplementary Table S1** for the detailed baseline data of our samples. Total RNA was isolated with RNAiso Plus reagent (Waltham, MA, United States) as described previously (Liu et al., 2021b; Liu et al., 2021c; Liu et al., 2021d; Zhang et al., 2021). The primer sequences of the included six genes and *GAPDH* are shown in **Supplementary Table S2**. The details of the qRT-PCR analysis are described in the **Supplementary Material**.

Statistical Analysis

The *survival* package was used to conduct Kaplan–Meier and multivariate Cox regression analyses. The optimal cut-off value was determined by the *survminer* package. The *timeROC* package was performed to plot time-dependent receiver operating characteristic (ROC) curves, and comparisons of the concordance index (C-index) among the variables were conducted by using the *compareC* package. All data processing, statistical analyses, and plotting were conducted in R 4.0.5 software.

RESULTS

Construction and Validation of Microsatellite Stable–Associated Signature

A total of 307 overlapped genes were obtained with differential analyses for these three cohorts. After univariate Cox regression, 11 genes in the three cohorts were defined as stable MSS-related prognostic genes (**Supplementary Table S3**). Subsequently, LASSO Cox regression was performed to identify eight key

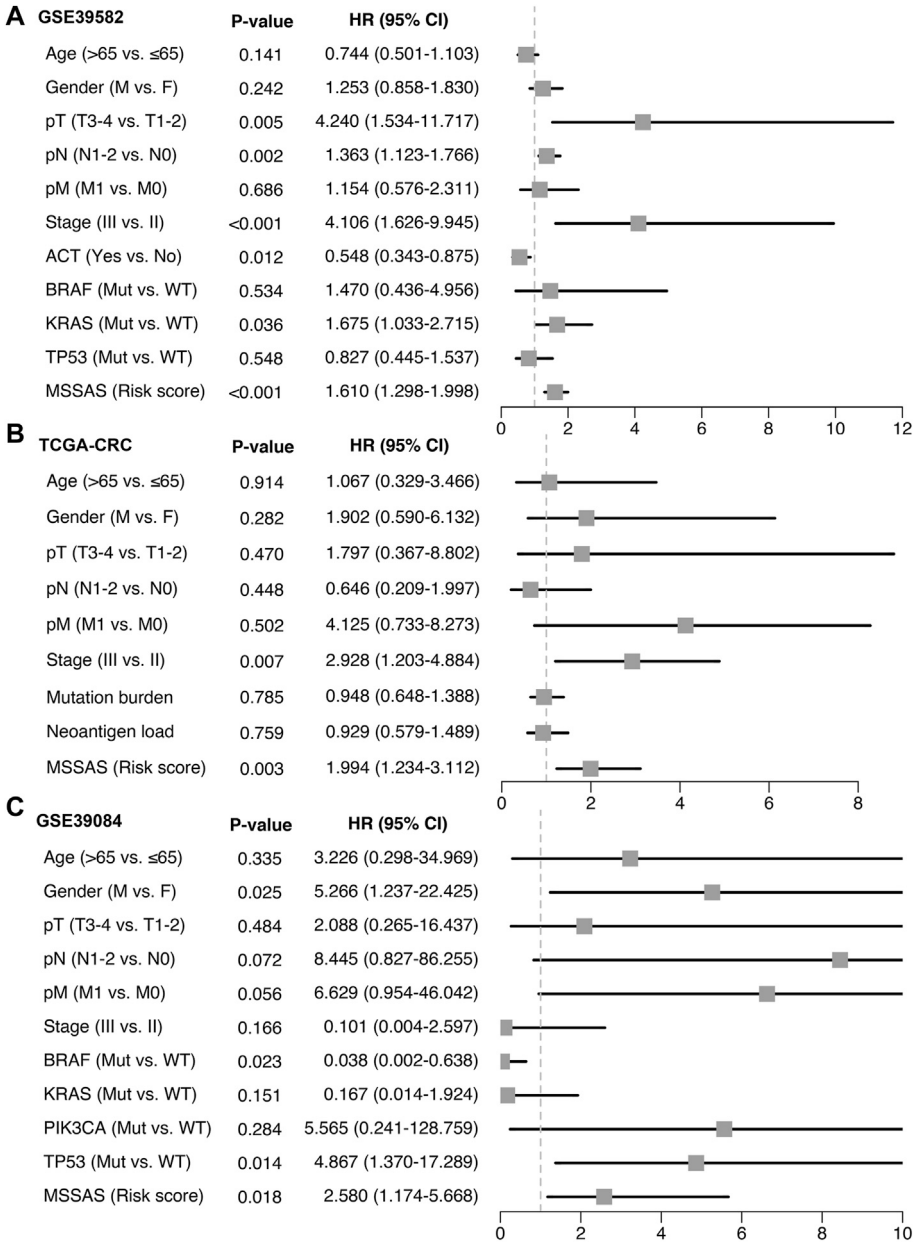


FIGURE 6 | Multivariable Cox regression analysis of RFS. Multivariable Cox regression analysis of MSSAS for RFS in GSE39582 **(A)**, TCGA–CRC **(B)**, and GSE39084 **(C)**.

genes, and these genes were subjected to stepwise Cox regression (**Figures 2A,B**). Ultimately, the MSSAS composed of six genes was developed (**Figure 2C**). A risk score was established according to the expression of the six genes weighted by their regression coefficients in a penalized Cox model as follows: risk score = −0.189*Exp (*ITGB8-AS1*) −0.270*Exp (*EPHB2*) −0.319*Exp (*ATOH1*) −0.196*Exp (*HACL1*) +0.237*Exp (*Lnc-ZFAT-1*) +0.311*Exp (*ERICH3*) (**Figure 2C**). The risk scores of all patients with non–MSI-H/pMMR CRC were calculated from this formula. The overall survival (OS) was significantly worse in the high-risk group across GSE39582 [HR: 2.490 (2.126–2.917); log-rank test, $p < 0.0001$; **Figure 3A**], TCGA-CRC [HR: 2.69 (2.116–3.425); log-rank test, $p < 0.0001$; **Figure 3B**], GSE39084 [HR: 2.930 (1.864–4.604); log-rank test, $p < 0.0001$; **Figure 3C**], and Meta-Cohort [HR: 2.584 (2.282–2.925); log-rank test, $p < 0.0001$; **Figure 3D**] than in the low-risk group. After controlling for gender; age; T stage, N stage, M stage, and AJCC stage; adjuvant chemotherapy (ACT); tumor mutation burden (TMB); neoantigen; and *TP53*, *KRAS*, *BRAF*, and *PIK3CA* mutations, MSSAS remained statistically significant for predicting the OS in the three cohorts, which suggested that MSSAS was an independent risk factor for the OS (**Figures 4A–C**).

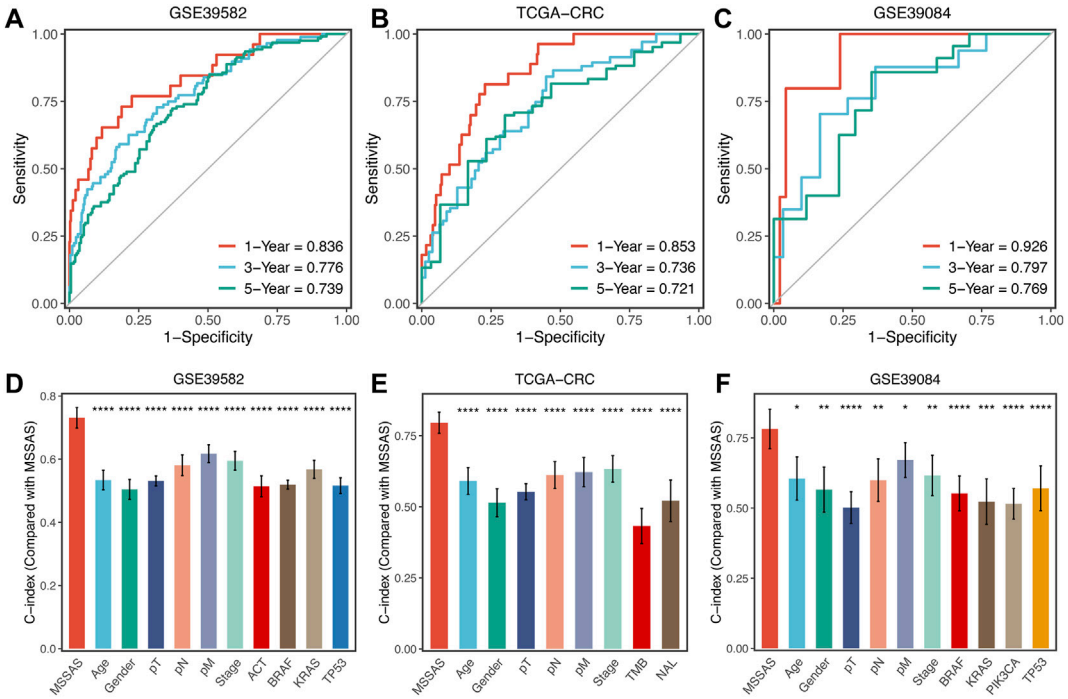


FIGURE 7 | Evaluation of MSSAS in three cohorts. Time-dependent ROC analysis for predicting OS at 1, 3, and 5 years in GSE39582 **(A)**, TCGA–CRC **(B)**, and GSE39084 **(C)**. The performance of MSSAS was compared with common clinical and molecular variables in predicting prognosis across GSE39582 **(D)**, TCGA-CRC **(E)**, and GSE39084 **(F)**.

Furthermore, we next explored the clinical significance of MSSAS in assessing the relapse-free survival (RFS). Similarly, the high-risk group also has dramatically dismal RFS in GSE39582 [HR: 1.674 (1.468–1.909); log-rank test, $p < 0.0001$; **Figure 5A**], TCGA-CRC [HR: 1.796 (1.037–3.110); log-rank test, $p = 0.0027$; **Figure 5B**], GSE39084 [HR: 1.892 (1.069–3.349); log-rank test, $p = 0.0067$; **Figure 5C**], and Meta-Cohort [HR: 1.697 (1.550–1.917); log-rank test, $p < 0.0001$; **Figure 5D**]. After the adjustments of available clinical variables, MSSAS remained statistically significant, which revealed that MSSAS was also an independent risk factor for RFS in non–MSI-H/pMMR CRC (**Figures 6A–C**).

Signature Predictive Capability

Subsequently, we evaluated the predictive capability of MSSAS in patients with non–MSI-H/pMMR CRC. The ROC analyses displayed that the areas under the curve of 1, 3, and 5 years were 0.836, 0.776, and 0.739 for GSE39582 (**Figure 7A**); 0.853, 0.736, and 0.721 for TCGA-CRC (**Figure 7B**); and 0.926, 0.797, and 0.769 for GSE39084 (**Figure 7C**), respectively. The C-indexes were 0.731 (0.689–0.773) in GSE39582, 0.796 (0.747–0.844) in TCGA-CRC, and 0.782 (0.691–0.873) in GSE39084, separately. These results revealed the high precision and value of MSSAS in forecasting the prognosis of non–MSI-H/pMMR CRC patients within 5 years. Moreover, MSSAS showed better performance than gender; age; T, N, M, and AJCC stages; ACT; TMB; neoantigen; and *TP53*, *KRAS*, *BRAF*, and *PIK3CA* mutations (**Figures 7D–F**), which indicated that MSSAS might be a promising surrogate in clinical settings.

Validation of the Microsatellite Stable–Associated Signature in Our Clinical Cohort

A total of 146 non–MSI-H CRC tissues were collected. The qRT-PCR assays were conducted to confirm the clinical translations of this six-genes signature. Consistently, patients in the high-risk group presented much worse OS and RFS than those in the low-risk group (**Figures 8A,B**). Also, multivariate Cox regression analysis indicated that MSSAS remained remarkably significant for the OS [HR: 3.966 (2.626–5.990); log-rank test, $p < 0.001$; **Figure 8C**] and RFS [HR: 3.095 (2.185–4.382); log-rank test, $p < 0.001$; **Figure 8D**] after adjusting for other clinical traits in non–MSI-H/pMMR CRC. The areas under the curve were 0.887, 0.812, and 0.713 at 1, 3, and 5 years, respectively (**Figure 8E**), and the C-index was also much more significant than other traditional factors (**Figure 8F**). These results indicate that MSSAS displayed a robust and powerful performance in assessing the prognosis of non–MSI-H/pMMR CRC patients.

DISCUSSION

CRC is a highly heterogeneous malignancy with complicated pathological processes and mechanisms (Picard et al., 2020; Liu et al., 2021d). 5-Fluorouracil chemotherapy alone or chemotherapy in combination with radiotherapy is the standard treatment for advanced CRC (Maiuthed et al., 2018; Dekker et al.,

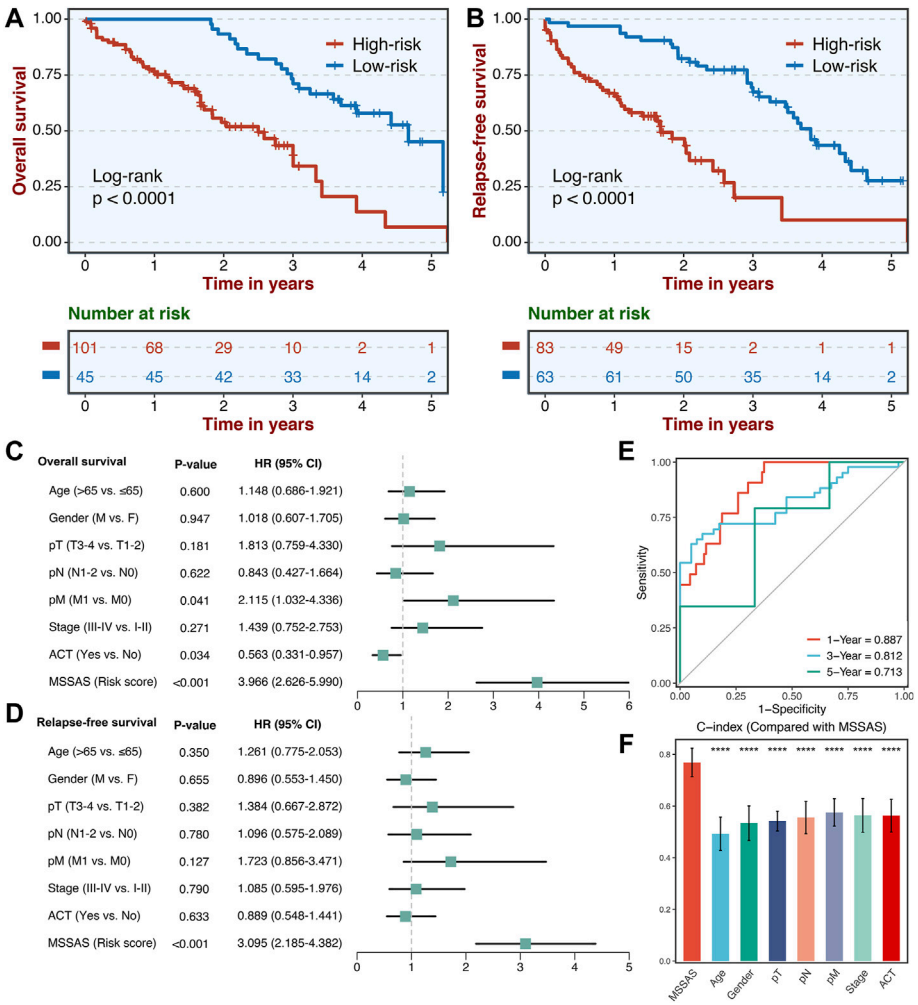


FIGURE 8 | Validation of MSSAS in our clinical cohort. Kaplan–Meier curves of OS **(A)** and RFS **(B)** according to the MSSAS. Multivariate Cox regression analysis of MSSAS for OS **(C)** and RFS **(D)**. **(E)** Time-dependent ROC analysis for predicting OS at 1, 3, and 5 years. **(F)**. The performance of MSSAS was compared with common clinical and molecular variables in predicting prognosis.

2019). Compared with MSI-H/dMMR CRC, non–MSI-H/pMMR CRC generally has a higher response rate to 5-fluorouracil–based chemotherapy (Webber et al., 2015). Unfortunately, the prognosis and recurrence rate of non–MSI-H/pMMR CRC patients are significantly worse (Vilar and Gruber, 2010). In recent years, immunotherapy has caused a great sensation due to its significant effect on the treatments of solid tumors (Liu et al., 2021a). Immunotherapy has improved the prognosis for patients with MSI-H/dMMR CRC indeed, but it has not been effective in non–MSI-H/pMMR CRC, which accounts for 95% of metastatic CRC (Le et al., 2015; Eng et al., 2019). Therefore, in view of the worse prognosis, higher recurrence rate, and limitations of immunotherapy in non–MSI-H/pMMR CRC, evaluating the risk of non–MSI-H/pMMR CRC to select the optimized clinical intervention is necessary.

Benefiting from the emergence of bioinformatics and artificial intelligence, the use of machine learning algorithms can screen out several core indicators that are important for predicting prognosis from a large scale of gene sets (Tibshirani, 1997). This actually conforms to a biological scale-free network dominated by several hub nodes (Zhang and Horvath, 2005). Hence, a new signature to assess the prognosis and recurrence risk of non–MSI-H/pMMR CRC was developed for the first time. The stable and reproducible results of MSSAS on multiple data sets and our qRT-PCR experimental data suggest that it not only is a stable and highly reliable signature but also has outstanding specificity and sensitivity. In addition, our model has a high value of clinical application given that qRT-PCR is simple and inexpensive to perform.

In our work, a prognosis and the relapse signature consisting of six genes were constructed, including *ITGB8-AS1*, *EPHB2*, *ATOH1*, *HACL1*, *ERICH3*, and *Lnc-ZFAT-1*. Studies have confirmed that most of these genes are associated with the development of tumors. For instance, *ITGB8-AS1* has been shown to be a biomarker for the growth and migration of CRC (Lin et al., 2021). The inactivation of *EPHB2* accelerates the occurrence of CRC. Besides, *EPHB2* mutations are closely associated with tumor suppression in gastric cancer and involved

in the prognosis and metastasis of other tumors such as Ewing sarcoma (Davalos et al., 2007; Keskin et al., 2021). *ATOH1* is a target of the *JNK1* and *MUC2* pathways and closely associated with the proliferation, invasion, and metastasis of CRC cells (Shen et al., 2018). *HACL1* is closely related to small-cell lung cancer, and *ERICH3* may be a key biomarker in nasopharyngeal carcinoma (Owonikoko et al., 2014; Zhang et al., 2019). The relationship of *Lnc-ZFAT-1* with cancer has not been reported in studies and needs to be further explored. Based on these six key genes, the MSSAS signature was finally established, which has been excellently displayed in predicting the prognosis of non–MSI-H/pMMR CRC patients. Multiple Cox regression analysis demonstrated that MSSAS was an independent feature after adjusting for other traditional clinical characteristics. What is more important, in three external data sets, MSSAS illustrated high calibration and discrimination in estimating the prognosis of non–MSI-H/pMMR CRC at 1, 3, and 5 years. In order to test the extrapolation ability of our signature, we performed qRT-PCR experiments on 146 MSI-H FFPE CRC tissues, and then validated the model according to its results, which showed that our signature is reliable and has high practicability in different cohorts. Besides, as we have previously demonstrated, patients with a score showed poor prognosis and RFS.

As far as we know, no one has yet developed prognostically relevant gene markers for non–MSI-H/pMMR CRC patients, and our signature fills the gap in this field. Prior to us, investigators have developed a number of biomolecular markers to predict the clinical outcomes of CRC patients (Jorissen et al., 2009; Salazar et al., 2011; Li et al., 2020). Compared with these studies, our signature has several advantages and innovations: 1) fewer genes have been included in the signature, making MSSAS easier to implement; 2) a variety of statistical algorithms were performed to estimate the calibration and discrimination of our signature. This maintained robust and powerful results at 1, 3, and 5 years; 3) external validation was performed using qRT-PCR experiments to confirm the accuracy and extrapolation ability of MSSAS. Despite the great potential clinical application value of MSSAS signature, it also has some limitations, for example, a prospective multicenter cohort is still necessary to further verify its performance.

In summary, using a systematic and comprehensive biomarker discovery and validation approach, we developed and experimentally validated a six-genes signature with a stable and powerful performance in evaluating the prognosis of non–MSI-H/pMMR CRC. It has demonstrated that our MSSAS signature might be a promising biomarker to facilitate the clinical management for non–MSI-H/pMMR CRC patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, and further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The human cancer tissues used in this study were approved by the Ethics Committee of The First Affiliated Hospital of Zhengzhou University. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Our work was designed by ZL and XH. The data were integrated and analyzed by ZL, XG, and XH. This manuscript was written by HX, XG, SW, and QD. ZL, XH, and ZL edited and revised this manuscript. All authors agree to publish this manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (81972663); Henan Province Young and Middle-Aged Health Science and Technology Innovation Talent Project (YXKC2020037); and Henan Provincial Health Commission Joint Youth Project (SB201902014).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.790214/full#supplementary-material

REFERENCES

Biller, L. H., and Schrag, D. (2021). Diagnosis and Treatment of Metastatic Colorectal Cancer. *JAMA* 325 (7), 669–685. doi:10.1001/jama.2021.0106

Carethers, J. M., and Doubeni, C. A. (2020). Causes of Socioeconomic Disparities in Colorectal Cancer and Intervention Framework and Strategies. *Gastroenterology* 158 (2), 354–367. doi:10.1053/j.gastro.2019.10.029

Davalos, V., Dopeso, H., Velho, S., Ferreira, A. M., Cirnes, L., Díaz-Chico, N., et al. (2007). High EPHB2 Mutation Rate in Gastric but Not Endometrial Tumors with Microsatellite Instability. *Oncogene* 26 (2), 308–311. doi:10.1038/sj.onc.1209780

Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M., and Wallace, M. B. (2019). Colorectal Cancer. *The Lancet* 394 (10207), 1467–1480. doi:10.1016/S0140-6736(19)32319-0

Endo, E., Okayama, H., Saito, K., Nakajima, S., Yamada, L., Ujiie, D., et al. (2020). A TGFβ-dependent Stromal Subset Underlies Immune Checkpoint Inhibitor Efficacy in DNA Mismatch Repair-Deficient/Microsatellite Instability-High Colorectal Cancer. *Mol. Cancer Res.* 18 (9), 1402–1413. doi:10.1158/1541-7786.MCR-20-0308

Eng, C., Kim, T. W., Bendell, J., Argilés, G., Tebbutt, N. C., Di Bartolomeo, M., et al. (2019). Atezolizumab with or without Cobimetinib versus Regorafenib in Previously Treated Metastatic Colorectal Cancer (IMblaze370): a Multicentre, Open-Label, Phase 3, Randomised, Controlled Trial. *Lancet Oncol.* 20 (6), 849–861. doi:10.1016/S1470-2045(19)30027-0

Ganesh, K., Stadler, Z. K., Cercek, A., Mendelsohn, R. B., Shia, J., Segal, N. H., et al. (2019). Immunotherapy in Colorectal Cancer: Rationale, Challenges and Potential. *Nat. Rev. Gastroenterol. Hepatol.* 16 (6), 361–375. doi:10.1038/s41575-019-0126-x

Jorissen, R. N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., et al. (2009). Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clin. Cancer Res.* 15 (24), 7642–7651. doi:10.1158/1078-0432.ccr-09-1431

Kather, J. N., Halama, N., and Jaeger, D. (2018). Genomics and Emerging Biomarkers for Immunotherapy of Colorectal Cancer. *Semin. Cancer Biol.* 52 (Pt 2), 189–197. doi:10.1016/j.semcancer.2018.02.010

Keskin, T., Rucci, B., Cornaz-Buros, S., Martin, P., Fusco, C., Broye, L., et al. (2021). A Live Single-Cell Reporter Assay Links Intratumor Heterogeneity to Metastatic Proclivity in Ewing Sarcoma. *Sci. Adv.* 7 (27), 7. doi:10.1126/sciadv.abf9394

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., et al. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* 372 (26), 2509–2520. doi:10.1056/NEJMoa1500596

Li, J., Zhang, J., Hu, H., Cai, Y., Ling, J., Wu, Z., et al. (2020). Gene Expression Signature to Predict Prognosis and Adjuvant Chemosensitivity of Colorectal Cancer Patients. *Cmar* Vol. 12, 3301–3310. doi:10.2147/CMAR.S243490

Lin, X., Zhuang, S., Chen, X., Du, J., Zhong, L., Ding, J., et al. (2021). lncRNA ITGB8-AS1 Functions as a ceRNA to Promote Colorectal Cancer Growth and Migration through Integrin-Mediated Focal Adhesion Signaling. *Mol. Ther.* S1525-0016, 00405–00406. doi:10.1016/j.ymthe.2021.08.011

Liu, Z., Lu, T., Li, J., Wang, L., Xu, K., Dang, Q., et al. (2021). Development and Clinical Validation of a Novel Six-Gene Signature for Accurately Predicting the Recurrence Risk of Patients with Stage II/III Colorectal Cancer. *Cancer Cel Int* 21 (1), 359. doi:10.1186/s12935-021-02070-z

Liu, Z., Lu, T., Li, J., Wang, L., Xu, K., Dang, Q., et al. (2021). Clinical Significance and Inflammatory Landscape of aNovel Recurrence-Associated Immune Signature in Stage II/III Colorectal Cancer. *Front. Immunol.* 12, 702594. doi:10.3389/fimmu.2021.702594

Liu, Z., Lu, T., Wang, Y., Jiao, D., Li, Z., Wang, L., et al. (2021). Establishment and Experimental Validation of an Immune miRNA Signature for Assessing Prognosis and Immune Landscape of Patients with Colorectal Cancer. *J. Cel Mol Med* 25 (14), 6874–6886. doi:10.1111/jcmm.16696

Liu, Z., Zhang, Y., Dang, Q., Wu, K., Jiao, D., Li, Z., et al. (2021). Genomic Alteration Characterization in Colorectal Cancer Identifies a Prognostic and Metastasis Biomarker: FAM83A|Ido1. *Front. Oncol.* 11, 632430. doi:10.3389/fonc.2021.632430

Lizardo, D. Y., Kuang, C., Hao, S., Yu, J., Huang, Y., and Zhang, L. (2020). Immunotherapy Efficacy on Mismatch Repair-Deficient Colorectal Cancer: From Bench to Bedside. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1874 (2), 188447. doi:10.1016/j.bbcan.2020.188447

Maiuthed, A., Ninsontia, C., Erlenbach-Wuensch, K., Ndreshkjana, B., Muenzner, J., Caliskan, A., et al. (2018). Cytoplasmic P21 Mediates 5-Fluorouracil Resistance by Inhibiting Pro-apoptotic Chk2. *Cancers* 10 (10), 373. doi:10.3390/cancers10100373

Owonikoko, T. K., Zhang, G., Deng, X., Rossi, M. R., Switchenko, J. M., Doho, G. H., et al. (2014). Poly (ADP) Ribose Polymerase Enzyme Inhibitor, Veliparib, Potentiates Chemotherapy and Radiation *In Vitro* and *In Vivo* in Small Cell Lung Cancer. *Cancer Med.* 3 (6), 1579–1594. doi:10.1002/cam4.317

Picard, E., Verschoor, C. P., Ma, G. W., and Pawelec, G. (2020). Relationships between Immune Landscapes, Genetic Subtypes and Responses to Immunotherapy in Colorectal Cancer. *Front. Immunol.* 11, 369. doi:10.3389/fimmu.2020.00369

Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., et al. (2011). Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer. *Jco* 29 (1), 17–24. doi:10.1200/JCO.2010.30.1077

Shen, P., Yang, S., Sun, H., Li, G., Wu, B., Ji, F., et al. (2018). SCF/c-KIT Signaling Increased Mucin2 Production by Maintaining Atoh1 Expression in Mucinous Colorectal Adenocarcinoma. *Ijms* 19 (5), 1541. doi:10.3390/ijms19051541

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statist. Med.* 16 (4), 385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385:aid-sim380>3.0.co;2-3

Vilar, E., and Gruber, S. B. (2010). Microsatellite Instability in Colorectal Cancer-The Stable Evidence. *Nat. Rev. Clin. Oncol.* 7 (3), 153–162. doi:10.1038/nrclinonc.2009.237

Webber, E. M., Kauffman, T. L., O'Connor, E., and Goddard, K. A. (2015). Systematic Review of the Predictive Effect of MSI Status in Colorectal Cancer Patients Undergoing 5FU-Based Chemotherapy. *BMC Cancer* 15, 156. doi:10.1186/s12885-015-1093-4

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. Article17. doi:10.2202/1544-6115.1128

Zhang, J.-Z., Wu, Z.-H., and Cheng, Q. (2019). Screening and Identification of Key Biomarkers in Nasopharyngeal Carcinoma. *Medicine (Baltimore)* 98 (48), e17997. doi:10.1097/MD.0000000000017997

Zhang, Y., Liu, Z., Li, X., Liu, L., Wang, L., Han, X., et al. (2021). Comprehensive Molecular Analyses of a Six-Gene Signature for Predicting Late Recurrence of Hepatocellular Carcinoma. *Front. Oncol.* 11, 732447. doi:10.3389/fonc.2021.732447

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Liu, Xu, Ge, Weng, Dang and Han. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 23 February 2022
doi: 10.3389/fcell.2022.828415

Based on Molecular Subtypes, Immune Characteristics and Genomic Variation to Constructing and Verifying Multi-Gene Prognostic Characteristics of Colorectal Cancer

Lei Gu[†], *Chunhui Jiang*[†], *Chunjie Xu, Ye Liu and Hong Zhou* *

Department of Gastrointestinal Surgery, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Christina Stuelten,
National Cancer Institute (NIH), United States
Ming Jun Zheng,
ShengJing Hospital of China Medical University, China

***Correspondence:**
Hong Zhou
renjizhouhong@163.com

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *03 December 2021*
Accepted: *31 January 2022*
Published: *23 February 2022*

Citation:
Gu L, Jiang C, Xu C, Liu Y and Zhou H (2022) Based on Molecular Subtypes, Immune Characteristics and Genomic Variation to Constructing and Verifying Multi-Gene Prognostic Characteristics of Colorectal Cancer. Front. Cell Dev. Biol. 10:828415. doi: 10.3389/fcell.2022.828415

Background: Colon cancer (COAD) has been identified as being among the most prevalent tumors globally and ranked the third major contributor to cancer-related mortality. COAD is a molecularly heterogeneous disease. There are great differences in clinical manifestations and prognosis among different molecular subtypes.

Methods:379 TCGA-COAD samples were divided into four subtypes: primary proliferative, with collective, crypt-like, and EMT invasion. The differences among the four subtypes were analyzed from the multidimensional perspectives of immunity, genomic variation, and prognosis. The limma package was utilized to identify differentially expressed genes (DEGs) amongst different molecular subtypes. Phenotype-related coexpressed gene modules were identified using WGCNA. The polygenic prognosis model was created utilizing the lasso Cox analysis and verified by time-dependent subject operating characteristics (ROC).

Results: There are some differences in prognosis, TMB and common gene variation, immune score, and immunotherapy/chemotherapy between proliferative and three invasive molecular subtypes. 846 differential genes (DEGs) were obtained by limma packet analysis. Differential gene analysis was utilized to screen the DEGs among distinct subtypes, which were significantly enriched in the pathways related to tumorigenesis and development. Co-expression network analysis found 46 co-expressed genes correlated with proliferative and three invasive phenotypes. Based on differentially co-expressed genes, we developed a prognostic risk model of 8-genes signature, which exhibited strong stability regardless of external and internal validation. RT-PCR experiments proved the expression of eight genes in tumor and normal samples.

Conclusion: We have developed an eight-gene signature prognostic stratification system. Furthermore, we proposed that this classifier can serve as a molecular diagnostic tool to assess the prognosis of colon cancer patients.

Keywords: COAD, molecular subtype, TCGA, multi-gene signature, prognosis

INTRODUCTION

Colon cancer is extensively recognized to be among the most prevalent tumors globally and has been ranked as the third major contributor to cancer mortality (Labianca et al., 2010). COAD is the main pathological type of colon cancer. A growing body of data indicates that COAD is a molecularly diverse illness in which unique molecular alterations influence the growth and survival of tumor cells, as well as their differentiation, apoptosis, and distant metastasis (Comprehensive molecular c, 2012). The conventional TNM staging may not be able to evaluate its metastatic potential because of the unique heterogeneity of COAD, and its staging may not be completely positively correlated with survival (Chu et al., 2016). Although the survival rate of COAD patients has been improved with the increase of treatment methods, the prognosis is still very poor (Neri et al., 2010; Roncucci and Mariani, 2015).In addition to targeted therapy, immunotherapy seems to be a promising treatment for advanced CRC. Recent research has demonstrated that immunotherapy could be beneficial for colon cancer patients who have DMMR/MSI-H (Asaoka et al., 2015; Overman et al., 2017), and the benefit is not obvious in MSS/MSI-L patients. Unfortunately, the proportion of DMMR/MSI-H in metastatic colon cancer is only about 5% (Jung et al., 2020), and the effective rate in such patients is only 30–40% (Le et al., 2017), which indicates that there are some limitations in the application of MSI status as an immune checkpoint inhibitor. On the other hand, it shows that due to the obvious heterogeneity of advanced CRC, a single molecular expression state may not be enough to reflect the information of the whole tumor. Therefore, it is urgent to explore a reliable biomarker or prognostic system, and further clarify the specific information of COAD typing, so as to provide the basis for clinical individualized treatment.

According to the spontaneous model of canine colorectal cancer and the characteristics of human CRC, Shayingzhao et al. divided CRC into the proliferative type and metastatic type, in which the metastatic type was specifically divided into collective type, crypt-like type, and EMT. These four subtypes have their unique biological behavior, pathway activation, and molecular mutation. Proliferative type shows that abnormal Wnt/β-catenin signaling pathway activation results in cell cycle and proliferation, which is the most significant feature of proliferative tumors. In a series of mechanisms, epithelial cells may be changed into cells with a mesenchymal phenotype in a process known as epithelial-mesenchymal transformation (EMT). This transformation makes it easier for tumor cells to separate from the primary tissue and metastasize (Thiery, 2002; Friedl and Gilmour, 2009). Collective invasion is defined as the migration of a group of cells while maintaining intercellular contact. These cells are usually epithelial. CryptLike invasion refers to the CryptLike invasion of cancer cells through CryptLike structures. At present, the research is limited. Usually, these cancer cells are MYC positive, similar to crypt stem cells or progenitor cells. The researchers believe that due to the significant proliferation of fibroblasts, they can develop crypts in non-mucosal sites, making the microenvironment more like mucosa (Wang et al., 2018).In this study, only the pathway and some specific molecular mutation characteristics of each subtype were preliminarily analyzed, but there was no overall comprehensive analysis of the prognostic characteristics and potential therapeutic targets of each subtype.

In this study, according to the molecular typing determined by Zhao et al., we analyzed the differences between the proliferative and three invasive molecular subtypes in prognosis, TMB and common gene variation, immune score, and the efficacy of immunotherapy/chemotherapy.The co-expressed genes related to proliferative phenotype and three invasive phenotypes were identified by co-expression network analysis. Based on the differentially co-expressed genes, we built the prognostic risk model of 8-gene signature (SPARCL1, HAND1, CRIP2, ZNF385A, CXCL1, CLEC10A, PTGS1, and PTN). Regardless of external and internal validation, the 8-gene signature has strong stability, indicating that our risk prediction model can play a stable predictive effect. Based on this, we provide a stable prognostic model for colorectal cancer and provide a basis for individualized treatment of different subtypes of COAD.

MATERIALS AND METHODS

Data Sources

TCGA GDC API was employed to obtain the RNA SEQ data of TCGA-COAD. After screening, we included a sum of 343 samples. In addition, we downloaded the GSE17538 chip data set in conjunction with survival duration from the Gene Expression Omnibus (GEO) database and finally included 226 samples. The subtypes of TCGA-COAD were screened from the **Supplementary Materials** studied by Shaying Zhao et al. (Wang et al., 2018), after deleting redundant samples through the sample information in the 'case' column, a total of 366 samples were divided into four subtypes, of which 155 samples were collective subtypes, 76 samples were crypt-like subtypes, 62 samples were EMT subtypes, and 73 samples were identified as proliferative subtypes. As shown in **Supplementary Table S1**.

Data Preprocessing

TCGA Data Preprocessing

The RNA-seq data of TCGA were preprocessed in the steps below to obtain 343 colon cancer samples:

1) Extraction of primary colon cancer samples.
2) Patients having a survival duration of more than 30 days and a good survival status were included.
3) Patients with molecular typing results in the study of Shaying Zhao et al.
4) The expression profiles of 25483 genes were obtained by matching ENSG to GeneSymbol.

GEO Data Preprocessing

Preprocess geo dataset GSE17538 in several steps:

1) Download the standardized probe expression profile data of GSE17538 chip data from the GEO database;
2) According to the annotation file of the platform corresponding to each chip data, convert the expression profile at the probe level into the expression profile at the gene level. In this regard, when several probes correspond to the very same gene, the gene expression should be calculated

using the average of the multiple probes, and the probe should be removed in the case where only one probe corresponds to multiple genes;

3) The samples of primary colon cancer were extracted, and the patients having a survival duration of more than 30 days and a good survival status were included in the present study. Finally, GSE17538 included 229 samples.

Relationship Between Proliferative and Invasive Molecular Subtypes and TMB and Common Gene Variants

Further, we explored whether there were differences in genomic data between molecular subtypes. We acquired the mutation data set and copy number variation data processed by mutect2 software of TCGA-COAD. Then, the Fisher test was employed to screen the genes that have a substantial mutation, significant copy deletion, and significant copy amplification differences in each subtype.

Differential Analysis of Proliferative and Invasive Molecular Subtypes in the Efficacy of Immunotherapy/ Chemotherapy

We examined the differences among distinct molecular subtypes in immunotherapy and chemotherapy. TIDE(http://tide.dfci.harvard.edu/) Software was utilized to evaluate the potential clinical effects of immunotherapy in proliferative and invasive molecular subtypes. The higher the tide prediction score, the higher the possibility of immune escape, suggesting that the patient is less likely to benefit from immunotherapy.

Identify Phenotype Related Coexpressed Gene Modules

We used the WGCNA module in R software(Langfelder and Horvath, 2008) to detect phenotype-related co-expression modules. Specifically, we selected TCGA expression profile data set and screened genes having MAD more than 50 percent as gene expression profiles. Firstly, we clustered the samples together and screened them for co-expression modules. In the study, we discovered that the co-expression network corresponds to the scale-free network, indicating that the log(k) of the node with linkage degree k is inversely associated with the log(P (k), which is the probability of the node, and the correlation coefficient is higher than 0.85. We then transformed an adjacency matrix from an expression matrix, and subsequently transformed the adjacency matrix into a topology matrix. Subsequently, for the purpose of clustering genes, we utilized the average linkage hierarchical clustering technique as per the standards of a hybrid dynamic cut tree with the aid of Tom. We also specified 30 as the limit quantity of genes present in each gene network module. After determining the gene modules utilizing the dynamic cutting approach, we computed the eigengenes of each module one at a time and clustered the modules together utilizing the eigengenes.

Constructing a Prognostic Risk Model on the Basis of Subtype Differential Expression and Co-expressed Genes

To begin with, the 343 samples in the TCGA data set are classified into two groups: the training set as well as the verification set. As a precaution against random allocation variation having an adverse effect on the stability of succeeding modeling, all samples are clustered at random 100 times before being used. In this case, the grouping of samples is conducted in accordance with the ratio of 1:1 for the training set to the verification set. Furthermore, utilizing the training set data, the univariate Cox proportional hazards regression model was performed on differentially co-expressed genes utilizing the survival coxph function in the R-package. The level of $p < 0.05$ was used as the cutoff value for filtering in order to get the genes associated with prognosis.

To significantly reduce the gene range while retaining high accuracy, we performed additional experiments. We employed lasso regression to additionally narrow the prognostic genes, resulting in a reduction in the proportion of genes in the risk model. The Lasso technique is a kind of estimation that uses compression. When it constructs a penalty function, it results in certain coefficients being compressed while others are being set to zero, producing a more refined model. In this way, it keeps its advantages in terms of subset shrinking. It is a skewed estimating method for dealing with complicated collinearity data sets. When assessing variables, it may be used to optimize the selection of parameters, and it could be used to effectively handle the issue of multi-collinearity in regression analysis. The lasso Cox regression analysis was performed utilizing the R software program glmnet (Friedman et al., 2010). In this regard, we investigated the transformation trajectory of every independent variable separately. In addition, we employed a 10-fold cross-validation procedure to construct the model.

Univariate and Multivariate Cox Analysis and Establishment of a Nomogram

In the TCGA dataset, univariate and multivariable Cox regression analyses were utilized to identify the clinical independence regarding the model. Nomogram is a method that can intuitively and effectively present the data regarding the risk model and may be conveniently applied in predicting the outcomes. The nomogram makes use of the length of the straight line to depict the effect of several factors on the result, as well as the effect of various variable values on the outcomes. To enhance the predictive performance, we integrated multi-factor meaningful variables and established a new nomogram using the Cox model. Additionally, the calibration curve was utilized to determine its predictive performance. Finally, we performed DCA (Decision curve analysis) to determine its credibility.

Cell Culture

The Chinese Academy of Sciences (Shanghai, China) provided the NCM460 cell line and the human colon cancer cell lines including LoVo, HT29, and HCT116. All the obtained cell lines were grown in RPMI 1640 accompanied by 10 percent fetal bovine serum.

RNA Isolation and RT-PCR Analysis

The TRIzol reagent (Invitrogen, Carlsbad, CA, United States) was utilized to extract RNA from the tissue samples. Subsequently, QuantiTect Reverse Transcription Kit (Qiagen, Valencia, CA, United States) was employed to convert the isolated RNA into cDNA. SYBR Green (Takara, Otsu, Shiga, Japan) was employed to quantify the results of real-time PCR analyses, and the levels were subsequently standardized to GAPDH levels. The primers used in the upstream and downstream experiments have the following sequences: SPARCL1-forward: 5′-CAACTGCTGAAACGG TAGCA-3'; SPARCL1-reverse 5′-GAACTCTTGCCCTGTTCTGC-3'; HAND1-forward: 5′- AGCCACCAGCTACATCGCCTAC-3'; HAND1-reverse: 5′- GCGATCCGCCTTCTTGAGTTC-3'; CLEC10A-Forward: 5′- TACACCTGGATGGGCCTCAG -3'; CLEC10A- reverse: 5′- TGTTCCATCCACCCACTTCC -3; PTGS1- forward: 5′- ATCGCCATGGAATTCAACCA-3′; PTGS1- reverse: 5′- GTGAGCCCACTTGGAAGGAA -3′; PTN- forward: 5′- CCATTTCCCTTCCGTTCC-3'; PTN - reverse: 5′- AGGTTGCTACCGCTGAGTCC -3′;CXCL1- forward: 5′- CTCGAGGCCCCTGGGGCAGAAGCCTC-3′; CXCL1- reverse: 5′- GATATCGGGGCTCAGCAGGCGGGTCT -3′; CRIP2- forward: 5′- ACTGATGCCTCCTCACCATC-3′; CRIP2- reverse: 5′- TGTTTGTGAGCCAACCAGAG-3′;

RESULTS

Relationship Between Proliferative and Invasive Molecular Subtypes and Prognosis of Colon Cancer

We analyzed the relationship between proliferative and invasive molecular subtypes and patient prognosis. First, we examined the correlation between four molecular subtypes and the OS prognosis. In **Figure 1A**, it can be observed that there exist marginal considerable differences among the four molecular subtypes (p = 0.075). Among the three invasive molecular subtypes, EMT and CryptLike subtypes have the worst prognosis, and Collective has the best prognosis; additionally, we evaluated the correlation between the four subtypes and PFS and found the same findings, that is, EMT and CryptLike subtypes exhibited the most unfavorable prognosis, and Collective exhibited favorable prognosis (**Figure 1B**); Finally, we analyzed the relationship with DSS. EMT and CryptLike exhibited the most unfavorable prognosis, and Collective exhibited an improved prognosis (**Figure 1C**). Our results show that there are invasion-related phenotypes in colon cancer, in which EMT and CryptLike are more aggressive.

Relationship Between Proliferative and Invasive Molecular Subtypes and TMB and Common Gene Variants

We acquired the mutation data set as well as copy number variation data processed by mutect2 software of TCGA-COAD. Then, we used the Fisher test to screen the genes with significant mutation, significant copy deletion, and significant copy amplification differences in each subtype. The mutation characteristics of top10 in each subtype are as shown in **Figures 2A–D**. It can be observed that the copy number amplification of DCAF12, AK3, ARHGEF6, GLIS3, MAGEA3, RCL1, RLN1, SLC9A6, and SPANXB1 genes in EMT molecular subtypes are substantially elevated in contrast with that in other subtypes.

Difference in Immune Score Between Proliferative and Invasive Molecular Subtypes

From the TCGA dataset, the proportion of 22 distinct immune cells in each sample was evaluated by the CIBERSORT method, and the distribution of these immune cell proportions in the four molecular subtypes was as shown in Supplement **Figure 1A**. The box diagram of immune cells is shown in **Figure 3A**, and the heat diagram of immune cells is shown in Supplement **Figure 1B**. It indicated that the Collective, CryptLike, and EMT in macrophases_ M1, Macrophages_ M2 was significantly higher than the proliferative molecular subtype. M1 macrophages play an anti-tumor role by activating the immune system and releasing tumor necrosis factor, nitric oxide, and reactive oxygen species, while macrophages M2 is the main participant and coordinator in promoting tumor progression in the tumor microenvironment(Atri et al., 2018). Immune infiltration analysis showed that CryptLike and EMT molecular subtypes had high immune microenvironment infiltration, and proliferative had the lowest immune microenvironment infiltration. Interestingly, CryptLike and EMT have the highest immune infiltration scores, but the prognosis is the worst, shown in **Figures 3B–D**.

Differential Analysis of Proliferative and Invasive Molecular Subtypes in the Efficacy of Immunotherapy/Chemotherapy

We use TIDE (http://tide.dfci.harvard.edu/) Software to evaluate the potential clinical effects of immunotherapy in proliferative and invasive molecular subtypes. The higher the TIDE prediction score, the higher the possibility of immune escape, suggesting that patients are less likely to benefit from immunotherapy. As illustrated in **Figure 4A**, we discovered that there were substantial differences in TIDE scores between distinct molecular subtypes. The TIDE score of the proliferative subtype is the lowest, which may benefit more from immunotherapy. At the same time, we also compared the predicted T cell rejection score and T cell dysfunction score in different molecular subtypes. As depicted in **Figures 4B,C**. it is evident that CryptLike and EMT have elevated T cell rejection scores and T cell dysfunction scores than collective and proliferative. It may be that CryptLike and EMT

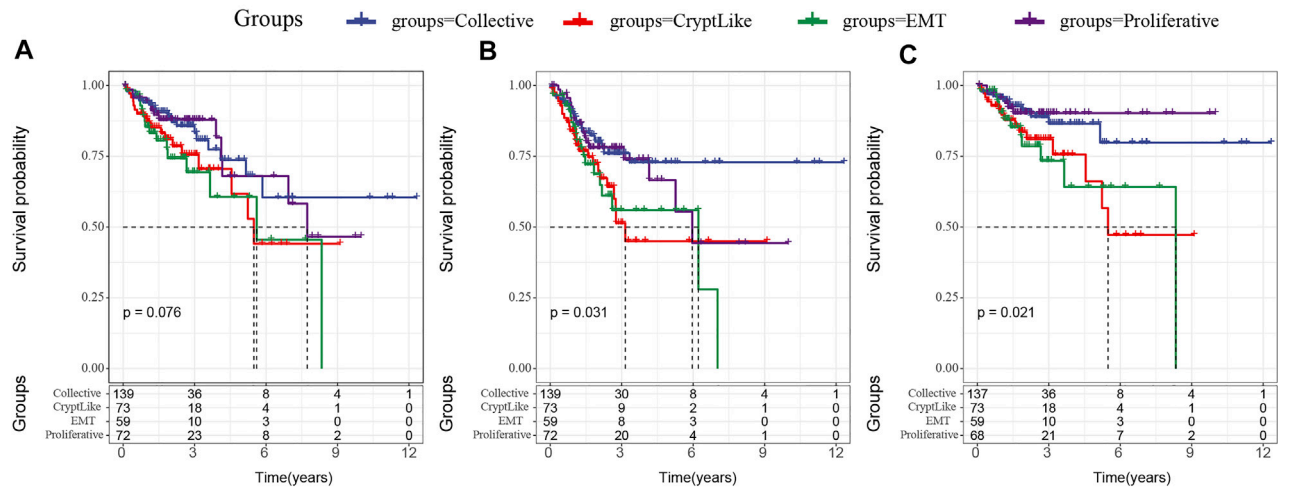


FIGURE 1 | (A): Km curve of the relationship between four subtypes and OS; **(B)**: Km curve of the relationship between four subtypes and PFS; **(C)**: Km curve of the relationship between four subtypes and DSS.

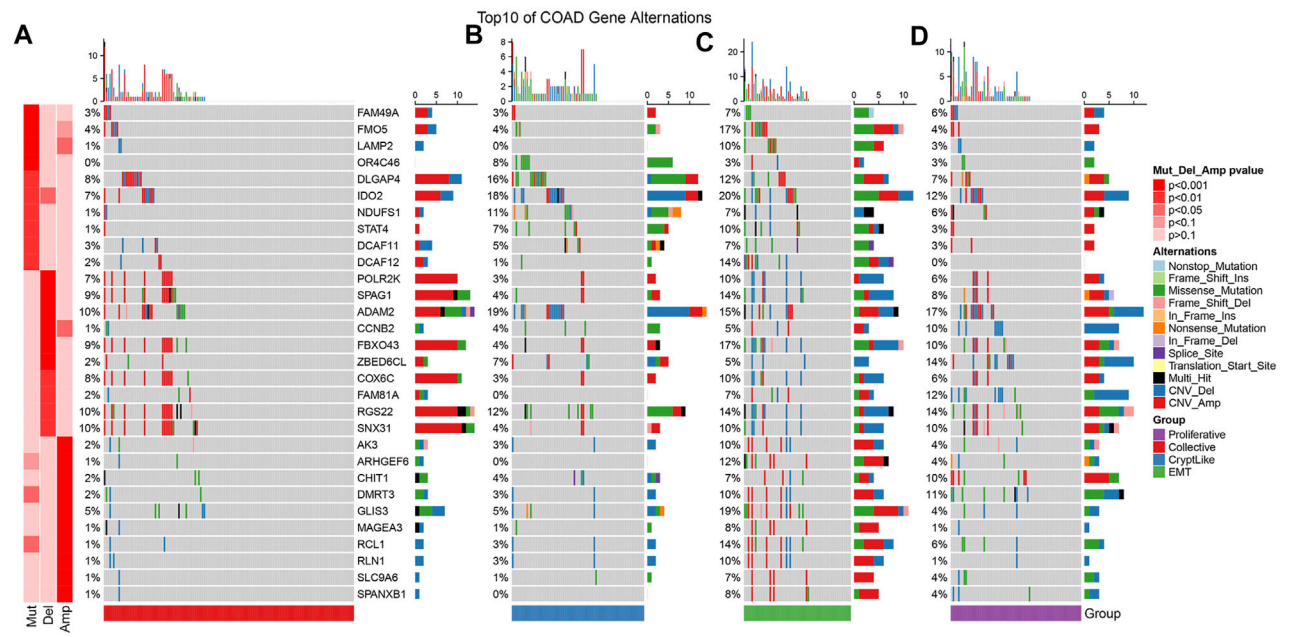


FIGURE 2 | Mutation features of the topmost 10 significantly mutated genes from each subtype sample. **(A–D)** represent collective, cryptlike, EMT, and proliferative subtypes respectively.

have high immune infiltration scores, but there are T cell dysfunction and T cell rejection, resulting in a poor prognosis of CryptLike and EMT. The distribution difference of different response states to immunotherapy predicted by tide software in different immune molecular subtypes is illustrated in **Figure 4D**. We discovered that the proportion of response to immunotherapy in collective and proliferative subtypes is substantially elevated as opposed to that in CryptLike and EMT subtypes. In addition, we evaluated the response of distinct subtypes to conventional chemotherapeutic drugs 5-FU and cisplatin. It was found that Collective and Proliferative subtypes were more sensitive to cisplatin than other subtypes (**Figure 4E**).

Analysis of Differentially Expressed Genes Among Subtypes

The DEGs among Collective ~ Other, CryptLike ~ Other, EMT ~ Other. and Proliferative-Other were obtained utilizing the limma package. Filtering was carried out according to the thresholds of | log2fc | > 1 and FDR <0.05. Collective ~ Other up-regulated and

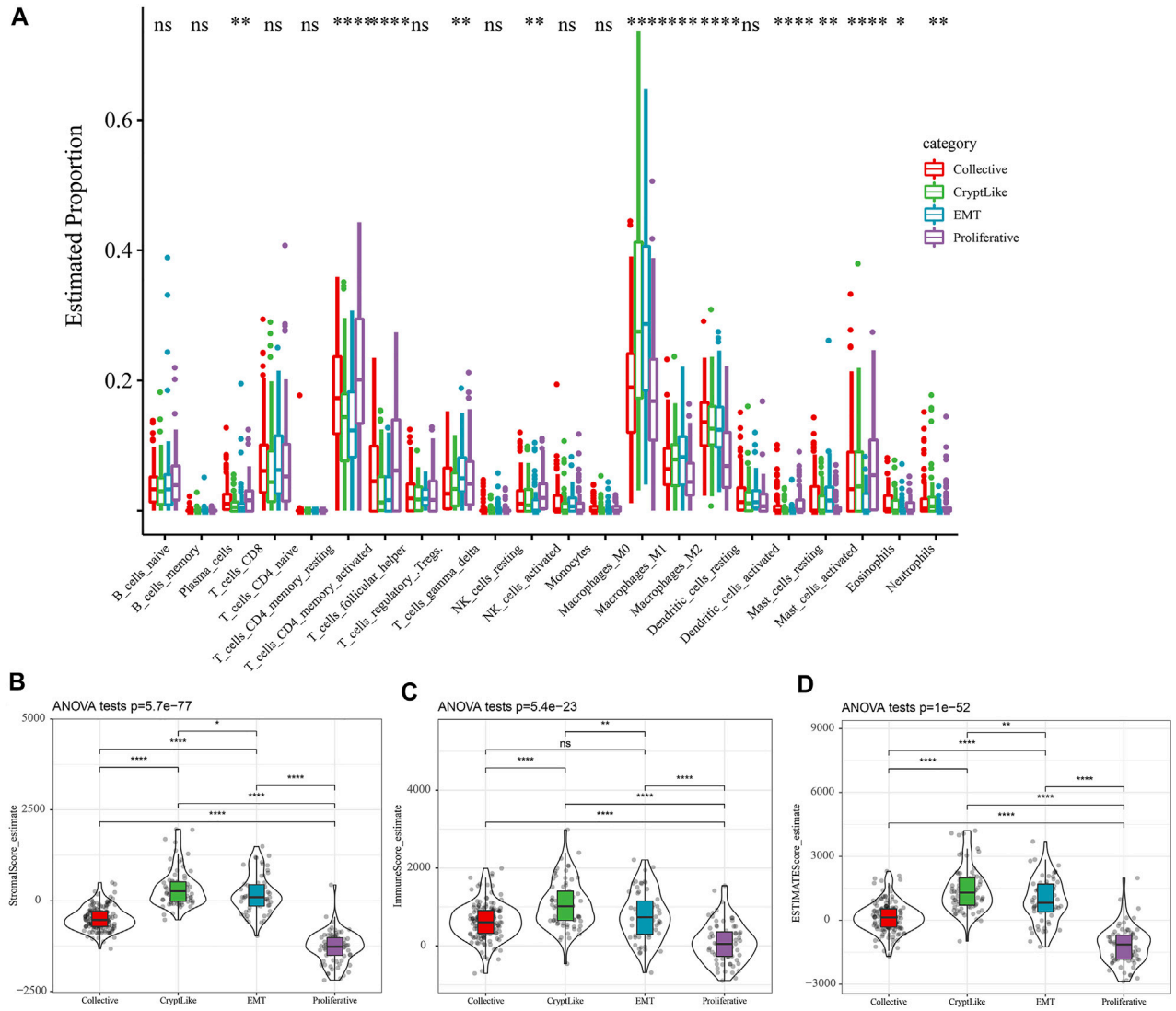


FIGURE 3 | (A): The difference of 22 immune cell components; **(B)**: Differences in matrix scores in different subtypes; **(C)**: The scores of immune infiltration in different subtypes were different; **(D)**: The estimate scores of different subtypes were different.

down-regulated DEGs are illustrated as the volcano figure as shown in **Figure 5A**, which include 11 down-regulated and 28 up-regulated genes, indicating that Collective ~ Other is mainly up-regulated; CryptLike ~ Other up-regulated and down regulated DEGs are illustrated as the volcano figure as shown in **Figure 5B**, which include one down-regulated gene and 394 up-regulated genes, indicating that CryptLike ~ Other is mainly up-regulated; EMT ~ Other down-regulated and up-regulated DEGs are illustrated as the volcano figure as shown in **Figure 5C**, which include 58 down-regulated genes and 232 up- regulated genes. The results show that EMT ~ Other is mainly up-regulated; The down- regulated and up-regulated DEGs of Proliferative-Other are shown in **Figure 5D**, including 730 down- regulated genes and one up- regulated gene. The results show that Proliferative-Other is mainly down-regulated. The heat map of DEGs among subtypes is shown in **Figure 5E**. Furthermore, we performed functional enrichment analysis on 846 DEGs after eliminating redundancy of differential genes among the four subtypes via the clusterProfiler function in R software, and defined the cutoff value of FDR <0.05. The KEGG is enriched in 55 pathways, mainly Focal adhesion, Cytokine-cytokine receptor interaction, PI3K-Akt signaling pathway, Cell adhesion molecules (CAMs), NF-kappa B signaling pathway, Proteoglycans in cancer, and other pathways associated with tumor occurrence and progression as depicted in **Figure 5F**; There are 1165 GO-BP enrichment results, of which the first six GO-BP are shown in **Figure 5G**.

Identify Phenotype Related Coexpression Gene Modules

We employed the WGCNA module in R software to identify phenotype-related coexpression modules. Specifically, we

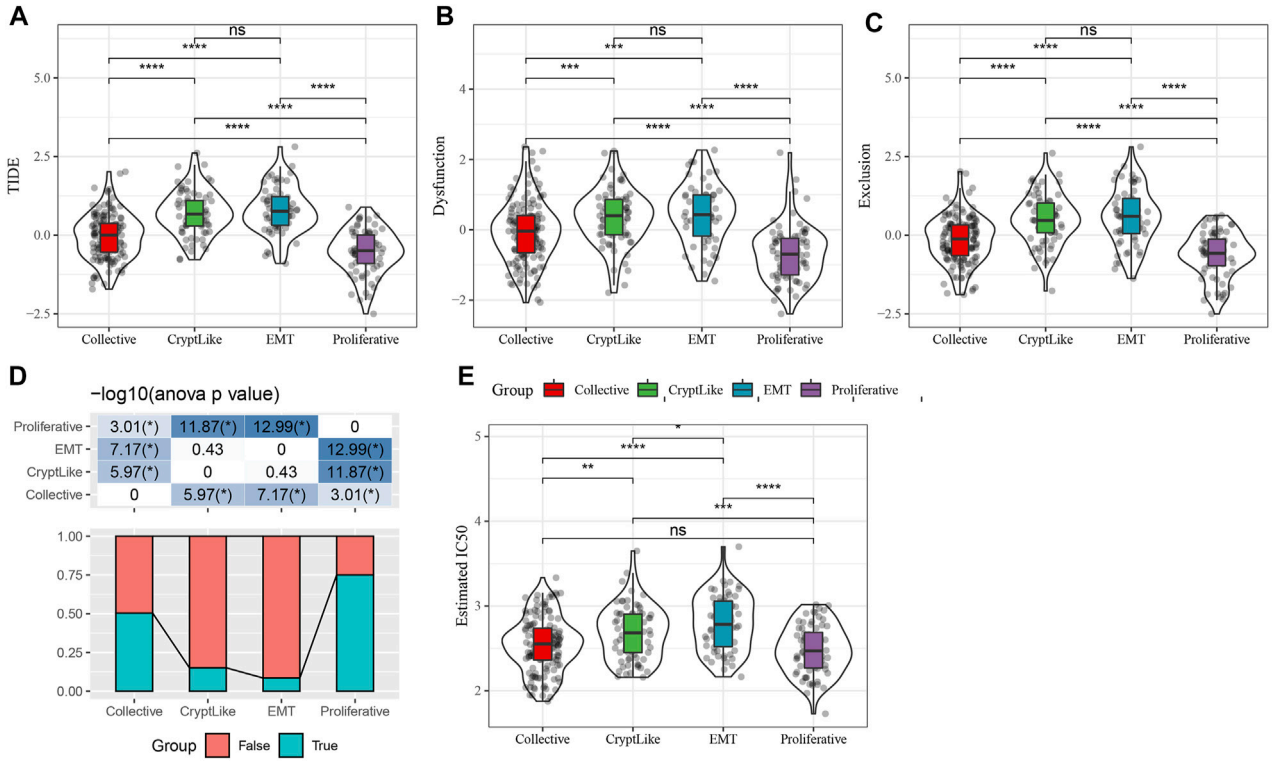


FIGURE 4 | (A): The tide scores of different molecular subtypes were different; **(B)**: T cell dysfunction scores of different molecular subtypes were different; **(C)**: The T cell rejection scores of different molecular subtypes were different; **(D)**: The difference of immune response of different molecular subtypes; **(E)**. The box plots represent the predicted IC50 inhibition concentration for Cisplatin.

selected TCGA expression profile data set and screened genes that had MAD of over 50 percent as gene expression profiles. Firstly, we clustered the samples and chose the soft threshold of 3 (**Figure 6A**). To guarantee a scale-free network, we choose β = 3(**Figures 6B–C**). Next, the adjacency matrix was generated by transforming the expression matrix. The adjacency matrix was subsequently turned into a topological matrix. To determine the gene module, we first computed the eigenvectors for all modules one by one. Then, we clustered the modules and merged those that are near to one another into newer modules. We next specified the minModuleSize = 30, DeepSplit = 2, and height = 0.25 parameters for the gene module. Finally, a sum of 17 modules was obtained (**Figure 6D**). Notably, however, the grey module is comprised of a gene set that cannot be combined with other modules. **Figure 6E** depicts the number of transcripts produced from each module, in which the grey module is a gene module that cannot be allocated. We examined the association between each module and the patients' M stage, gender, N stage, T stage, age, Stage, and proliferative and invasive molecular subtypes, as illustrated in **Figure 6F**. It is evident that greenyellow, purple, yellow and red modules have a substantial positive association with CryptLike and EMT, and an inverse association with collective and proliferative. The results of correlation analysis between GS and MM of genes in the module are shown in **Figures 6G–J** the results show that GS and MM of green, yellow, purple, yellow, and red modules are highly positively correlated.

Developing a Prognostic Risk Model According to Subtype Differential Expression and Co-expression Genes

Construction of Training Set Sample Risk Model

The 343 samples from the TCGA data set have been classified into two categories: the training set as well as the verification set. As a precaution against random allocation bias having an adverse effect on the stability of succeeding modeling, all samples were clustered at random 100 times before being used. To group the samples, a ratio of 1:1 was utilized for the training set to the verification set. The final training set comprises 171 samples, while the test set contains 172 samples.

Further, using the training set data, for the differentially co-expressed genes (399 in total) and survival data, the survival coxph function of the R-package was utilized to carry out a univariate Cox proportional hazards regression model. The filtration level was set at $p < 0.05$. Finally, there were 46 genes related to prognosis.

At present, 46 differentially co-expressed genes related to prognosis have been identified. Nevertheless, because of the substantial proportion of these genes and the inconducive impact on clinical detection, we must additionally reduce the

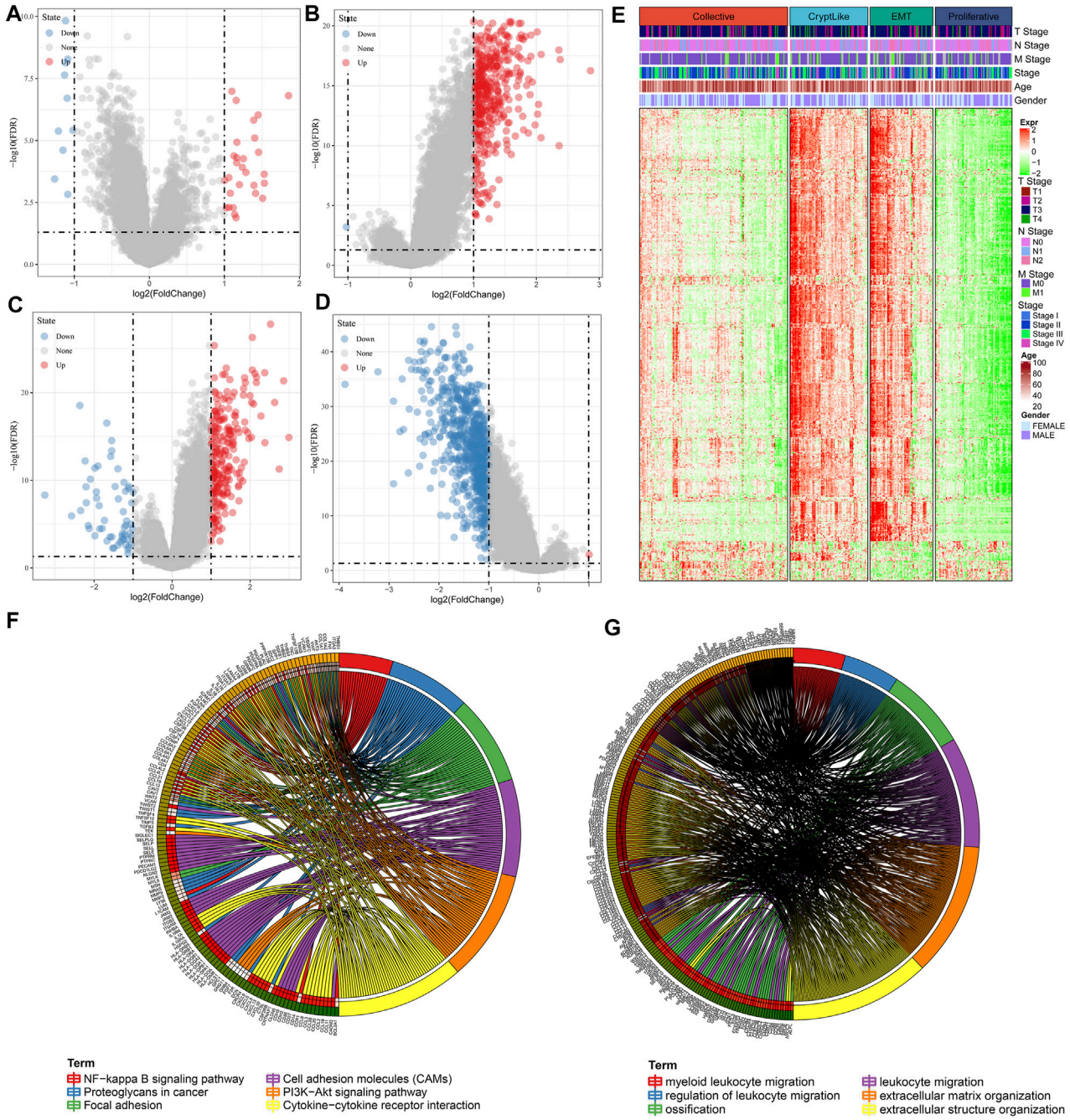


FIGURE 5 | (A): Differences in gene expression between Collective and non-Collective populations are depicted on a volcano map; **(B)**: CryptLike and non-CryptLike groups were compared using a volcano map to identify genes that differed; **(C)**: Volcano map of differential genes between EMT and non-EMT groups; **(D)**: Volcanic map of differential genes between Proliferative and non-Proliferative groups. **(E)**. Heat map of DEGs among subtypes; **(F)**: The results of KEGG enrichment of 846 differential genes; **(G)**: GO-BP enrichment results of 846 different genes, where distinct colors denote distinct pathways, and connecting lines denote genes correlated with pathways.

gene range while retaining high accuracy in the process. We further compressed these 46 genes utilizing lasso regression to minimize the proportion of the risk model genes. For the lasso Cox regression analysis, we utilized the glmnet package of R software. As illustrated in **Figure 7A**, we examined the change trajectory of each independent variable. It is evident that when lambda is progressively increased, the proportion of independent variable coefficients that are on the verge of zero grows progressively as well. To construct the model and examine the confidence interval within each lambda, we employed a 10-fold cross-validation procedure, as depicted in **Figure 7B**. It can be seen from the figure that when lambda = 0.03507542, the model

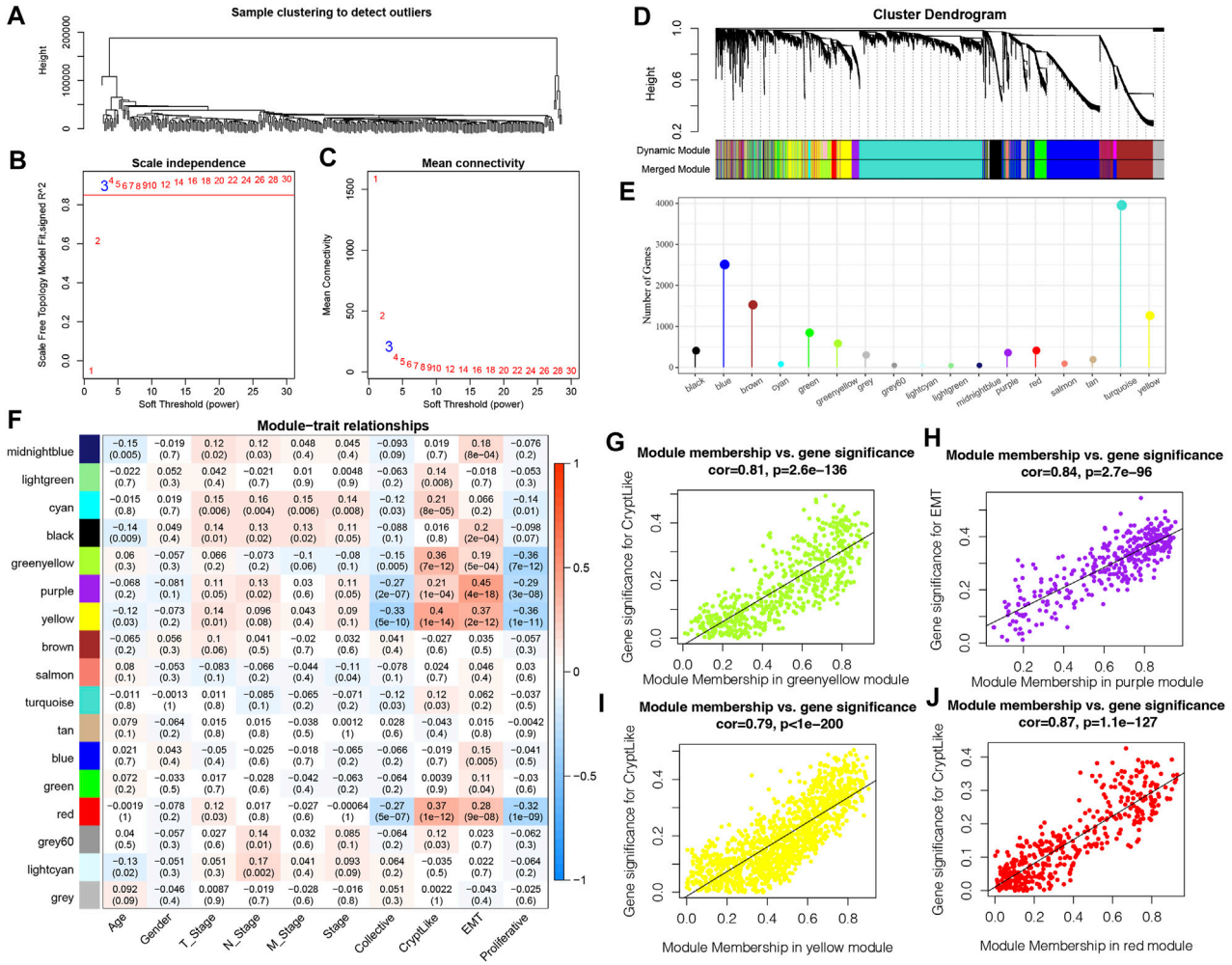


FIGURE 6 | (A): Clustering tree of each sample; **(B)**: An examination of the scale-free fit index for a variety of soft-threshold values (β). **(C)**: For different soft-threshold values, an analysis of the mean connectedness was carried out. **(D)**: Dendrogram illustrating all DEGs/lncRNAs clustered according to a dissimilarity measure (1-TOM); **(E)**: Statistics of the proportion of each module genes; **(F)**: Relationship between clinical information and each module; **(G)**: In the green-yellow module, a scatter plot showing the relationship between module membership and gene value for CryptLike is shown; **(H)**: In the purple module, the scatter graphic shows the relationship between module membership and gene importance for EMT; **(I)**: In the yellow module, a scatter plot showing the relationship between module membership and gene importance for CryptLike is presented; **(J)**: For CryptLike in the red module, the scatter graphic shows the relationship between module membership and gene importance.

reaches the optimum. In view of this, in the following stage, we picked 14 genes when lambda = 0.03507542, which were then utilized as targets for the subsequent step.

As part of its stepwise regression procedure, the AIC data criterion is utilized, which takes into account both the statistical fit as well as the total number of components employed for fitting. AIC is reduced using a stepping technique in the Stats package. We begin with the most complicated model and remove one variable at a time to lower AIC. The model performs better when the value is lower than a certain threshold, demonstrating that the model may get a satisfactory fit using fewer features. With the help of this algorithm, we were able to reduce 14 genes to only eight genes: SPARCL1, CXCL1, HAND1, CRIP2, CLEC10A, PTGS1, PTN, and ZNF385A.

Eventually, the signature equation consists of eight genes and is illustrated below: risk score = SPARCL1*1.8291043-CXCL1*0.4539938 + HAND1*0.4792704 + CRIP2*1.5098001-CLEC10A*1.2404183-PTGS1*1.2728945-PTN*2.2581275 + ZNF385A*0.6856336.

As illustrated in **Figure 7C**, we computed the risk score of each sample based on the level of expression of the sample and plotted the risk score distribution for the sample. From the figure, it is evident that the mortality rate for samples having an elevated risk score was substantially greater in contrast with the death rate for samples having a reduced risk score, indicating that high-risk score samples have a poorer prognosis. The expression of eight separate signature genes differs in response to an increase in risk value. The elevated expression of SPARCL1, HAND1, CRIP2 and ZNF385A was identified as risk variables. The high expression of CXCL1,

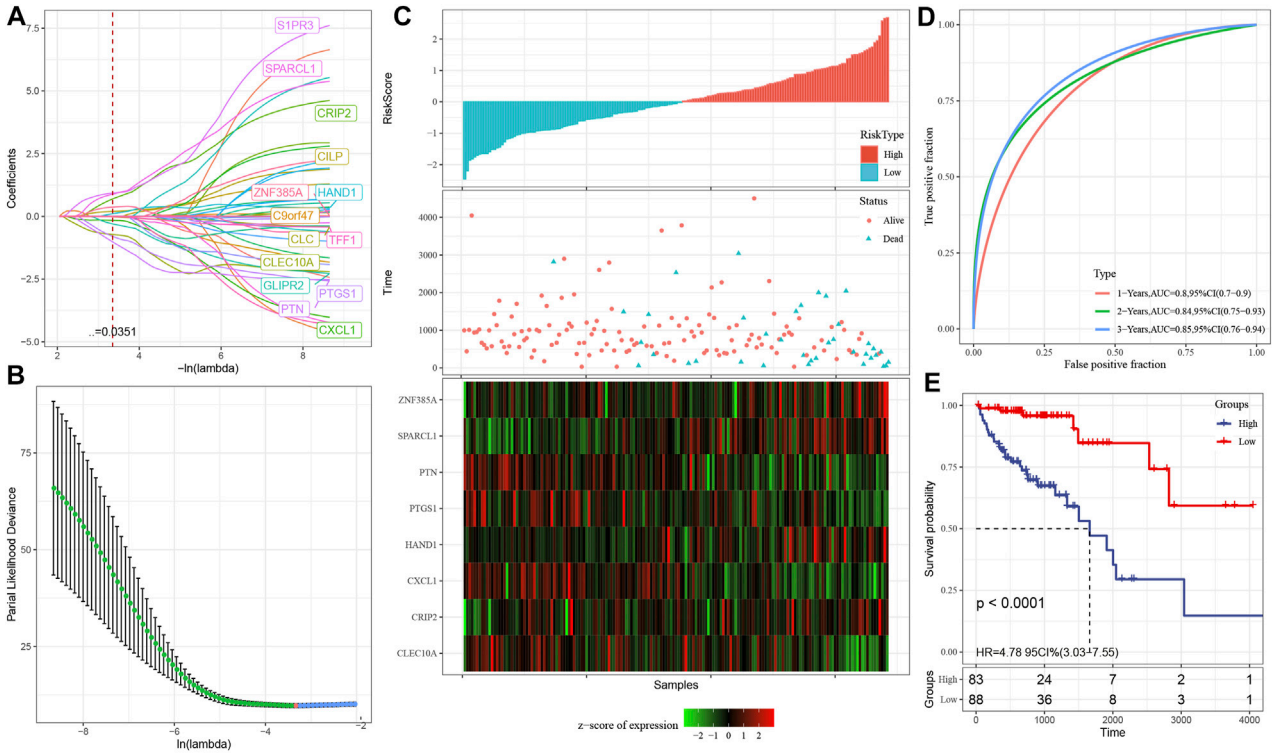


FIGURE 7 | (A): The change trajectory of each independent variable, where the independent variable lambda log value is represented by the horizontal axis, while the independent variable coefficient is represented by the vertical axis; **(B)**: In each lambda, the confidence interval is calculated; **(C)**: The survival status, survival duration, risk score, and eight-gene expression in the TCGA training set; **(D)**: The classification of the ROC curve and AUC is performed based on an 8-gene signature; **(E)**: In the training set, the survival curve distribution of an 8-gene signature was calculated using the KM algorithm.

CLEC10A, PTGS1 and PTN were correlated with low risk, indicating that these genes have a protective function. In addition, we utilized the timeROC function of the R software to examine the ROC of the prognostic categorization of risk score. We examined the categorization effectiveness of prognostic predictions in three different time periods: one-, two-, and three-year (**Figure 7D**). It is evident that the model has a large area underneath the AUC line; Eventually, we computed the zscore on the risk scores and classified the samples with risk scores larger than zero into two groups: high- and low-risk groups. After plotting the KM curve as illustrated in **Figure 7E**, we identified a substantial difference between the two groups ($p < 0.0001$), of which 88 samples were categorized into the low-risk group while 83 samples were categorized into the high-risk group.

Multiple Data Sets Confirm the Stability of the Eight-Gene Signature

The validation set from TCGA as well as all other data sets were utilized to assess the robustness of the model, with a similar model and equivalent coefficients as the training set being used to make the determination. After determining the levels of expression of the risk score of each sample, the ROC analysis of the risk score of each sample was performed utilizing the timeROC function in the R software. We examined the categorization effectiveness of prognostic predictions in three different time periods: one-, two-, and three-year. We subsequently computed the zscore and the risk scores, classified the samples with risk scores larger than zero into two groups: high-and low-risk. Then, we created the KM curve. In the TCGA internal validation data, as depicted in **Figures 8A–B**, we found that the model exhibits an elevated AUC in the independent internal validation data set, and a substantial difference was observed between the low- and high-risk groups according to the risk value ($p = 0.016$). Specifically, 94 samples were categorized into the low-risk group while 78 samples were categorized into the high-risk group. In the whole TCGA data set, as depicted in **Figures 8C,D**, it was discovered that the AUC of the model in 3 years reaches 0.76, and there is also a considerable difference between the low- and high-risk groups according to the risk value of the samples ($p < 0.0001$), of which 162 samples are categorized into the high-risk groups while 181 samples fall under the low-risk group.

In the external verification set GSE17538, we employed a similar model and equivalent coefficients as those utilized in the training set. In a similar manner, the risk score of each sample was determined in accordance with the degree of expression of the sample. In addition, we utilized the timeROC function in the R software to investigate the ROC of the prognostic categorization of the risk score. We examined the categorization effectiveness of prognostic predictions in three different time periods: one-, two-, and three-year (**Figure 8E**); In the end, we calculated the risk score,

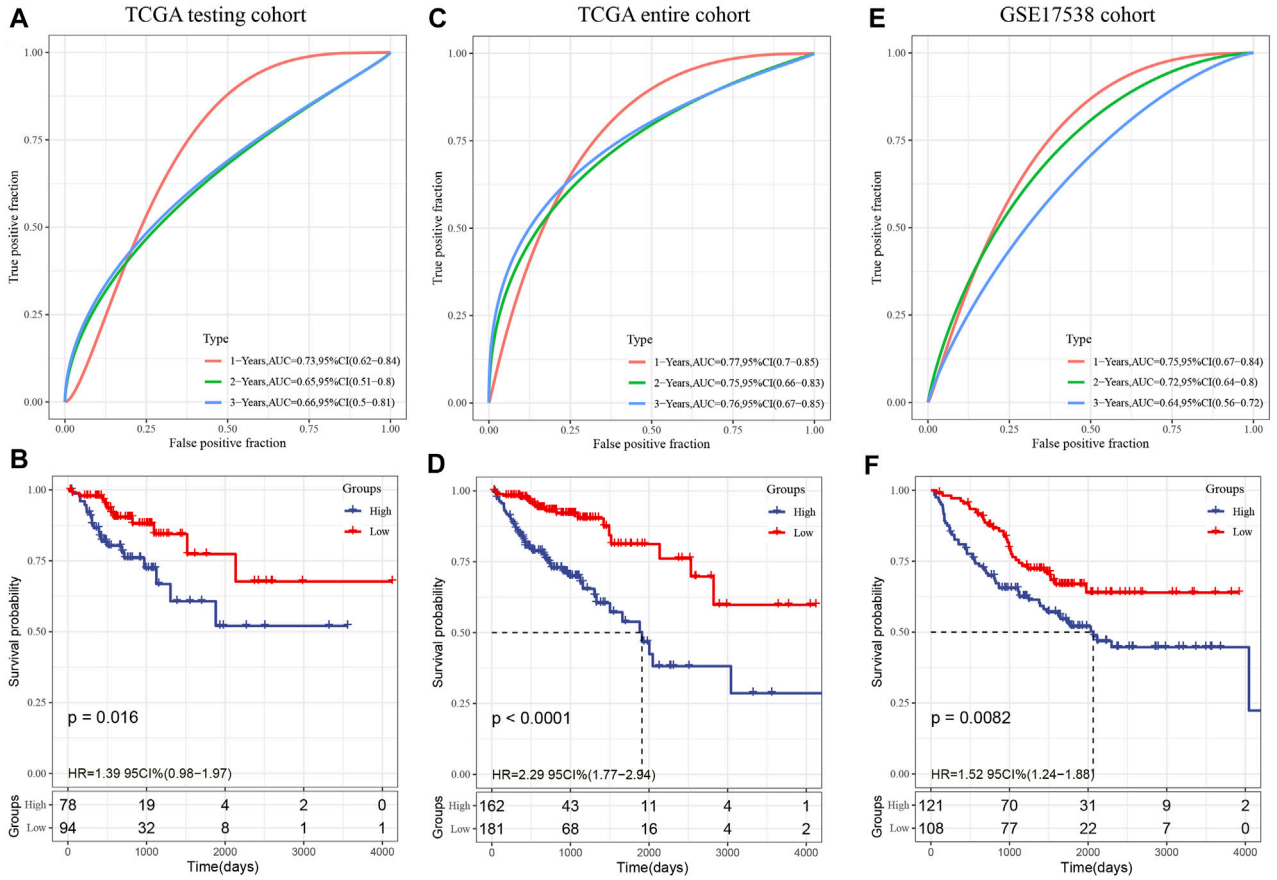


FIGURE 8 | (A): In the TCGA test set, the ROC curve and AUC were categorized utilizing an 8-gene signature; **(B)**: In the TCGA test set, the KM survival curve distribution of an eight-gene signature was examined; **(C)**: In the whole TCGA data set, the ROC curve and AUC were categorized utilizing an 8-gene signature; **(D)**: The distribution of the KM survival curve for an eight-gene signature throughout the whole TCGA data set. **(E)**: The classification of the ROC curve and AUC is based on an eight-gene signature; **(F)**: The KM survival curve illustrating the distribution of the eight-gene signature within the independent validation data set GSE17538.

classified the samples into low- and high-risk groups, and plotted the KM curve, as demonstrated in **Figure 8F**. The findings illustrated a statistically substantial difference between the two groups ($p = 0.0082$), of which 121 samples are categorized into the high-risk group while 208 samples fall under the low-risk group.

Cox Analysis and Nomogram Construction of 8-Gene Signature

The risk score was shown to be strongly linked to survival in the TCGA data set, according to univariable Cox regression analysis (**Figure 9A**). Corresponding multivariable Cox regression analysis found that risk score was still substantially linked to survival (**Figure 9B** HR = 1.86, 95% CI = 1.42—2.45, P < 1 e-5). The findings described above demonstrate that our eight-gene signature model exhibits strong predictive performance in terms of clinical applicability. Specifically, to examine the correlation between biological activities and risk scores among various samples, we utilized the gene expression profiles related to these samples, performed single sample GSEA analysis utilizing the GSVA package of the R software, measured the scores of each sample on various functions, and acquired the ssGSEA scores of each sample related to each function, and thereafter obtained the connection between any of these functions and risk scores. **Figure 9C** illustrates our selection of functions with correlations larger than 0.3. It could be observed that eight pathways were positively linked to the sample risk score, while three pathways were inversely linked to the sample risk score. As demonstrated in **Figure 9D**, we chose the 11 topmost correlated KEGG pathways and performed cluster analysis on them based on their enrichment ratings. We found that the tumor-related pathways such as KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION, KEGG_FOCAL_ADHESION, and KEGG_ECM_RECEPTOR_INTERACTION increased as the risk score continued to increase.

The univariate and multivariate analyses illustrated that in addition to the risk score, the clinical feature M stage was found to independently function as a prognostic marker, implying that they had complementary value. To additionally enhance the prediction performance, we integrated the M stage and risk score, established a new nomogram using the Cox model, and combined the two independent prognostic predictors (**Figure 9E**), according to this model, we found that risk score has the most contribution to OS, with the M stage coming second. In addition, we utilized the calibration

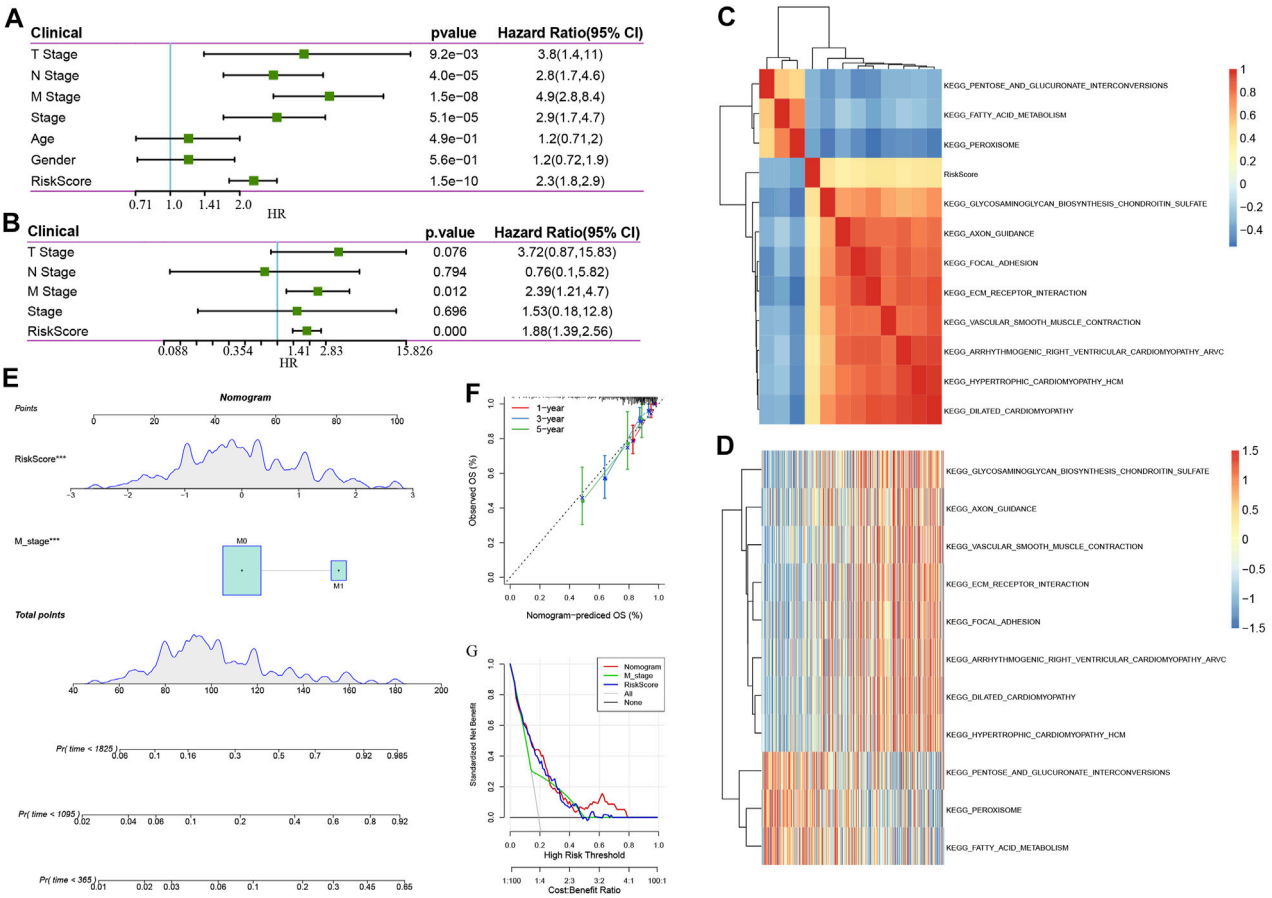


FIGURE 9 | (A): Findings of the univariate analysis and clinical characteristics of the risk score; **(B)**: Findings of the multivariate analysis and clinical characteristics of the risk score. **(C)**. KEGG pathways that are substantially linked to risk score are grouped together in a cluster; **(D)**: Changes in the association between KEGG pathways that are strongly associated with the risk score and ssGSEA score (the sample is represented by the abscissa, and the risk scores are elevated from the left side to the right side). **(E)**: Nomogram model of M stage and risk score combination; **(F)**: One-, three-, and five-year Calibration curve of the nomogram; **(G)**: Decision curve of M stage, risk score, and nomogram.

curve to determine the predictive accuracy regarding the model (**Figure 9F**). The findings illustrated that the predictive calibration curve of three estimation points in one-, three-, and five-year is proximate to the standard curve, indicating that the model possesses strong predictive accuracy. Furthermore, we utilized DCA (decision curve) to assess the reliability (**Figure 9G**) of the model. It was discovered that at a high-risk threshold of 0.1–0.3, the advantages of nomogram and risk score substantially outweigh those of M stage and extreme curve by a wide margin. When the high-risk threshold is 0.3–0.5, the benefits of risk score, nomogram, and M stage are similar. When the high-risk threshold is greater than 0.5, the net benefits of the nomogram are greater than 0, which has clinical significance. On the whole, risk score and nomogram have good reliability when the high-risk threshold is 0.1–0.5. When the high-risk threshold is greater than 0.5, the nomogram has a high benefit rate.

Comparison Between Risk Model and Other Models

Following a review of the literature, we eventually obtained five prognosis-related risk models to be used for comparison with our eight-gene model including 15-gene signature (Dai) (Dai et al., 2018), 9-gene signature(Mo) (Mo et al., 2019), 12-gene signature(Sun) (Sun et al., 2018), 13-gene signature(Tian) (Tian et al., 2017),15-gene signature(Xu) (Xu et al., 2017). In the attempt of making the model more similar to some degree, the risk score of each COAD sample from TCGA was computed as per the matching genes in the five models, utilizing the same approach, and the risk score was zscore. After the zscore, the samples whose risk score was found to exceed zero were categorized into low- and high-risk groups. The difference in COAD prognosis between the two groups was computed and compared. **Figures 10A–J** depicts the ROC and COAD-KM curves of the five models. It can be seen that the AUC values of the five models in 1, 2, and 3 years are lower than our model. The COAD prognosis of the low and high grouped samples of the five models except for the 9-gene signature (Mo) model was different (log-rank $p < 0.05$). Our model was discovered to be more rational and accurate even when fewer genes are used. Moreover, we also utilized the "rms" package in R to determine the concordance index (C-index) of various models in order to compare the predictive accuracy among models. According to **Figure 10K**, the C-index of the eight-genes model is the greatest, showing that the

overall performance of our model is better than the other five models.

Verification of 8 Gene Expression

In order to study the expression patterns of genes in tumors and normal cell lines, PCR was used to detect the mRNA expression of genes, and the results showed that the expression of SPARCL1, HAND1, CLEC10A, PTGS1 genes in tumor group was significantly lower than that in the normal group. However, the expression of CXCL1 in tumor group was significantly higher than that in the control group. There is no difference in the expression of ZNF385A and CRIP2 between the tumor and the control group (**Figures 11A–H**).

DISCUSSION

The tumor microenvironment performs an integral function in the progression of tumors and it is not only a barrier for tumor cell metastasis (Wang and Hielscher, 2017; Lin et al., 2019), but also a favorable "soil" for tumor cell reproduction, and endows tumor cells with the ability to metastasize to a distance (Mao and Schwarzbauer, 2005). In addition to tumor cells, the tumor microenvironment is composed of stromal cells, inflammatory cells, vascular system, and extracellular matrix (Frankel et al., 2017), in which tumor-related immune cells are an important part. Previous studies have shown that stromal-rich tumors have been proved to be unfavorable for prognosis, and the same conclusion has been reached in colorectal cancer (West et al., 2010; van Pelt et al., 2018; Geessink et al., 2019), which is consistent with our results in this study. The EMT and cryptlike subtypes with the worst prognosis have high immune microenvironment infiltration. Of course, our results also provide another possible reason. In these two subtypes, the T cell rejection score and T cell dysfunction score are high. It is speculated that the genetic abnormalities of tumor cells will reprogram the surrounding infiltrating cells, resulting in the tumor microenvironment conducive to tumor development. This gene abnormality is usually manifested in the abnormal amplification of some specific genes. These results show that COAD leads to different prognostic outcomes due to its high heterogeneity. At present, the clinical TNM stage may not fully reflect its prognosis and the choice of treatment methods. The 2020 ESMO colon cancer clinical practice guide recommends the application of immune scores to predict the prognosis of patients with colon cancer and guide clinical medication (Argilés et al., 2020). On the other hand, our results show that the proportion of response to immunotherapy in Collective and Proliferative subtypes is substantially elevated as opposed to that in CryptLike and EMT subtypes, which cannot be perfectly explained by the results of this study. However, we speculate that this may be related to the imbalance of intestinal-specific microbiota. Studies have shown that in crypt-like infiltration, mucosal microbiota still exists even if the tumor has distant metastasis. While in EMT invasion, the tumor seems to lose the mucosal microbiota (Bullman et al., 2017). The state of intestinal flora also performs an integral function in the efficacy of immune checkpoint inhibitors (Westdorp et al., 2021). Unfortunately, we lack the analysis of intestinal flora in this study. We will improve this part of data in further research to provide a basis for further guiding clinical medication.

Proliferation and metastasis are the two characteristics of tumors. There is a deviation in the gene expression of these two characteristics of tumors. We extracted the DEGs of these two groups of COAD to screen the genes that perform an instrumental function in tumors. The results of pathway enrichment suggest that the pathways involved by these genes are concentrated in Cell adhesion molecules (CAMs) (Bronikowska et al., 2021), cytokine receptor interaction (Wen et al., 2021), focal adhesion (Lu et al., 2021), NF kappa B signaling pathway (Takakura et al., 2021), proteoglycans in cancer (Jinmin Sun et al., 2021), PI3K Akt signaling pathway (Xue et al., 2021), etc, which are strongly associated with to the onset and progression of tumors. Further cluster analysis showed that CryptLike, EMT, Collective and Proliferative were distributed in different modules respectively. These results showed that different clinical subtypes of colorectal cancer not only showed differences in tumor behavior, but also the fundamental reason was the abnormal activation of different genes and pathways, which was not only an important reason for tumor progression but also a key clinical target.

In order to further enhance clinical applicability, we established prognostic gene signatures related to differential genes. After lasso regression analysis, the combination with the maximum occurrence frequency includes 8 genes: SPARCL1 (SPARC-like protein 1), CXCL1 (Growth-regulated alpha protein), HAND1 (Heart- and neural crest derivatives-expressed protein 1), CRIP2 (Cysteine-rich protein 2), CLEC10A(C-type lectin domain family 10, member A), PTGS1 (Prostaglandin G/H synthase 1), PTN (Pleiotrophin), ZNF385A (Zinc finger protein 385A), The role of these genes in tumors has been preliminarily studied. SPARCL1 is considered a tumor suppressor, which inhibits tumor progression in a variety of tumors. SPARCL1 may prevent the activity of colorectal cancer through its DNA methylation (Hu et al., 2021). The expression of SPARCL1 in HeLa cells is low at the protein level and transcription level. When SPARCL1 is overexpressed in HeLa cells the proliferation, migration, and invasion of cells are strongly inhibited (Zhang et al., 2021). CXCL1, as a chemokine, has been proved to promote tumor progression in a variety of tumors. The CXCL1 overexpression has a positive relationship with the migration and invasive activity of osteosarcoma cell lines (Lee et al., 2021). In acute leukemia, bladder cancer, and other tumors, its expression is increased, and its prognosis is poor, (Xiaoqi Sun et al., 2021; Yazdani et al., 2021). HAND1 is a mesodermal marker (Yi et al., 2020). At present, there are many studies on its role in EMT, but there are contradictions about its role in tumors. HAND1 is expressed in invasive gastrointestinal stromal tumors (GIST) (Hemming et al., 2021). However, in medulloblastoma, HAND1 expression may be the key to weaken EMT, which may be different from previous research conclusions (Asuthkar et al., 2016), but it may be related to different apparent modifications of HAND1 in tumors (Tan et al., 2014). CRIP2 is a transcription factor, which is an unfavorable prognostic factor for breast

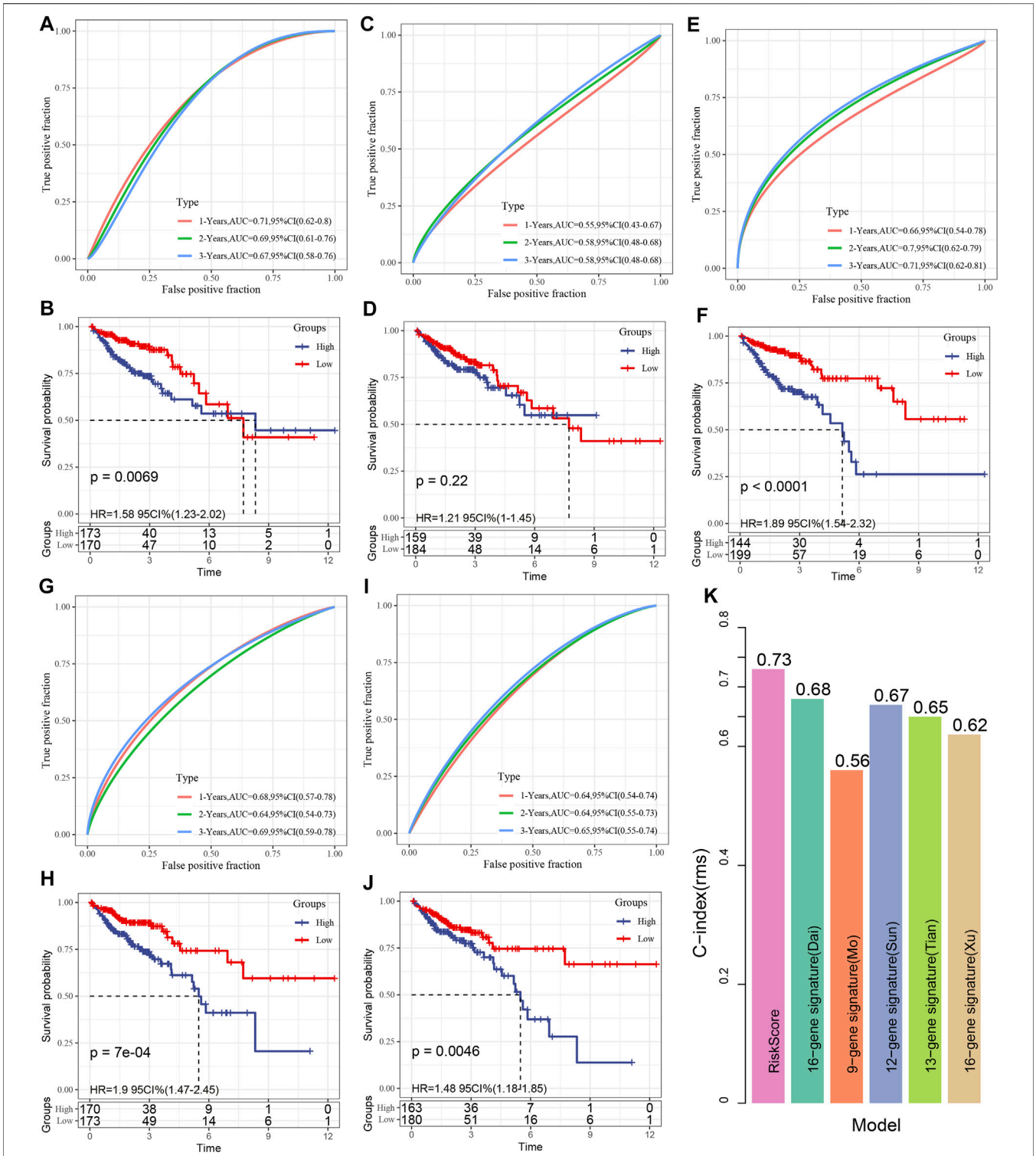


FIGURE 10 | (A, B): The receiver operating characteristic (ROC) of the 15 gene signature (Dai) risk model, as well as the COAD KM curve of low- and high-risk clustered samples; **(C, D)**: The ROC curve of the 9-gene signature (Mo) and the COAD KM curve of the low/high; **(E, F)**: The ROC of 12-gene signature(Sun) as well as COAD KM curve of Low/High; **(G, H)**: ROC of 13-gene signature(Tian) and COAD KM curve of Low/High; **(I, J)**: ROC of 15-gene signature(Xu) and COAD KM curve of Low/High/; **(K)**. C-indexes of six prognostic risk models.

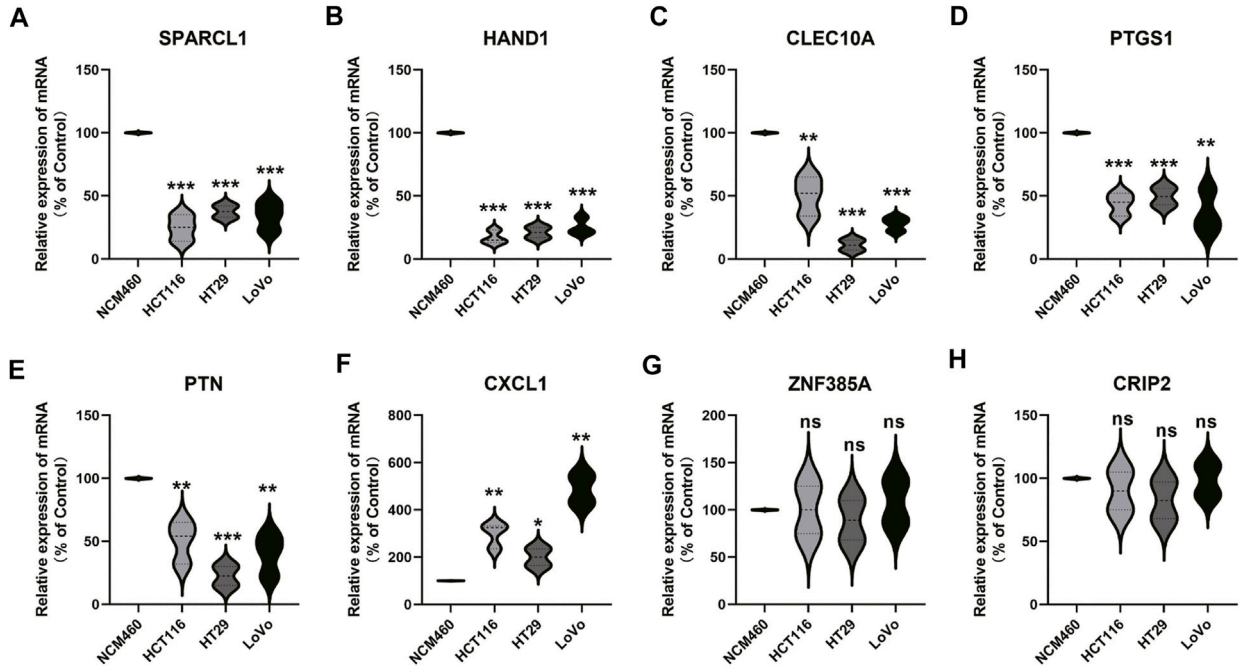


FIGURE 11 | Gene expression verification in normal samples and tumor samples. From **(A to H)** represent SPARCL1, HAND1, CLEC10A, PTGS1, PTN, CXCL1, ZNF385A, and CRIP2, respectively.

cancer as well as colorectal cancer (Shi et al., 2016; Zhang et al., 2020). CRIP2 is an autophagy inhibitory protein. CRIP2 mediated copper metabolism activates autophagy in cancer cells (Chen et al., 2021). Moreover, CRIP2 was confirmed to be up-modulated at the mRNA and protein levels of anti-radiation cells, which is a potential diagnostic biomarker and a key biomarker for predicting prognosis (Li et al., 2021). CLEC10A, also known as MGL (macrophage galactose type C lectin), has recently been reported to perform a crucial part in enhancing immune cell activities. CLEC10A can recognize tumor-associated antigens and pass them to CD4 T cells (van Vliet et al., 2007). Moreover, CLEC10A has been shown to greatly enhance the stimulation of antigen-specific CD8 T cells (Napoletano et al., 2012). The role of CLEC10A in enhancing the anticancer effects of immune cells has undoubtedly drawn the interest of researchers, and it has been identified as a potential target for cancer immunotherapy treatment (Eggink et al., 2018). Constitutive cyclooxygenase (COX) - 1 (gene PTGS1) is significantly related to the concentration of PGE2 in the colon and is highly expressed in colon cancer. Therefore, PTGS1 may be regarded as another potential target for colon cancer prevention in high-risk groups (Sidahmed et al., 2016; Ayiomamitis et al., 2019). In colorectal cancer, PTN may play a role as the downstream of PRPH and promote tumor progression (Huang et al., 2021). Tumor-related macrophages increase the proportion of tumor stem cells in lymphoma via the secretion of PTN (Wei et al., 2019). Upregulation of PTN by activating the NF- κ B pathway promotes tumor cell proliferation, inhibits apoptosis and chemosensitivity (Huang et al., 2018). According to the results of a meta-analysis, increased expression of PTN was substantially associated with advanced TNM stage and dismal OS in cancer patients (Zhou et al., 2018). ZNF385A, as RNA binding proteins (RBPs), is a constituent factor of the COAD prognosis model, but its specific mechanism is not clear (Chang et al., 2021). These results show that although these genes play different roles in tumors, according to the current research results, more genes may participate in apparent modification, which may also be why there exist obvious differences in the expression of these genes in distinct phenotypes of COAD, which performs an instrumental function in the prognosis of COAD, but this needs our further verification.

In conclusion, in this study, we further analyzed the differences between molecular subtypes of prognosis of patients with COAD, TMB and common gene variants, immune score, and efficacy of immunotherapy/chemotherapy based on the research of Shayingzhao et al. Moreover, the risk score prognostic model of 8 genes is constructed according to the differential genes. The risk score constructed according to these 8 genes can divide the patients into low- and high-risk groups. The death proportion of the samples with a high-risk score is substantially elevated as opposed to that with the low-risk score, implying that the high risk score samples exhibit unfavorable prognoses. Further application of the training set as well as verification set have higher AUC shows that risk score constructed by our 8 genes has a stable role in predicting prognosis and provides a basis for clinical precision treatment.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. LG: Conceptualization, Methodology, Software, Writing-Reviewing and Editing. CJ: Conceptualization, Methodology, Software, Writing-Reviewing and Editing. CX: Methodology, Software. Dongyang Wang: Methodology, Software. Yuguang Shen: Writing-Reviewing and Editing. YL: Data curation, Writing-Original draft preparation. HZ: Supervision.

ACKNOWLEDGMENTS

The authors thank all contributors of the present study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.828415/full#supplementary-material

Supplementary Table S1 | A total of 366 samples were divided into four subtypes based on TCGA-COAD.

REFERENCES

Argilés, G., Tabernero, J., Labianca, R., Hochhauser, D., Salazar, R., Iveson, T., et al. (2020). Localised colon Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* 31 (10), 1291–1305. doi:10.1016/j.annonc.2020.06.022

Asaoka, Y., Ijichi, H., and Koike, K. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* 373 (20), 1979. doi:10.1056/NEJMc1510353

Asuthkar, S., Guda, M. R., Martin, S. E., Antony, R., Fernandez, K., Lin, J., et al. (2016). Hand1 Overexpression Inhibits Medulloblastoma Metastasis. *Biochem. biophysical Res. Commun.* 477 (2), 215–221. doi:10.1016/j.bbrc.2016.06.045

Atri, C., Guerfali, F., and Laouini, D. (2018). Role of Human Macrophage Polarization in Inflammation during Infectious Diseases. *Int. J. Mol. Sci.* 19 (6), 1801. doi:10.3390/ijms19061801

Ayiomamitis, G. D., Notas, G., Vasilakaki, T., Tsavari, A., Vederaki, S., Theodosopoulos, T., et al. (2019). Understanding the Interplay between COX-2 and hTERT in Colorectal Cancer Using a Multi-Omics Analysis. *Cancers (Basel)* 11 (10), 1536. doi:10.3390/cancers11101536

Bronikowska, J., Kłósek, M., Janeczko, T., Kostrzewa-Susłow, E., and Czuba, Z. P. (2021). The Modulating Effect of Methoxy-Derivatives of 2'-hydroxychalcones on the Release of IL-8, MIF, VCAM-1 and ICAM-1 by colon Cancer Cells. *Biomed. Pharmacother. = Biomedecine pharmacotherapie* 145, 112428. doi:10.1016/j.biopha.2021.112428

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., et al. (2017). Analysis of Fusobacterium Persistence and Antibiotic Response in Colorectal Cancer. *Science* 358 (6369), 1443–1448. doi:10.1126/science.aal5240

Chang, K., Yuan, C., and Liu, X. (2021). A New RBPs-Related Signature Predicts the Prognosis of Colon Adenocarcinoma Patients. *Front. Oncol.* 11, 627504. doi:10.3389/fonc.2021.627504

Chen, L., Li, N., Zhang, M., Sun, M., Bian, J., Yang, B., et al. (2021). APEX2-based Proximity Labeling of Atox1 Identifies CRIP2 as a Nuclear Copper-binding Protein that Regulates Autophagy Activation. *Angew. Chem. Int. Ed.* 60 (48), 25346–25355. doi:10.1002/anie.202108961

Chu, Q. D., Zhou, M., Medeiros, K., and Peddi, P. (2016). Positive Surgical Margins Contribute to the Survival Paradox between Patients with Stage IIB/C (T4N0) and Stage IIIA (T1-2N1, T1N2a) colon Cancer. *Surgery* 160 (5), 1333–1343. doi:10.1016/j.surg.2016.05.028

Comprehensive Molecular Characterization of Human colon and Rectal Cancer. *Nature* 2012, 487(7407):330–337.doi:10.1038/nature11252

Dai, W., Li, Y., Mo, S., Feng, Y., Zhang, L., Xu, Y., et al. (2018). A Robust Gene Signature for the Prediction of Early Relapse in Stage I-III colon Cancer. *Mol. Oncol.* 12 (4), 463–475. doi:10.1002/1878-0261.12175

Eggink, L. L., Roby, K. F., Cote, R., and Kenneth Hoober, J. (2018). An Innovative Immunotherapeutic Strategy for Ovarian Cancer: CLEC10A and Glycomimetic Peptides. *J. Immunotherapy Cancer* 6 (1), 28. doi:10.1186/s40425-018-0339-5

Frankel, T., Lanfranca, M. P., and Zou, W. (2017). The Role of Tumor Microenvironment in Cancer Immunotherapy. *Adv. Exp. Med. Biol.* 1036, 51–64. doi:10.1007/978-3-319-67577-0_4

Friedl, P., and Gilmour, D. (2009). Collective Cell Migration in Morphogenesis, Regeneration and Cancer. *Nat. Rev. Mol. Cel Biol* 10 (7), 445–457. doi:10.1038/nrm2720

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33 (1), 1–22. doi:10.18637/jss.v033.i01

Geessink, O. G. F., Baidoshvili, A., Klaase, J. M., Ehteshami Bejnordi, B., Litjens, G. J. S., van Pelt, G. W., et al. (2019). Computer Aided Quantification of Intratumoral Stroma Yields an Independent Prognosticator in Rectal Cancer. *Cell Oncol.* 42 (3), 331–341. doi:10.1007/s13402-019-00429-z

Hemming, M. L., Coy, S., Lin, J.-R., Andersen, J. L., Przybyl, J., Mazzola, E., et al. (2021). HAND1 and BARX1 Act as Transcriptional and Anatomic Determinants of Malignancy in Gastrointestinal Stromal Tumor. *Clin. Cancer Res.* 27 (6), 1706–1719. doi:10.1158/1078-0432.ccr-20-3538

Hu, H., Wu, D., Liu, X., Yu, H., Xu, J., Cai, W., et al. (2021). SPARCL1 Exhibits Different Expressions in Left- and Right-Sided colon Cancer and Is Downregulated via DNA Methylation. *Epigenomics* 13 (16), 1269–1282. doi:10.2217/epi-2021-0231

Huang, H., Xu, S., Chen, A., Li, F., Wu, J., Tu, X., et al. (2021). Identification of a 5-Gene-Based Scoring System by WGCNA and LASSO to Predict Prognosis for Rectal Cancer Patients. *Anal. Cell. Pathol. (Amsterdam)* 2021, 6697407. doi:10.1155/2021/6697407

Huang, P., Ouyang, D.-j., Chang, S., Li, M.-y., Li, L., Li, Q.-y., et al. (2018). Chemotherapy-driven Increases in the CDKN1A/PTN/PTPRZ1 axis Promote Chemoresistance by Activating the NF-Kb Pathway in Breast Cancer Cells. *Cell Commun Signal* 16 (1), 92. doi:10.1186/s12964-018-0304-4

Jung, G., Benítez-Ribas, D., Sánchez, A., and Balaguer, F. (2020). Current Treatments of Metastatic Colorectal Cancer with Immune Checkpoint Inhibitors-2020 Update. *J. Clin. Med.* 9 (11), 3520. doi:10.3390/jcm9113520

Labianca, R., Beretta, G. D., Kildani, B., Milesi, L., Merlin, F., Mosconi, S., et al. (2010). Colon Cancer. *Crit. Rev. oncology/hematology* 74 (2), 106–133. doi:10.1016/j.critrevonc.2010.01.010

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Le, D. T., Durham, J. N., Smith, K. N., Wang, H., Bartlett, B. R., Aulakh, L. K., et al. (2017). Mismatch Repair Deficiency Predicts Response of Solid Tumors to PD-1 Blockade. *Science* 357 (6349), 409–413. doi:10.1126/science.aan6733

Lee, C.-W., Chiang, Y.-C., Yu, P.-A., Peng, K.-T., Chi, M.-C., Lee, M.-H., et al. (2021). A Role of CXCL1 Drives Osteosarcoma Lung Metastasis via VCAM-1 Production. *Front. Oncol.* 11, 735277. doi:10.3389/fonc.2021.735277

Li, F., Bing, Z., Chen, W., Ye, F., Liu, Y., Ding, L., et al. (2021). Prognosis Biomarker and Potential Therapeutic Target CRIP2 Associated with Radiosensitivity in

NSCLC Cells. *Biochem. biophysical Res. Commun.* 584, 73–79. doi:10.1016/j.bbrc.2021.11.002

Lin, T. C., Yang, C. H., Cheng, L. H., Chang, W. T., Lin, Y. R., and Cheng, H. C. (2019). Fibronectin in Cancer: Friend or Foe. *Cells* 9 (1), 27. doi:10.3390/cells9010027

Lu, X., Zhou, B., Cao, M., Shao, Q., Pan, Y., and Zhao, T. (2021). CTEN Inhibits Tumor Angiogenesis and Growth by Targeting VEGFA through Down-Regulation of β-Catenin in Breast Cancer. *Techn. Cancer Res. Treat.* 20, 15330338211045506. doi:10.1177/15330338211045506

Mao, Y., and Schwarzbauer, J. E. (2005). Fibronectin Fibrillogenesis, a Cell-Mediated Matrix Assembly Process. *Matrix Biol.* 24 (6), 389–399. doi:10.1016/j.matbio.2005.06.008

Mo, S., Dai, W., Xiang, W., Li, Y., Feng, Y., Zhang, L., et al. (2019). Prognostic and Predictive Value of an Autophagy-Related Signature for Early Relapse in Stages I-III colon Cancer. *Carcinogenesis* 40 (7), 861–870. doi:10.1093/carcin/bgz031

Napoletano, C., Zizzari, I. G., Rughetti, A., Rahimi, H., Irimura, T., Clausen, H., et al. (2012). Targeting of Macrophage Galactose-type C-type Lectin (MGL) Induces DC Signaling and Activation. *Eur. J. Immunol.* 42 (4), 936–945. doi:10.1002/eji.201142086

Neri, E., Faggioni, L., Cini, L., and Bartolozzi, C. (2010). Colonic Polyps: Inheritance, Susceptibility, Risk Evaluation, and Diagnostic Management. *Cancer Manag. Res.* 3, 17–24. doi:10.2147/cmar.s15705

Overman, M. J., McDermott, R., Leach, J. L., Lonardi, S., Lenz, H.-J., Morse, M. A., et al. (2017). Nivolumab in Patients with Metastatic DNA Mismatch Repair-Deficient or Microsatellite Instability-High Colorectal Cancer (CheckMate 142): an Open-Label, Multicentre, Phase 2 Study. *Lancet Oncol.* 18 (9), 1182–1191. doi:10.1016/s1470-2045(17)30422-9

Roncucci, L., and Mariani, F. (2015). Prevention of Colorectal Cancer: How many Tools Do We Have in Our Basket. *Eur. J. Intern. Med.* 26 (10), 752–756. doi:10.1016/j.ejim.2015.08.019

Shi, W., Bruce, J., Lee, M., Yue, S., Rowe, M., Pintilie, M., et al. (2016). MiR-449a Promotes Breast Cancer Progression by Targeting CRIP2. *Oncotarget* 7 (14), 18906–18918. doi:10.18632/oncotarget.7753

Sidahmed, E., Sen, A., Ren, J., Patel, A., Turgeon, D. K., Ruffin, M. T., et al. (2016). Colonic Saturated Fatty Acid Concentrations and Expression of COX-1, but Not Diet, Predict Prostaglandin E2in Normal Human Colon Tissue. *Nutr. Cancer* 68 (7), 1192–1201. doi:10.1080/01635581.2016.1213866

Sun, D., Chen, J., Liu, L., Zhao, G., Dong, P., Wu, B., et al. (2018). Establishment of a 12-gene Expression Signature to Predict colon Cancer Prognosis. *PeerJ* 6, e4942. doi:10.7717/peerj.4942

Sun, J., Xu, Y., Liu, J., Cui, H., Cao, H., and Ren, J. (2021). PDRG1 Promotes the Proliferation and Migration of GBM Cells by MEK/ERK/CD44 Pathway. *Cancer Sci.* 113 (2), 500–516. doi:10.1111/cas.15214

Sun, X., Chen, Q., Zhang, L., Chen, J., and Zhang, X. (2021). Exploration of Prognostic Biomarkers and Therapeutic Targets in the Microenvironment of Bladder Cancer Based on CXC Chemokines. *Math. Biosci. Eng.* 18 (5), 6262–6287. doi:10.3934/mbe.2021313

Takakura, Y., Hori, N., Terada, N., Machida, M., Yamaguchi, N., Takano, H., et al. (2021). VGLL3 Activates Inflammatory Responses by Inducing Interleukin-1α Secretion. *FASEB J.* 35 (11), e21996. doi:10.1096/fj.202100679rr

Tan, J., Yang, X., Jiang, X., Zhou, J., Li, Z., Lee, P. L., et al. (2014). Integrative Epigenome Analysis Identifies a Polycomb-Targeted Differentiation Program as a Tumor-Suppressor Event Epigenetically Inactivated in Colorectal Cancer. *Cell Death Dis* 5 (7), e1324. doi:10.1038/cddis.2014.283

Thiery, J. P. (2002). Epithelial-mesenchymal Transitions in Tumour Progression. *Nat. Rev. Cancer* 2 (6), 442–454. doi:10.1038/nrc822

Tian, X., Zhu, X., Yan, T., Yu, C., Shen, C., Hu, Y., et al. (2017). Recurrence-associated Gene Signature Optimizes Recurrence-free Survival Prediction of Colorectal Cancer. *Mol. Oncol.* 11 (11), 1544–1560. doi:10.1002/1878-0261.12117

van Pelt, G. W., Kjær-Frifeldt, S., van Krieken, J. H. J. M., Al Dieri, R., Morreau, H., Tollenaar, R. A. E. M., et al. (2018). Scoring the Tumor-Stroma Ratio in colon Cancer: Procedure and Recommendations. *Virchows Arch.* 473 (4), 405–412. doi:10.1007/s00428-018-2408-z

van Vliet, S. J., Aarnoudse, C. A., Broks-van den Berg, V. C. M., Boks, M., Geijtenbeek, T. B. H., and van Kooyk, Y. (2007). MGL-mediated Internalization and Antigen Presentation by Dendritic Cells: a Role for Tyrosine-5. *Eur. J. Immunol.* 37 (8), 2075–2081. doi:10.1002/eji.200636838

Wang, J., Wang, T., Sun, Y., Feng, Y., Kisseberth, W. C., Henry, C. J., et al. (2018). Proliferative and Invasive Colorectal Tumors in Pet Dogs Provide Unique Insights into Human Colorectal Cancer. *Cancers (Basel)* 10 (9), 330. doi:10.3390/cancers10090330

Wang, J. P., and Hielscher, A. (2017). Fibronectin: How its Aberrant Expression in Tumors May Improve Therapeutic Targeting. *J. Cancer* 8 (4), 674–682. doi:10.7150/jca.16901

Wei, X., Yang, S., Pu, X., He, S., Yang, Z., Sheng, X., et al. (2019). Tumor-associated Macrophages Increase the Proportion of Cancer Stem Cells in Lymphoma by Secreting Pleiotrophin. *Am. J. Transl Res.* 11 (10), 6393–6402.

Wen, J., Mao, X., Cheng, Q., Liu, Z., and Liu, F. (2021). A Pan-Cancer Analysis Revealing the Role of TIGIT in Tumor Microenvironment. *Sci. Rep.* 11 (1), 22502. doi:10.1038/s41598-021-01933-9

West, N. P., Dattani, M., McShane, P., Hutchins, G., Grabsch, J., Mueller, W., et al. (2010). The Proportion of Tumour Cells Is an Independent Predictor for Survival in Colorectal Cancer Patients. *Br. J. Cancer* 102 (10), 1519–1523. doi:10.1038/sj.bjc.6605674

Westdorp, H., Sweep, M. W. D., Gorris, M. A. J., Hoentjen, F., Boers-Sonderen, M. J., Post, R. S. v. d., et al. (2021). Mechanisms of Immune Checkpoint Inhibitor-Mediated Colitis. *Front. Immunol.* 12, 768957. doi:10.3389/fimmu.2021.768957

Xu, G., Zhang, M., Zhu, H., and Xu, J. (2017). A 15-gene Signature for Prediction of colon Cancer Recurrence and Prognosis Based on SVM. *Gene* 604, 33–40. doi:10.1016/j.gene.2016.12.016

Xue, C., Li, G., Lu, J., and Li, L. (2021). Crosstalk between circRNAs and the PI3K/AKT Signaling Pathway in Cancer Progression. *Sig Transduct Target. Ther.* 6 (1), 400. doi:10.1038/s41392-021-00788-w

Yazdani, B., Hassanshahi, G., Mousavi, Z., Ahmadi, Z., Khorramdelazad, H., Moradabadi, A., et al. (2021). CXCL1, CXCL10 and CXCL12 Chemokines Are Variously Expressed in Acute Myeloid Leukemia Patients Prior and Post Bone Marrow Transplantation. *Asian Pac. J. Cancer Prev.* 22 (10), 3377–3384. doi:10.31557/apjcp.2021.22.10.3377

Yi, J.-K., Park, S., Ha, J.-J., Kim, D.-H., Huang, H., Park, S.-J., et al. (2020). Effects of Dimethyl Sulfoxide on the Pluripotency and Differentiation Capacity of Mouse Embryonic Stem Cells. *Cell reprogramming* 22 (5), 244–253. doi:10.1089/cell.2020.0006

Zhang, S., Zhang, F., and Feng, L. (2021). The Inhibition of HeLa Cells Proliferation through SPARCL1 Mediated by SPP1. *Cytotechnology* 73 (1), 71–78. doi:10.1007/s10616-020-00443-2

Zhang, X., Zhao, H., Shi, X., Jia, X., and Yang, Y. (2020). Identification and Validation of an Immune-Related Gene Signature Predictive of Overall Survival in colon Cancer. *Aging* 12 (24), 26095–26120. doi:10.18632/aging.202317

Zhou, J., Yang, Y., Zhang, Y., Liu, H., and Dou, Q. (2018). A Meta-Analysis on the Role of Pleiotrophin (PTN) as a Prognostic Factor in Cancer. *PloS one* 13 (11), e0207473. doi:10.1371/journal.pone.0207473

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gu, Jiang, Xu, Liu and Zhou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 25 February 2022
doi: 10.3389/fcell.2022.816153

Integrated Analysis of Multi-Omics Alteration, Immune Profile, and Pharmacological Landscape of Pyroptosis-Derived lncRNA Pairs in Gastric Cancer

Chunguang Guo[1†], Zaoqu Liu[2†], Yin Yu[3†], Shirui Liu[1], Ke Ma[1], Xiaoyong Ge[2], Zhe Xing[4], Taoyuan Lu[5], Siyuan Weng[2], Libo Wang[6], Long Liu[6], Zhaohui Hua[1], Xinwei Han[2]* and Zhen Li[1]**

[1]Department of Endovascular Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2]Department of Interventional Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [3]Department of Pathophysiology, School of Basic Medical Sciences, The Academy of Medical Science, Zhengzhou University, Zhengzhou, China, [4]Department of Neurosurgery, The Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [5]Department of Cerebrovascular Disease, Zhengzhou University People's Hospital, Zhengzhou, China, [6]Department of Hepatobiliary and Pancreatic Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

Background: Recent evidence demonstrates that pyroptosis-derived long non-coding RNAs (lncRNAs) have profound impacts on the initiation, progression, and microenvironment of tumors. However, the roles of pyroptosis-derived lncRNAs (PDLs) in gastric cancer (GC) remain elusive.

Methods: We comprehensively analyzed the multi-omics data of 839 GC patients from three independent cohorts. The previous gene set enrichment analysis embedding algorithm was utilized to identify PDLs. A gene pair pipeline was developed to facilitate clinical translation *via* qualitative relative expression orders. The LASSO algorithm was used to construct and validate a pyroptosis-derived lncRNA pair prognostics signature (PLPPS). The associations between PLPPS and multi-omics alteration, immune profile, and pharmacological landscape were further investigated.

Results: A total of 350 PDLs and 61,075 PDL pairs in the training set were generated. Cox regression revealed 15 PDL pairs associated with overall survival, which were utilized to construct the PLPPS model *via* the LASSO algorithm. The high-risk group demonstrated adverse prognosis relative to the low-risk group. Remarkably, genomic analysis suggested that the lower tumor mutation burden and gene mutation frequency (e.g., *TTN*, *MUC16*, and *LRP1B*) were found in the high-risk group patients. The copy number variants were not significantly different between the two groups. Additionally, the high-risk group possessed lower immune cell infiltration abundance and might be resistant to a few chemotherapeutic drugs (including cisplatin, paclitaxel, and gemcitabine).

Conclusion: PDLs were closely implicated in the biological process and prognosis of GC, and our PLPPS model could serve as a promising tool to advance prognostic management and personalized treatment of GC patients.

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Xin Hu,
Tianjin Medical University Cancer Institute and Hospital, China
Yijing Su,
University of Pennsylvania, United States
Yufeng Gao,
First Affiliated Hospital of Anhui Medical University, China

***Correspondence:**
Zhaohui Hua
huazhaohuisfy@163.com
Xinwei Han
fcchanxw@zzu.edu.cn
Zhen Li
fccliz2@zzu.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *16 November 2021*
Accepted: *09 February 2022*
Published: *25 February 2022*

Citation:
Guo C, Liu Z, Yu Y, Liu S, Ma K, Ge X, Xing Z, Lu T, Weng S, Wang L, Liu L, Hua Z, Han X and Li Z (2022) Integrated Analysis of Multi-Omics Alteration, Immune Profile, and Pharmacological Landscape of Pyroptosis-Derived lncRNA Pairs in Gastric Cancer. Front. Cell Dev. Biol. 10:816153. doi: 10.3389/fcell.2022.816153

Keywords: gastric cancer, pyroptosis, lncRNA, gene pair, prognosis, mutation, immune landscape

INTRODUCTION

Gastric cancer (GC) is a serious disease worldwide, ranking the fifth in incidence and fourth in mortality globally (Smyth et al., 2020; Sung et al., 2021). Recently, the treatment modalities for GC have been gradually refined and optimized, but surgical resection is still the main modality (Fuchs et al., 2017; Smyth et al., 2020; Sung et al., 2021). Disappointingly, due to the highly heterogeneous and occult course, most patients with GC are already in advanced stage when diagnosed, with 769,000 deaths globally in 2020 and the 5-y survival rate of only about 25% (Rugge et al., 2017; Eusebi et al., 2020; Yuan et al., 2020; Sung et al., 2021). Traditionally, clinical decisions of GC have been specified by the American Joint Committee on Cancer (AJCC) tumor node metastasis (TNM) staging system (Edition et al., 2017). However, sometimes the prognosis of GC patients with the same TNM stages who receive the same treatment modalities is not identical. The molecular heterogeneity between individuals might drive the clinical diversity of GC patients (Akshatha et al., 2021). Therefore, it is necessary to incorporate molecular characteristics into the prognostic evaluation of patients.

Pyroptosis is recognized as a new type of programmed cell death, which is inherently inflammatory and triggered by various pathological stimuli (including infection, cancer, and cardiovascular events) (Ayers et al., 2017). Pyroptosis is characterized by gasdermin family-mediated cell swelling, lysis, and the release of pro-inflammatory intracellular contents including interleukin (IL)-18 and adenosine triphosphate (Kovacs and Miao, 2017). Pyroptosis was initially an important pathway to antagonize cell infection, and with the deepening of research, it was found to play a linchpin role in influencing the tumor microenvironment (TME) (Ye et al., 2021). Accumulating evidence has indicated that pyroptosis impacts the prognosis of patients in multiple cancers by promoting or inhibiting the proliferation, invasion, and metastasis of tumors (Fang et al., 2020; Al Mamun et al., 2021; Lin et al., 2021). For instance, recent research revealed that pyroptosis promoted the proliferation of esophageal squamous carcinoma cells through the *PKM2-caspase-8/3-GSDME* pathway (Jiang et al., 2021). Tian et al. (2020) reported that *GSDME*-mediated pyroptosis promoted the progression of colorectal cancer by releasing *HMGB1*. Therefore, pyroptosis has profound impacts on cancers.

Long non-coding RNAs (lncRNAs), a type of non-coding RNA molecules greater than 200 nucleotides in length, play important roles in regulating the transcriptional, posttranscriptional, or epigenetic level (Al Mamun et al., 2021). Emerging evidence has indicated that the interactions between pyroptosis and lncRNAs are significant in the development of multiple complex human diseases, especially in cancers (He et al., 2020; Gao et al., 2021). Due to the important role of lncRNAs in the process of tumor cell proliferation, migration, and apoptosis, the diagnosis and prognosis of patients are affected (Liu S. J. et al., 2021). The aberrant expression of lncRNAs could regulate the biological behavior of pyroptosis-derived lncRNA (PDLs) through miRNAs and downstream pathways, further prolonging or shortening the overall survival (OS) of multiple cancers (He et al., 2020). Scientists reported a new finding that knockdown of *rp1-85f18.6* significantly promotes the pyroptosis of colorectal cancer cells by cleaving *GSDMD* (Ma et al., 2018). These findings demonstrated that identifying and exploring PDLs could promote the revelation of the complex mechanisms of cancer. Nevertheless, the clinical significance of PDLs in GC remains elusive and needs to be further explored.

Large-scale data open the gate to comprehensively explore molecular alterations and systematically develop useful biomarkers in cancers (Cai et al., 2020; Liu et al., 2021d; Liu et al., 2021f). However, cross-platforms remain the non-negligible obstacle in biomarker discovery. Thus, in this study, we proposed a novel gene pair pipeline based on relative expression orders. Using this pipeline, we successfully constructed and validated a pyroptosis-derived lncRNA pair prognostics signature (PLPPS) model with robust performance across multiple independent datasets. More importantly, we also revealed the multi-omics alteration, immune profile, and pharmacological landscape of PLPPS.

MATERIALS AND METHODS

Data Collection and Processing

The flowchart of this study is shown in **Figure 1**. Three cohorts were enrolled from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) and Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) datasets, including TCGA-STAD (n = 348), GSE62254 (n = 300), and GSE15458 (n = 191). For the TCGA-STAD cohort, "level 3" transcriptome profile and clinical data of patients were retrieved. All GEO datasets were annotated by the Affymetrix® GPL570 platform, and raw transcriptome data were processed by the robust multiarray averaging (RMA) algorithm implemented in "affy" R package (Deandres-Galiana et al., 2016). In three cohorts, primary tumor tissue samples without chemoradiotherapy and cancer-adjacent normal tissues were retained. The somatic mutation and copy number variation (CNV) data were all downloaded from the TCGA portal and cBioPortal, respectively. The baseline characteristics are summarized in **Supplementary Table S1**.

Pyroptosis Gene List

The pyroptosis gene list (containing a total of 51 genes) was required from the MSigDB (Liberzon et al., 2015) (version 7.4, REACTOME_PYROPTOSIS) and prior research (Shao et al., 2021; Ye et al., 2021). The gene list is illustrated in **Supplementary Table S2**.

Identification of lncRNAs Associated With Pyroptosis Genes

Using our previous integrated algorithm (Liu et al., 2021c; Liu et al., 2021e), we aimed to identify the potential lncRNA regulators of pyroptosis. The summary is given as follows: 1) total mRNAs were ranked in descending order *via* their correlation with a specific

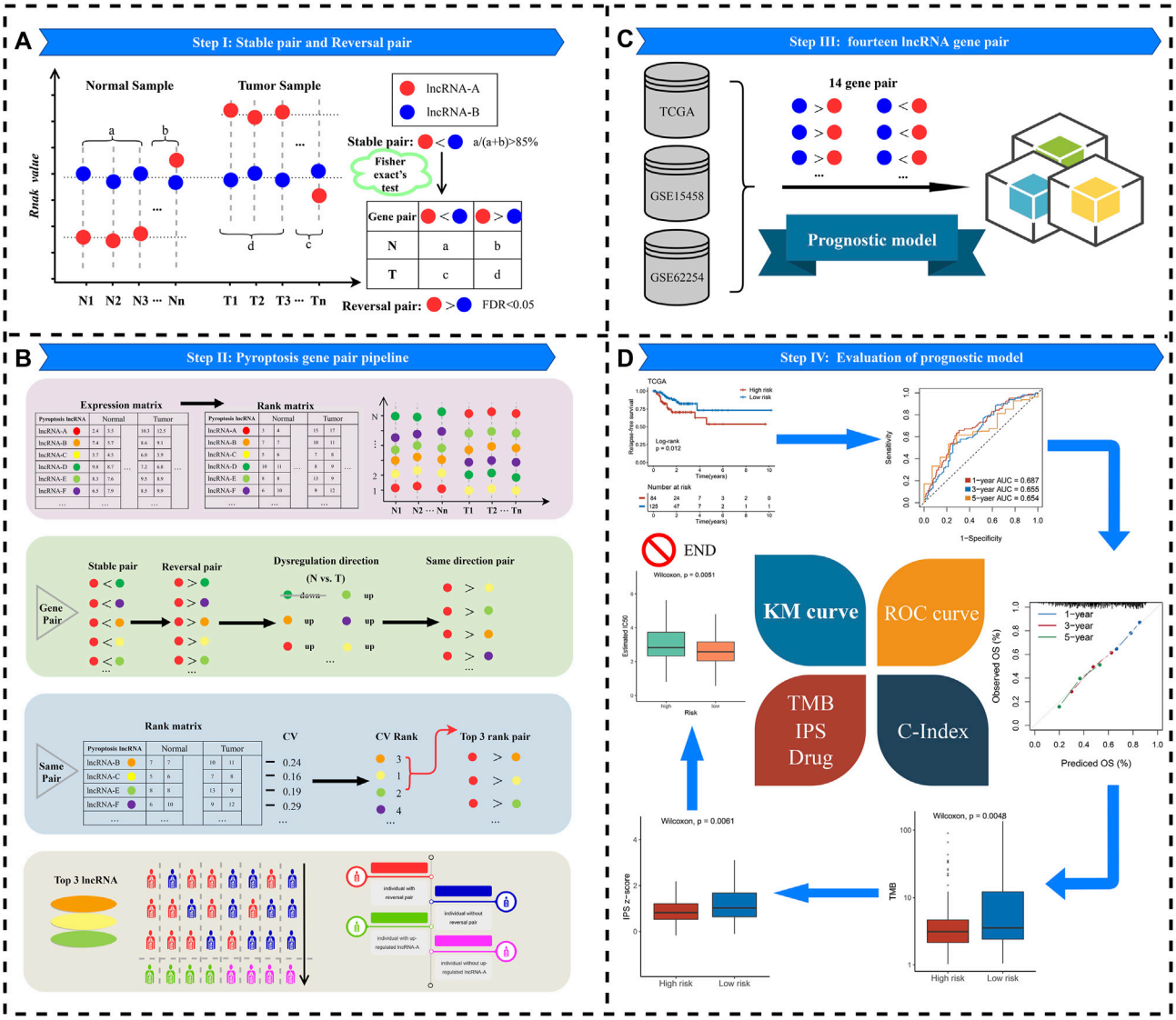


FIGURE 1 | Flowchart of analysis procedure.

lncRNA. 2) The order gene list was further entered into "fgsea" R package to perform enrichment analysis and explore whether the genes of pyroptosis signaling were enriched in the top or bottom of the ordered list. 3) The pyroptosis enrichment score (PES) was calculated for total lncRNAs. LncRNAs with significant PES were determined as pyroptosis-derived lncRNA by the permutation test framework. 4) The lncRNAs with the absolute value of PES > 0.995 and the false discovery rate (FDR) < 0.05 were determined as pyroptosis-derived lncRNAs.

Definition of Stable and Reversal lncRNA Pairs

The lncRNA expression matrix was converted to the corresponding rank matrix (the greatest expression value was converted to the maximum rank, and the smallest expression value was converted to the minimum rank in each sample) (**Figure 1A**). Pairwise comparisons of all pyroptosis-derived lncRNAs were performed to determine stable pairs in normal samples and reversed pairs in tumor samples. When two lncRNAs exhibited the same rank pattern (e.g., lncRNA-A < lncRNA-B) in 85% of normal samples, these two lncRNAs were called stable pairs. Likewise, reversal pair was defined as displaying the uniform rank pattern in 85% of tumor samples and in the opposite direction to stable pairs (e.g., lncRNA-A < lncRNA-B).

Gene Pair Pipeline

In order to screen suitable reversal pyroptosis gene pairs, we performed the following steps accordingly (**Figure 1B**):

Step 1: The expression profile of lncRNA was converted to the rank profile, and that will reduce batch effects during the experiment and qualitative transcriptional characterization.

Step 2: The identification of lncRNA pairs with the same direction of deregulation. For instance, lncRNA-A with lncRNA-B, lncRNA-C, lncRNA-D, lncRNA-E, and lncRNA-F constituted stable pair and reversal pair relationships in normal vs. tumor samples, respectively. Compared with normal samples, the rank of lncRNA-A in tumor samples was "up." However, the direction of deregulation of lncRNA-D was "down." Then, the combination of lncRNA-A with lncRNA-D will be abandoned.

Step 3: Subsequently, we calculated the coefficient of variation (CV) of partner lncRNAs for each lncRNA-A through the rank change of lncRNAs in tumor and normal tissues. Assuming that if the order of lncRNA pair is approximately invariable in normal and tumor samples, the stable and reversal relationship between lncRNA-A and partner lncRNAs may be altered because of the order change of lncRNA-A in the two groups, which facilitates judging whether lncRNA-A exists different in individual tumor samples. Based on the aforementioned reasons, we calculated the CV of patten lncRNA for lncRNA-A and ranked it in descending order. After that, the top three patten lncRNAs of the CV (lncRNA-B, lncRNA-C, and lncRNA-E) were retained and included in the following analysis (Peng et al., 2017).

Step 4:Ultimately, three lncRNA pairs (lncRNA-A > lncRNA-B, lncRNA-A > lncRNA-C, and lncRNA-A > lncRNA-E) were considered as qualified reversal pairs in every single tumor sample. For instance, more than 60% of the tumor samples had lncRNA-A > lncRNA-B, and lncRNA-A > lncRNA-B was included in the final study (the green human shape in **Figure 1B**).

Step 5: Relative expression ordering. For each lncRNA pair (e.g., lncRNA-A and lncRNA-B), we defined lncRNA-A > lncRNA-B as number 1. In contrast, number 0 was used to denote lncRNA-A < lncRNA-B. Passing this transformation will reduce the experimental batch effect and improve the robustness. Then, we obtained a matrix of samples based on the relative expression order of lncRNA pairs for subsequent research.

Signature Generation

First, the stable PDL pairs were identified according to the following guideline: 1) *p*-value of univariate Cox regression less than 0.05. 2) Hazard ratio (HR) values were in the consistent direction (HR > 1 or <1) in three cohorts. Subsequently, based on the rank value of the PLPPS, we used the LASSO regression algorithm (Lockhart et al., 2014) to develop a PLPPS model to predict the OS of GC patients. To achieve this purpose, the calculation process was performed in "glmnet" R package and the lambda value was chosen when the partial likelihood deviance reached the minimum value by the 10-fold cross-validations (Friedman et al., 2010). Finally, the lncRNA pairs with non-zero coefficients were selected to establish the final prognosis model. Patients were divided into high- and low-risk groups according to the optimal cutoff point. The predictive ability of the PLPPS model was validated in GSE62254 and GSE15458.

Copy Number Variation Analysis

Afterward, we performed a comprehensive analysis of copy number deletion and amplification of two groups. The top 10 genes in the frequency of homozygously deleted (HOMDEL) and high-level gene amplification (AMP) events were depicted based on the CNV data from cBioPortal.

Immunogenicity Characterization Analysis

In order to decipher the immunogenicity of the two groups, tumor mutation burden (TMB) and immunophenoscore (IPS) of patients were explored. Using the "*tmb*" function in "maftool" R package, we calculated the difference of TMB between the two group patients. After that, the IPS was used to assess the immune state of each patient. IPS is an assessment protocol that quantifies tumor immunogenicity through multiple immune markers (Charoentong et al., 2017). This score has a positive correlation with immunogenicity.

Delineate the Mutation Landscape

We compared molecular mutation differences between high- and low-risk groups. The calculation of genes with the top 20 mutation frequencies and the depiction of mutation waterfall plots were performed by "maftools" R package.

Gene Set Enrichment Analysis

The differential genes between the high- and low-risk groups were identified by "limma" package and sequenced by the log2 (fold change) value. Subsequently, gene set enrichment analysis (GSEA) was used to decipher the underlying biological mechanisms of the GC sample using GO and KEGG terms (Molecular Signatures Database, version: c5.go.v7.4.symbols.gmt and c2.cp.kegg.v7.4.symbols.gmt).

Immune Landscape Description

Tumor infiltration by immune and stromal cells was evaluated by the MCPcounter algorithm (Becht et al., 2016), implemented in "MCPcounter" R package. Additionally, to analyze the response to immune checkpoint blockade, the two groups' samples were scored by T-cell inflammatory signature (TIS). TIS was constituted by 18 inflammatory genes associated with adaptive immune resistance, antigen presentation, and chemokine expression (Ayers et al., 2017).

Chemotherapy Drug Response

In order to assess the clinical significance of the PLPPS model, we analyzed the response [evaluated by half maximal inhibitory concentration (IC50)] to chemotherapy drugs between the high- and low-risk groups. This process was implemented in "pRRophetic" R package.

Statistical Analysis

All data processing, statistical analysis, and plotting were conducted in R 4.0.3 software. Continuous variables and categorical variables were compared between two groups using the Wilcoxon rank-sum test and Fisher's exact test, respectively. The optimal cutoff point was used to classify patients into high- and low-risk groups. The Kaplan–Meier method and the log-rank test were utilized to estimate the different OS and relapse-free survival (RFS) between two groups. The Benjamin–Hochberg method was used to further calculate the FDR. For every analysis, statistical significance was considered at $p < 0.05$.

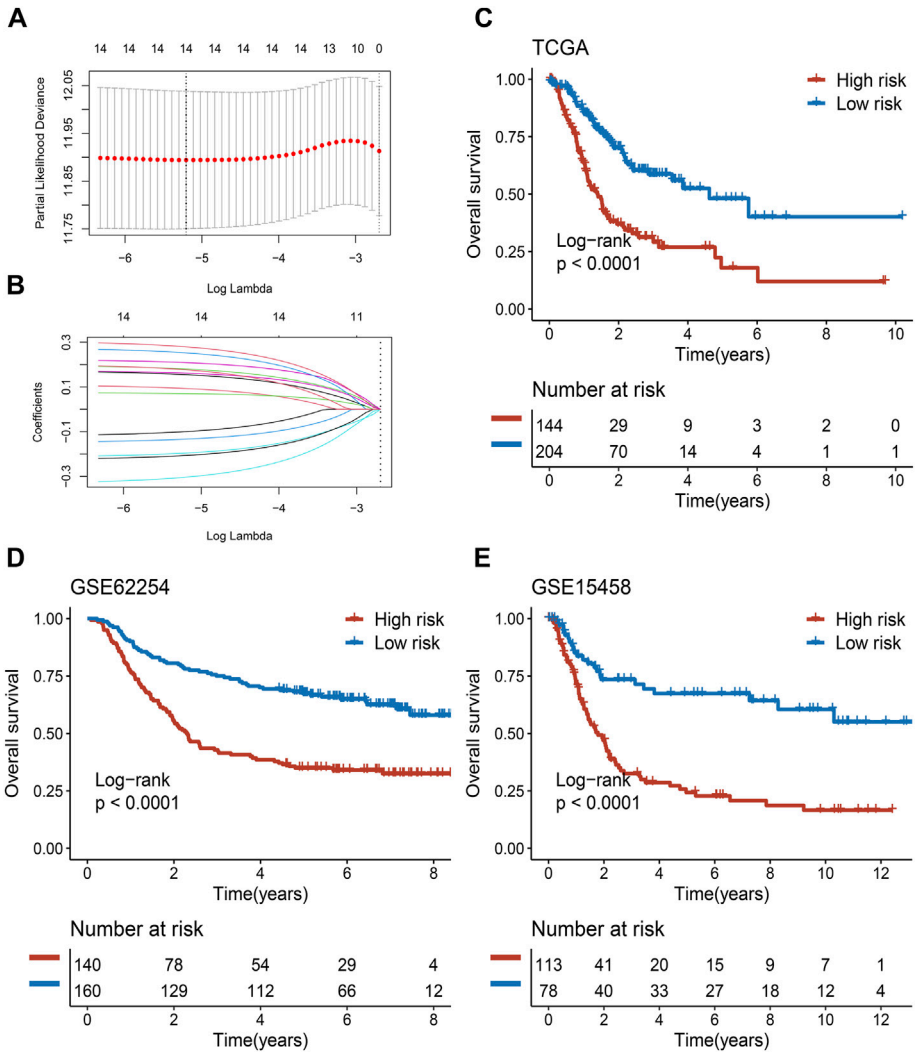


FIGURE 2 | Development and validation of the PPSM model in clinical samples. **(A,B)** Least absolute shrinkage and selection operator (LASSO) logistic regression algorithm to screen of gene pairs associated with prognostic. **(C–E)** Kaplan–Meier curves of OS according to the high- and low-risk groups in TCGA **(C)**, GSE62254 **(D)**, and GSE15458 **(E)** cohorts.

RESULTS

Identification of PDLs and Reversal lncRNA Pairs

We identified PDLs through the integration lncRNA pipeline. Based on the lncRNA and mRNA expression datasets, the pipeline could systematically and accurately trace lncRNA molecules which are potentially linked to pyroptosis genes. Thus, a total of 350 PDLs were selected for the following research. Subsequently, PDL expression profiles were transformed to rank profiles in each GC sample. A total of 61,075 lncRNA pairs were obtained from 350 PDLs. According to the pipeline in **Figure 1B**, we identified 5,046, 7,753, and 2,373 reversal lncRNA pairs from TCGA-STAD, GSE62254, and GSE15458, respectively. The three lists of lncRNA pairs had 430 overlaps. Afterward, to evaluate the performance of reversal pairs between different platforms, the common 430 lncRNA pairs between sequencing and microarray datasets underwent all analyses.

Construction and Validation of the PLPPS Model

Among the 430 lncRNA pairs, we identified 15 PDLs correlated with OS by univariate Cox analysis in the three cohorts. Based on the rank profile of these lncRNA pairs in TCGA-STAD, we fitted a LASSO Cox regression model and identified 14 PDLs that were highly predictive of OS (**Figures 2A,B**). The specific calculation formula of risk score is given in **Table 1**. All patients were assigned to the high- and low-risk groups according to the optimal cutoff point. Compared with the low-risk group, patients in the high-risk group had dismal OS in all cohorts

TABLE 1 | List of gene pairs and corresponding coefficient.

Signature	Gene A	Gene B	Coefficient
Pair 1	LINC00607	C5orf17	0.2793
Pair 2	TUSC8	PITRM1-AS1	0.2525
Pair 3	TRPM2-AS	RP11-579D7.4	0.2088
Pair 4	AC074286.1	AC058791.1	0.1802
Pair 5	MMP25-AS1	RP11-876N24.5	0.1766
Pair 6	LINC01094	RP4-680D5.8	0.1616
Pair 7	AC013275.2	RP11-567C2.1	0.1548
Pair 8	MLLT4-AS1	RP11-21L23.2	0.0930
Pair 9	LINC00607	RP11-109E24.1	0.0700
Pair 10	C10orf91	TRPM2-AS	–0.0999
Pair 11	LINC01588	RP11-73K9.2	–0.1337
Pair 12	RP3-522D1.1	LINC01094	–0.1959
Pair 13	RP11-61A14.1	RP11-416I2.1	–0.2067
Pair 14	CTD-2377D24.6	LINC01169	–0.3038

(log-rank test, all $p < 0.05$; **Figures 2C–E**). In addition, PLPPS remained an independent prognostic factor for GC patients after controlling the available clinical characteristic in the three cohorts (all $p < 0.05$; **Figures 3A–C**). Of note, patients in the high-risk group also had significantly worse RFS (**Supplementary Figure S1**). Multivariate Cox regression analysis displayed that PLPPS was still an independent prognostic factor for RFS ($p < 0.05$; **Supplementary Figure S2**).

Assessment of the PLPPS Model

In our research, we assessed this model from two perspectives: discrimination and calibration. To achieve this purpose, ROC curves and calibration plots were utilized to evaluate the PLPPS model. As shown in **Figures 3D–F**, the AUCs for predicting OS at 1, 3, and 5 years were 0.792, 0.779, and 0.777 in TCGA-STAD; 0.788, 0.804, and 0.790 in GSE62254; and 0.772, 0.813, and 0.806 in GSE15458, respectively. To evaluate the accuracy of the model's prediction results, we calculated the concordance index of the PLPPS model, which were 0.737 (95% confidence interval (CI): 0.694–0.781), 0.722 (95%CI: 0.684–0.760), and 0.723 (95%CI: 0.676–0.769) in the three cohorts, respectively (**Supplementary Figure S3**). In addition, calibration plots displayed excellent calibration of the model that indicated the predicted probability of OS at 1, 3, and 5 years was accurate (**Figures 3G–I**).

Copy Number Variation Analysis

Subsequently, we performed a comprehensive analysis of CNV in the two groups of patients. As presented in **Figure 4A**, we calculated the top ten genes with deletion and amplification in the two groups, respectively. We found that there was no significant difference in copy number deletion between the groups, and a high similarity was presented in the top ten genes of deletion between the two groups of patients, such as *WWOX*, *CCSER1*, *PDE4D*, and *PTPRD*. Notably, compared with low-risk group patients, patients in the high-risk group had scarcer AMP, such as *PGAP3*, *MIEN1*, *GRB7*, and *ERBB2*. The significant difference in AMP could be a potential factor for dismal prognosis in high-risk group patients.

Immunogenicity Characterization

Tumors with higher TMB are more likely to express recognizable neoantigens that enhance the immune response. Similarly, IPS is a scoring scheme to assess the immunogenicity of tumor samples. Based on the earlier reasons, we further assessed the difference in TMB and IPS scores between the high- and low-risk groups. Notably, compared with the low-risk group, patients in the high-risk group showed inferior TMB and IPS scores (**Figures 4B,C**).

Landscape of Gene Mutations in STAD

The waterfall plot delineated the top 20 high-frequency genes in the high- and low-risk groups (**Figure 4D**). Subsequently, we compared the frequency difference of the top 20 genes in the two groups. Overall, patients in the high-risk group displayed lower mutation frequency, especially in *MUC16*, *ARID1A*, *FAT4*, *DNAH5*, etc. (**Figure 4E**).

Gene Set Enrichment Analysis

GSEA revealed significant GO biological processes (**Figures 5A,B**) and KEGG pathways (**Figures 5C,D**) between the two risk groups. For biological processes, the high-risk group mainly focused on supporting and development of cellular structures, such as "chondrocyte differentiation," "collagen fibril organization," and "extracellular matrix organization" (**Figure 5A**); the low-risk group mainly acted on the generation and transmission of energy, such as "ATP synthesis coupled electron transport" and "mitochondrial respiratory chain complex assembly" (**Figure 5B**). For KEGG pathways, the high-risk group mainly focused on the intercellular adhesion, such as "cell adhesion molecules" and "focal adhesion" (**Figure 5C**); the low-risk group mainly acted on the inflammatory and immune infiltration-related functions, such as "B-cell activation" and "interleukin-6 production" (**Figure 5D**). Together, compared with the low–high group, high-risk group patients were not satisfactory in the enrichment of immune.

Immune Infiltration Analysis and Prediction of Chemotherapy Response

To further explore the infiltration abundance of immune cells between the high- and low-risk groups, the MCPcounter algorithm was utilized. We found higher overall infiltration in the low-risk group than in the high-risk group (**Figure 6A**); especially, the high-risk group possessed inferior levels of macrophage monocytes, myeloid dendritic cells, endothelial cells, and cancer-associated fibroblasts compared with the low-risk group (**Figure 6B**). Ultimately, we have chosen the cisplatin (CDDP, platinum chemotherapy drug), paclitaxel (PTX, taxus chemotherapy drug), gemcitabine (GEM, pyrimidines drug), and doxorubicin (ADM, anthracycline chemotherapy drug) to compare the responses of patients in the two subtypes to chemotherapy drugs. Remarkably, compared with the high-risk group, IC50 of the four drugs was lower in the low-risk group (**Figures 6C–F**).

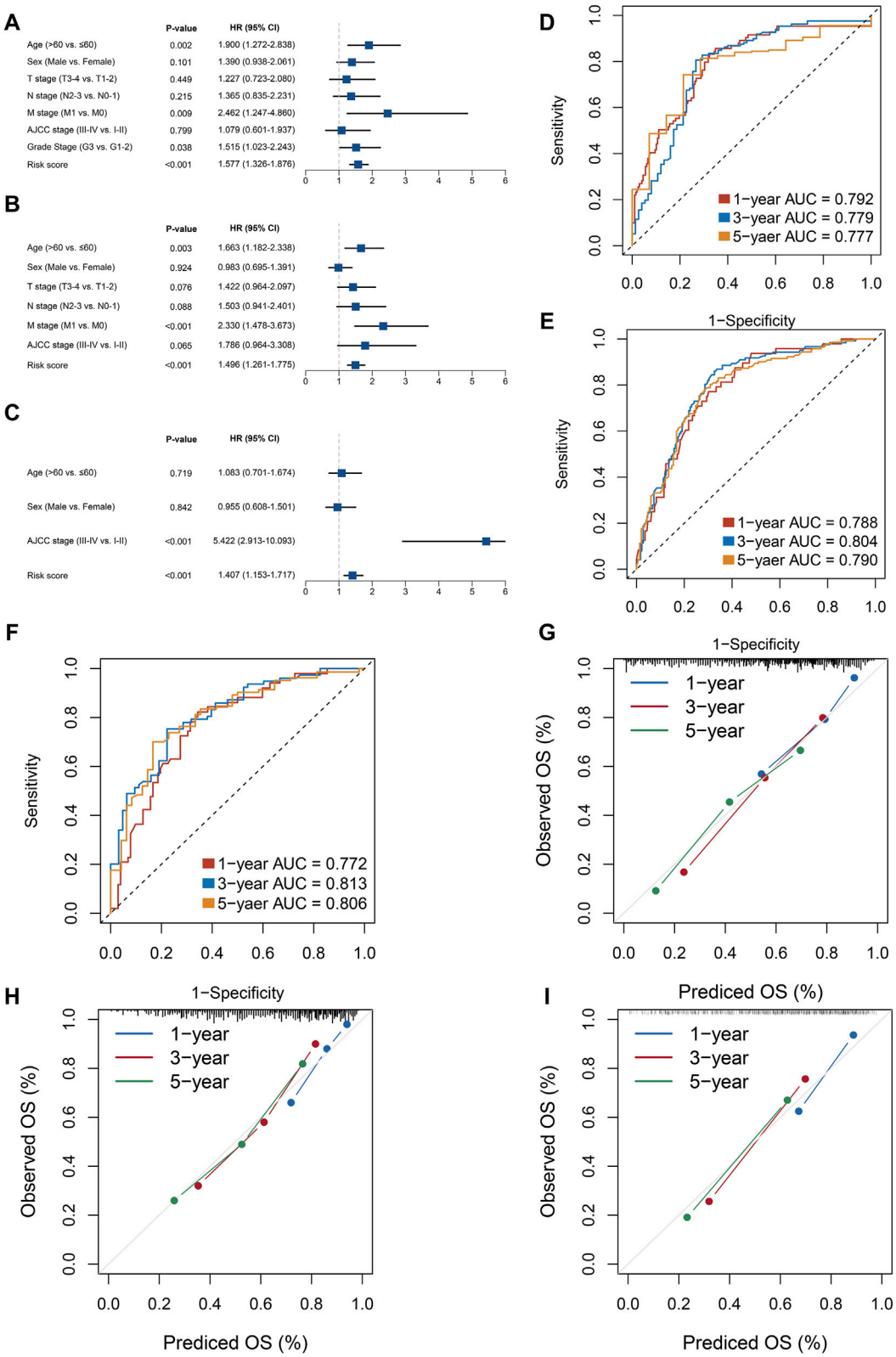


FIGURE 3 | Evaluation of the PPSM model effectiveness in three cohorts. **(A–C)** Multivariate COX regression analysis of the risk score in the three cohorts: TCGA **(A)**, GSE62254 **(B)**, and GSE15458 **(C)**. **(D–F)** ROC analysis for the three cohorts: TCGA **(D)**, GSE62254 **(E)**, and GSE15458 **(F)**. **(G–I)** Calibration plots were used to compare the actual probabilities and the predicted probabilities of OS in the three cohorts: TCGA **(G)**, GSE62254 **(H)**, and GSE15458 **(I)**.

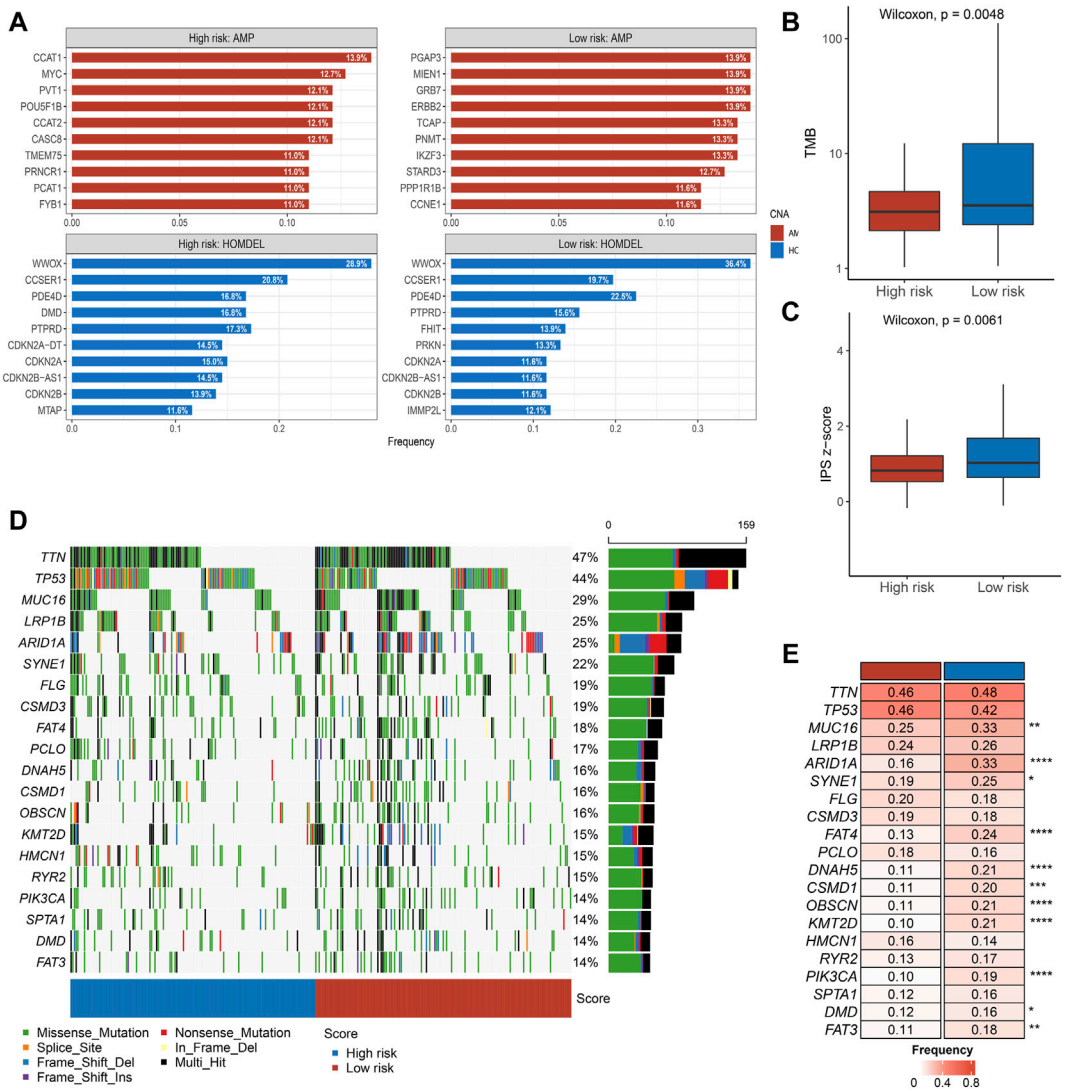


FIGURE 4 | Multi-omics alteration of the PLPPS model **(A–C)**. The differences in CNV **(A)**, TMB **(B)**, and IPS **(C)** between the two groups. **(D)**. Waterfall plot of 20 frequently segments in the two subtypes **(E)**. Comparison of mutation frequency of top 20 segments.

DISCUSSION

Numerous reports indicated that pyroptosis exerts a dual role in tumor progression and treatment (He et al., 2020; Loveless et al., 2021). The double-edged sword roles were manifested in the phenomenon that pyroptosis promotes the evolution of normal cells into tumor cells by releasing inflammatory factors and further lead to programmed cell death of tumor cells through the infectious pathway to inhibit tumor progression (Kovacs and Miao, 2017; Ruan et al., 2020; Al Mamun et al., 2021). However, most research studies related to pyroptosis in tumors have focused on the encoded protein level, while non-coding RNAs that play an important role in apoptosis vs. cycle regulation processes have been neglected (Song et al., 2021; Ye et al., 2021; Zhao et al., 2021). Likewise, PDLs in GC also awaits further investigation. In order to eliminate the errors caused by sequencing platforms and different sample cohorts and to further improve the rigor of our research, we applied a new algorithm (Qi et al., 2016; Peng et al., 2017; Chen et al., 2018; Zhao et al., 2021) based on the gene expression ranks to reveal for the first time the impact of PDLs on patient prognosis in GC and delineate the different immune and mutational landscapes of the high- and low-risk groups.

Prognostic prediction is crucial for further interventions. In our study, we first identified 14 PDLs and elucidated their prognostic value in three independent GC cohorts. The prognostic pyroptosis signature can successfully categorize patients into subtypes with different survival outcomes. Subsequently, we found that genomic differences in patients with subgroups were abundant. As is well known, the TMB is associated with tumor immunity, and low TMB is a poor prognostic factor in GC (Roh et al., 2017; Ma et al., 2021).

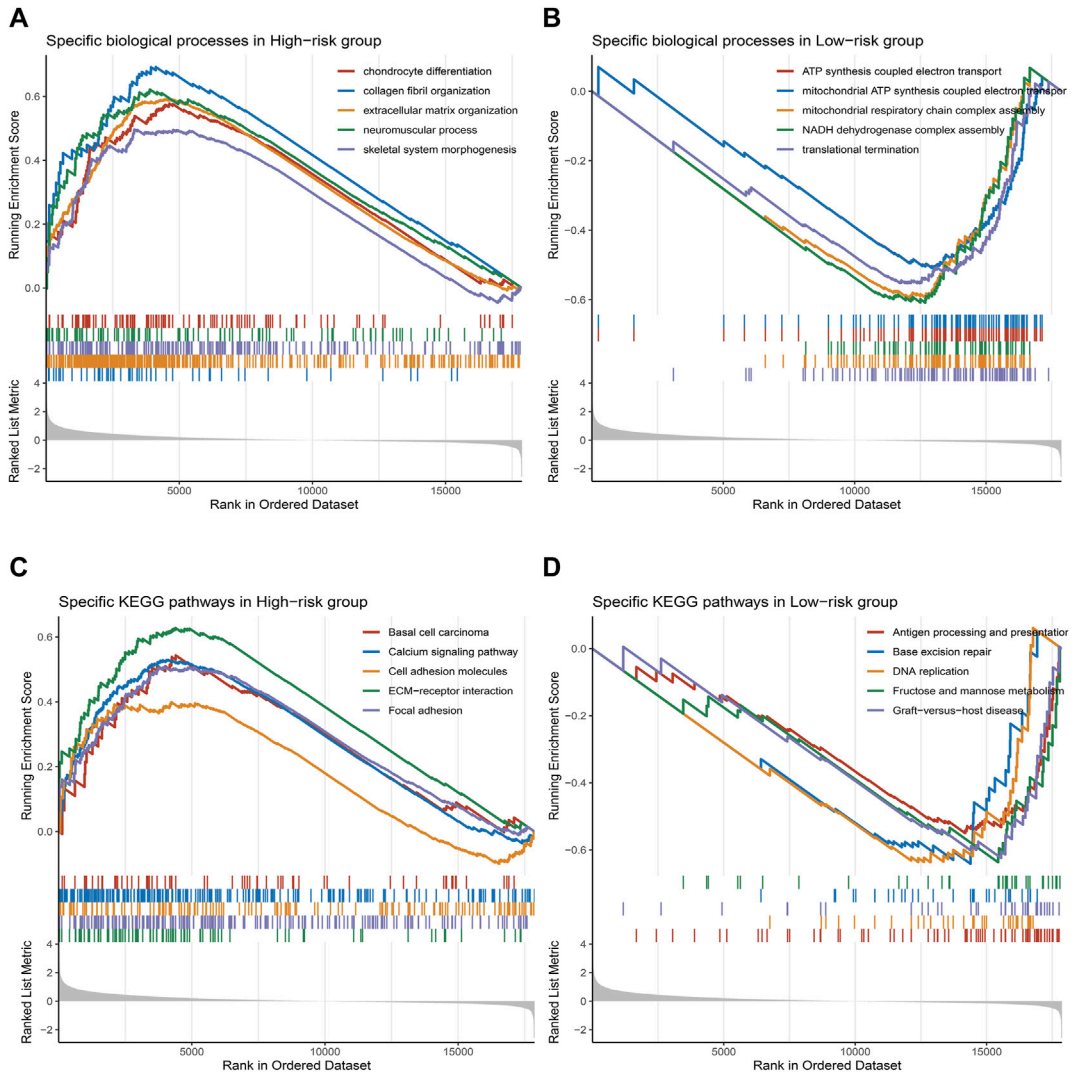


FIGURE 5 | GSEA and prediction of chemotherapy response. **(A–D)**. The top five GO terms **(A,B)** and KEGG pathways **(C,D)** in the two subtypes.

This is consistent with the results of the correlation analysis in our study, where patients in the high-risk group had lower TMB. Additionally, patients in the high-risk group had a higher frequency of *TP53* mutation, which is perhaps the reason for the shorter OS and RFS of patients in the high-risk group (Wang et al., 2021).

To further reveal the intrinsic biological mechanisms underlying the differences in prognostic among patients with subtypes, we performed GSEA for patients in the high- and low-risk groups. We found that inflammation and immune response-related pathways were particularly scarce in high-risk patients. Interestingly, as reported by Rao et al. (2021), "immune desert" tumors indicate adverse clinical outcomes. In addition, numerous studies indicate that high immune cell infiltration in GC patients was associated with a better long-term prognosis (Liu et al., 2020; Akshatha et al., 2021; Li et al., 2021). Notably, our investigation further suggested that differences between the abundance of immune cell infiltrates and the expression levels of immune checkpoints among the two groups. On comparing these levels in the high- and low-risk groups, we found that the levels of macrophage monocytes, myeloid dendritic cells, T cells, and CD8+ T cells were deserted in the high-risk group. On the other hand, we found that IPS scores representing immune checkpoints were similarly inferior in the high-risk patients. Our findings were consistent with the previous studies (Liu et al., 2020; Li et al., 2021; Rao et al., 2021), which indicate that PLPPS could guide the prognosis of GC patients. Different from the previous study focusing on the prognosis of GC, we further predicted sensitivity to chemotherapeutic agents in high- and low-risk group patients. The first-line chemotherapy drugs for GC such as cisplatin, paclitaxel, and gemcitabine (Joshi and Badgwell, 2021) showed higher sensitivity in the low-risk group. This further demonstrates that PLPPS has the

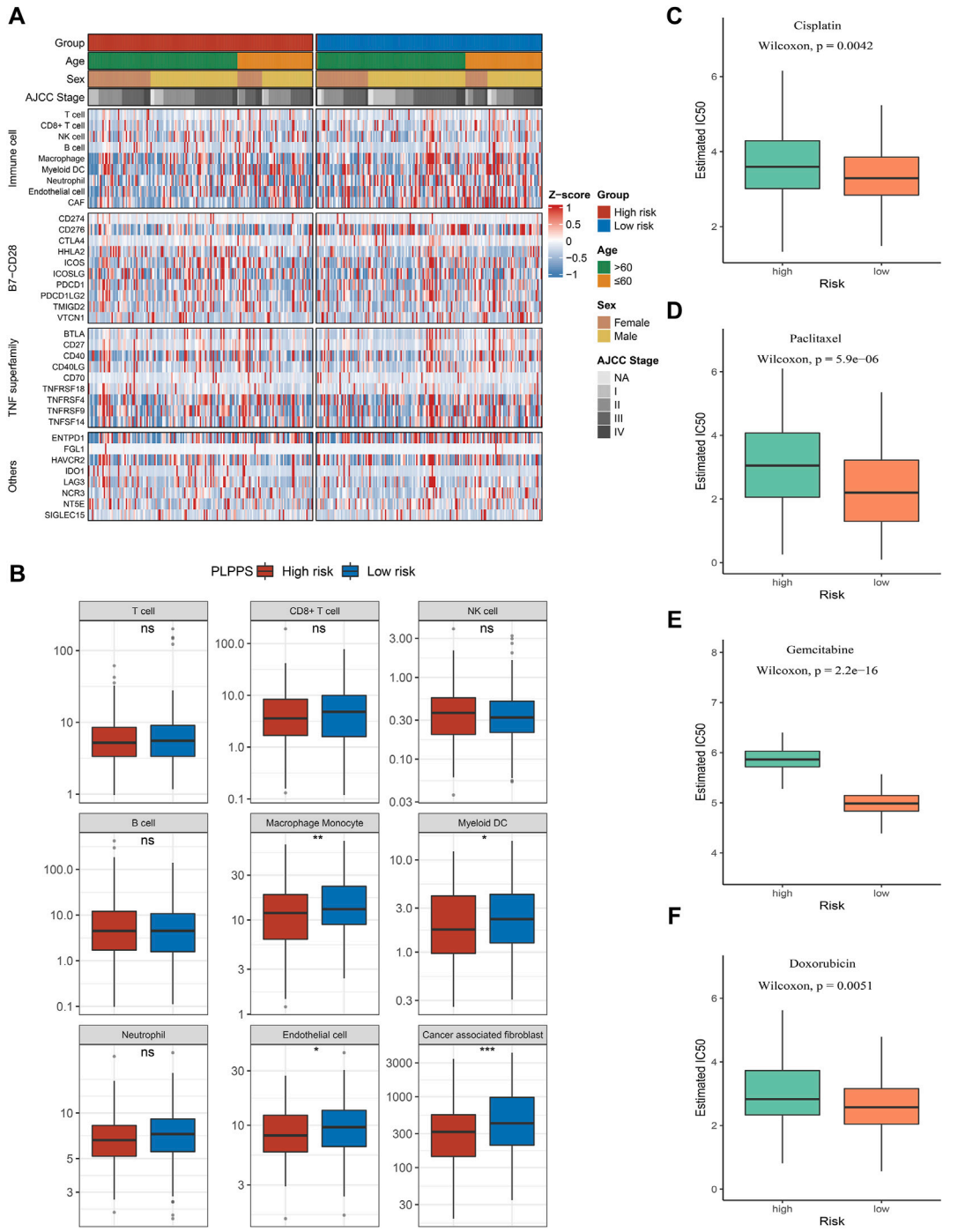


FIGURE 6 | Immune infiltration analysis and prediction of chemotherapy response in the two groups. **(A)** Heatmap for immune analysis among the two subtypes. **(B)** Boxplots of nine immune cell enrichment levels **(C–F)**. Sensitivity of cisplatin **(C)**, paclitaxel **(D)**, gemcitabine **(E)**, and doxorubicin **(F)** in the two subtypes.

potential to significantly contribute to identifying high-risk patients in the clinics.

Following the development of high-throughput genetic sequencing technology, we are faced with a large amount of biological data. With the combination of multi-omics with advanced algorithms, researchers can explore the molecular characteristics of GC patients in detail. This has led to the emergence of abundant prognostic and predictive gene expression signatures (Liu et al., 2021b; Wang et al., 2021; Zhang et al., 2021; Zhao et al., 2021). Nevertheless, these signatures are difficult to implement clinically due to the difficulty in unifying sequencing platforms and samples. To

address these issues, we used the gene pair algorithm (Qi et al., 2016; Peng et al., 2017). As the gene pair was calculated based on the gene expression ranking of tumor samples, they could be personalized for various platform data without normalization. This indicates that our signature built from 14 PDLs more clinically applicable.

Our work is the first to comprehensively evaluate the prognostic characteristics of GC patients by PDLs. Our research has the following advantages: 1) we identified PDLS by a novel algorithm: lncRNA pipeline. 2) The gene expression order was used, instead of gene expression quantity, which eliminates the influence of different platforms and cohorts on the prognostic model and improves its applicability. 3) The PLPPS model was validated in multiple independent cohorts, ROC curves, concordance index, and calibration plots all displaying that the model has high accuracy. 4) This study comprehensively depicted the immune and mutational landscape of patients in two GC subgroups and predicted their differential sensitivity to multiple chemotherapeutic agents that demonstrate the accuracy and clinical applicability of the model from multiple aspects. Although further biological validation is necessary, we believe that comprehensive analysis based on the multicenter and the larger sample can compensate for the shortcoming.

CONCLUSION

In conclusion, we proposed a gene pair pipeline that could ignore the platform batch. Based on this pipeline, the PLPPS consisting of 14 PDL gene pairs was constructed and validated across multiple independent cohorts. This signature was verified to serve as a promising and reliable prognostic tool in GC. In addition, the multi-omics alteration, immune profile, and pharmacological landscape of PLPPS were further revealed, which could provide new insights for individualized treatment and management of GC patients.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

CG and ZL designed this work. CG, YY, SL, KM, XG, and ZX integrated and analyzed the data. CG, TL, SW, LW, LL, ZH, and XH wrote this manuscript. CG and ZL edited and revised the manuscript. All authors approved this article.

FUNDING

This study was supported by the National Natural Science Foundation of China (81873527).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.816153/full#supplementary-material

REFERENCES

Akshatha, C. R., Bhat, S., Sindhu, R., Shashank, D., Rose Sommano, S., Tapingkae, W., et al. (2021). Current Therapeutic Options for Gastric Adenocarcinoma. *Saudi J. Biol. Sci.* 28, 5371–5378. doi:10.1016/j.sjbs.2021.05.060

Al Mamun, A., Mimi, A. A., Aziz, M. A., Zaeem, M., Ahmed, T., Munir, F., et al. (2021). Role of Pyroptosis in Cancer and its Therapeutic Regulation. *Eur. J. Pharmacol.* 910, 174444. doi:10.1016/j.ejphar.2021.174444

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., et al. (2017). IFN-γ-related mRNA Profile Predicts Clinical Response to PD-1 Blockade. *J. Clin. Invest.* 127, 2930–2940. doi:10.1172/JCI91190

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., et al. (2016). Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression. *Genome Biol.* 17, 218. doi:10.1186/s13059-016-1070-5

Cai, W.-Y., Dong, Z.-N., Fu, X.-T., Lin, L.-Y., Wang, L., Ye, G.-D., et al. (2020). Identification of a Tumor Microenvironment-Relevant Gene Set-Based Prognostic Signature and Related Therapy Targets in Gastric Cancer. *Theranostics* 10, 8633–8647. doi:10.7150/thno.47938

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cel Rep.* 18, 248–262. doi:10.1016/j.celrep.2016.12.019

Chen, P.-f., Wang, F., Zhang, Z.-x., Nie, J.-y., Liu, L., Feng, J.-r., et al. (2018). A Novel Gene-Pair Signature for Relapse-free Survival Prediction in colon Cancer. *Cmar* Vol. 10, 4145–4153. doi:10.2147/CMAR.S176260

Deandrés-Galiana, E. J., Fernández-Martínez, J. L., Saligan, L. N., and Sonis, S. T. (2016). Impact of Microarray Preprocessing Techniques in Unraveling Biological Pathways. *J. Comput. Biol.* 23, 957–968. doi:10.1089/cmb.2016.0042

Edition, S., Edge, S., and Byrd, D. (2017). *AJCC Cancer Staging Manual*.

Eusebi, L. H., Telese, A., Marasco, G., Bazzoli, F., and Zagari, R. M. (2020). Gastric Cancer Prevention Strategies: A Global Perspective. *J. Gastroenterol. Hepatol.* 35, 1495–1502. doi:10.1111/jgh.15037

Fang, Y., Tian, S., Pan, Y., Li, W., Wang, Q., Tang, Y., et al. (2020). Pyroptosis: A New Frontier in Cancer. *Biomed. Pharmacother.* 121, 109595. doi:10.1016/j.biopha.2019.109595

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. doi:10.18637/jss.v033.i01

Fuchs, C. S., Niedzwiecki, D., Mamon, H. J., Tepper, J. E., Ye, X., Swanson, R. S., et al. (2017). Adjuvant Chemoradiotherapy with Epirubicin, Cisplatin, and Fluorouracil Compared with Adjuvant Chemoradiotherapy with Fluorouracil and Leucovorin after Curative Resection of Gastric Cancer: Results from CALGB 80101 (Alliance). *Jco* 35, 3671–3677. doi:10.1200/JCO.2017.74.2130

Gao, J., Chen, X., Wei, P., Wang, Y., Li, P., and Shao, K. (2021). Regulation of Pyroptosis in Cardiovascular Pathologies: Role of Noncoding RNAs. *Mol. Ther. - Nucleic Acids* 25, 220–236. doi:10.1016/j.omtn.2021.05.016

He, D., Zheng, J., Hu, J., Chen, J., and Wei, X. (2020). Long Non-coding RNAs and Pyroptosis. *Clinica Chim. Acta* 504, 201–208. doi:10.1016/j.cca.2019.11.035

Jiang, M., Wu, Y., Qi, L., Li, L., Song, D., Gan, J., et al. (2021). Dihydroartemisinin Mediating PKM2-Caspase-8/3-GSDME axis for Pyroptosis in Esophageal Squamous Cell Carcinoma. *Chemico-Biological Interactions* 350, 109704. doi:10.1016/j.cbi.2021.109704

Joshi, S. S., and Badgwell, B. D. (2021). Current Treatment and Recent Progress in Gastric Cancer. *CA A. Cancer J. Clin.* 71, 264–279. doi:10.3322/caac.21657

Kovacs, S. B., and Miao, E. A. (2017). Gasdermins: Effectors of Pyroptosis. *Trends Cel Biol.* 27, 673–684. doi:10.1016/j.tcb.2017.05.005

Li, K., Zhang, A., Li, X., Zhang, H., and Zhao, L. (2021). Advances in Clinical Immunotherapy for Gastric Cancer. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1876, 188615. doi:10.1016/j.bbcan.2021.188615

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cel Syst.* 1, 417–425. doi:10.1016/j.cels.2015.12.004

Lin, W., Chen, Y., Wu, B., Chen, Y., and Li, Z. (2021). Identification of the Pyroptosis-related P-rognostic G-ene S-ignature and the A-ssociated R-egulation axis in L-ung A-denocarcinoma. *Cell Death Discov.* 7, 161. doi:10.1038/s41420-021-00557-2

Liu, C., Chen, B., Huang, Z., Hu, C., Jiang, L., and Zhao, C. (2020). Comprehensive Analysis of a 14 Immune-Related Gene Pair Signature to Predict the Prognosis and Immune Features of Gastric Cancer. *Int. Immunopharmacology* 89, 107074. doi:10.1016/j.intimp.2020.107074

Liu, S. J., Dang, H. X., Lim, D. A., Feng, F. Y., and Maher, C. A. (2021a). Long Noncoding RNAs in Cancer Metastasis. *Nat. Rev. Cancer* 21, 446–460. doi:10.1038/s41568-021-00353-1

Liu, Z., Lu, T., Li, J., Wang, L., Xu, K., Dang, Q., et al. (2021b). Clinical Significance and Inflammatory Landscape of aNovel Recurrence-Associated Immune Signature in Stage II/III Colorectal Cancer. *Front. Immunol.* 12, 702594. doi:10.3389/fimmu.2021.702594

Liu, Z., Lu, T., Wang, Y., Jiao, D., Li, Z., Wang, L., et al. (2021c). Establishment and Experimental Validation of an Immune miRNA Signature for Assessing Prognosis and Immune Landscape of Patients with Colorectal Cancer. *J. Cel Mol Med* 25, 6874–6886. doi:10.1111/jcmm.16696

Liu, Z., Wang, L., Liu, L., Lu, T., Jiao, D., Sun, Y., et al. (2021d). The Identification and Validation of Two Heterogenous Subtypes and a Risk Signature Based on Ferroptosis in Hepatocellular Carcinoma. *Front. Oncol.* 11, 619242. doi:10.3389/fonc.2021.619242

Liu, Z., Weng, S., Xu, H., Wang, L., Liu, L., Zhang, Y., et al. (2021e). Computational Recognition and Clinical Verification of TGF-β-Derived miRNA Signature with Potential Implications in Prognosis and Immunotherapy of Intrahepatic Cholangiocarcinoma. *Front. Oncol.* 11, 757919. doi:10.3389/fonc.2021.757919

Liu, Z., Zhang, Y., Dang, Q., Wu, K., Jiao, D., Li, Z., et al. (2021f). Genomic Alteration Characterization in Colorectal Cancer Identifies a Prognostic and Metastasis Biomarker: FAM83A|Ido1. *Front. Oncol.* 11, 632430. doi:10.3389/fonc.2021.632430

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A Significance Test for the Lasso. *Ann. Statist.* 42, 413–468. doi:10.1214/13-AOS1175

Loveless, R., Bloomquist, R., and Teng, Y. (2021). Pyroptosis at the Forefront of Anticancer Immunity. *J. Exp. Clin. Cancer Res.* 40, 264. doi:10.1186/s13046-021-02065-8

Ma, W., Li, W., Xu, L., Liu, L., Xia, Y., Yang, L., et al. (2021). Identification of a Gene Prognostic Model of Gastric Cancer Based on Analysis of Tumor Mutation Burden. *Pathol. Oncol. Res.* 27, 1609852. doi:10.3389/pore.2021.1609852

Ma, Y., Chen, Y., Lin, C., and Hu, G. (2018). Biological Functions and Clinical Significance of the Newly Identified Long Non-coding RNA RP1-85F18.6 in C-olorectal C-ancer. *Oncol. Rep.* 40, 2648–2658. doi:10.3892/or.2018.6694

Peng, F., Wang, R., Zhang, Y., Zhao, Z., Zhou, W., Chang, Z., et al. (2017). Differential Expression Analysis at the Individual Level Reveals a lncRNA Prognostic Signature for Lung Adenocarcinoma. *Mol. Cancer* 16, 98. doi:10.1186/s12943-017-0666-z

Qi, L., Chen, L., Li, Y., Qin, Y., Pan, R., Zhao, W., et al. (2016). Critical Limitations of Prognostic Signatures Based on Risk Scores Summarized from Gene Expression Levels: a Case Study for Resected Stage I Non-small-cell Lung Cancer. *Brief Bioinform* 17, 233–242. doi:10.1093/bib/bbv064

Rao, X., Jiang, J., Liang, Z., Zhang, J., Zhuang, Z., Qiu, H., et al. (2021). Down-Regulated CLDN10 Predicts Favorable Prognosis and Correlates with Immune Infiltration in Gastric Cancer. *Front. Genet.* 12, 747581. doi:10.3389/fgene.2021.747581

Roh, W., Chen, P.-L., Reuben, A., Spencer, C. N., Prieto, P. A., Miller, J. P., et al. (2017). Integrated Molecular Analysis of Tumor Biopsies on Sequential CTLA-4 and PD-1 Blockade Reveals Markers of Response and Resistance. *Sci. Transl. Med.* 9, eaah3560. doi:10.1126/scitranslmed.aah3560

Ruan, J., Wang, S., and Wang, J. (2020). Mechanism and Regulation of Pyroptosis-Mediated in Cancer Cell Death. *Chemico-Biological Interactions* 323, 109052. doi:10.1016/j.cbi.2020.109052

Rugge, M., Genta, R. M., Di Mario, F., El-Omar, E. M., El-Serag, H. B., Fassan, M., et al. (2017). Gastric Cancer as Preventable Disease. *Clin. Gastroenterol. Hepatol.* 15, 1833–1843. doi:10.1016/j.cgh.2017.05.023

Shao, W., Yang, Z., Fu, Y., Zheng, L., Liu, F., Chai, L., et al. (2021). The Pyroptosis-Related Signature Predicts Prognosis and Indicates Immune Microenvironment Infiltration in Gastric Cancer. *Front. Cel Dev. Biol.* 9, 676485. doi:10.3389/fcell.2021.676485

Smyth, E. C., Nilsson, M., Grabsch, H. I., van Grieken, N. C., and Lordick, F. (2020). Gastric Cancer. *The Lancet* 396, 635–648. doi:10.1016/S0140-6736(20)31288-5

Song, W., Ren, J., Xiang, R., Kong, C., and Fu, T. (2021). Identification of Pyroptosis-Related Subtypes, the Development of a Prognosis Model, and Characterization of Tumor Microenvironment Infiltration in Colorectal Cancer. *Oncoimmunology* 10, 1987636. doi:10.1080/2162402X.2021.1987636

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Tan, G., Huang, C., Chen, J., and Zhi, F. (2020). HMGB1 Released from GSDME-Mediated Pyroptotic Epithelial Cells Participates in the Tumorigenesis of Colitis-Associated Colorectal Cancer through the ERK1/2 Pathway. *J. Hematol. Oncol.* 13, 149. doi:10.1186/s13045-020-00985-0

Wang, H., Ding, Y., Chen, Y., Jiang, J., Chen, Y., Lu, J., et al. (2021). A Novel Genomic Classification System of Gastric Cancer via Integrating Multidimensional Genomic Characteristics. *Gastric Cancer* 24, 1227–1241. doi:10.1007/s10120-021-01201-9

Ye, Y., Dai, Q., and Qi, H. (2021). A Novel Defined Pyroptosis-Related Gene Signature for Predicting the Prognosis of Ovarian Cancer. *Cel Death Discov.* 7, 71. doi:10.1038/s41420-021-00451-x

Yuan, L., Xu, Z.-Y., Ruan, S.-M., Mo, S., Qin, J.-J., and Cheng, X.-D. (2020). Long Non-coding RNAs towards Precision Medicine in Gastric Cancer: Early Diagnosis, Treatment, and Drug Resistance. *Mol. Cancer* 19, 96. doi:10.1186/s12943-020-01219-0

Zhang, Y., Liu, Z., Li, X., Liu, L., Wang, L., Han, X., et al. (2021). Comprehensive Molecular Analyses of a Six-Gene Signature for Predicting Late Recurrence of Hepatocellular Carcinoma. *Front. Oncol.* 11, 732447. doi:10.3389/fonc.2021.732447

Zhao, E., Zhou, C., and Chen, S. (2021). A Signature of 14 Immune-Related Gene Pairs Predicts Overall Survival in Gastric Cancer. *Clin. Transl Oncol.* 23, 265–274. doi:10.1007/s12094-020-02414-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Guo, Liu, Yu, Liu, Ma, Ge, Xing, Lu, Weng, Wang, Liu, Hua, Han and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 28 February 2022
doi: 10.3389/fcell.2022.730132

A Novel Signature of Lipid Metabolism-Related Gene Predicts Prognosis and Response to Immunotherapy in Lung Adenocarcinoma

Kai Zhang[1], *Ying Qian*[1], *Xiaowei Quan*[2], *Tengteng Zhu*[2]* *and Biyun Qian*[2,3]*

[1]*Shanghai Tongren Hospital and Faculty of Public Health, Hongqiao International Institute of Medicine, Shanghai Jiao Tong University School of Medicine, Shanghai, China,* [2]*Hongqiao International Institute of Medicine, Shanghai Tongren Hospital and Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China,* [3]*Shanghai Hospital Development Center, Shanghai, China*

Background: Lipid metabolism disorder, a new hallmark of cancer initiation, has been involved in lung adenocarcinoma (LUAD). However, few biomarkers about lipid metabolism-related genes (LMRGs) have been developed for prognosis prediction and clinical treatment of LUAD patients.

Methods: In this study, we constructed and validated an effective prognostic prediction model for LUAD patients depending on LMRGs. Subsequently, we investigated the prediction model from immune microenvironment, genomic changes, and immunotherapy.

Results: Then, eleven LMRGs were identified and applied to LUAD subtyping. In comparison with the high-risk group, the low-risk group exhibited a remarkably favorable prognosis, along with a higher immune score and lower tumor purity. Moreover, the low-risk group presented higher levels of immune checkpoint molecules, lower tumor immune dysfunction and exclusion (TIDE) score and tumor mutation burden (TMB), and higher likelihood of benefiting from immunotherapy. Furthermore, the genomic changes of six LMRGs (CD79A, HACD1, CYP17A1, SLCO1B3, ANGPTL4, and LDHA) were responsible for the difference in susceptibility to LUAD by greatly influencing B-cell activation.

Conclusion: Generally speaking, the LMRG model is a reliable independent biomarker for predicting adverse outcomes in LUAD patients and has the potential to facilitate risk-stratified immunotherapy.

Keywords: lipid metabolism, lung adenocarcinoma, tumor microenvironment, immunotherapy, prognosis model, immune checkpoint

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Magdalena Julita Orzechowska,
Medical University of Lodz, Poland
Yi-Qing Qu,
Shandong University, China

***Correspondence:**
Tengteng Zhu
zhutengteng@gmail.com
Biyun Qian
qianbiyun@sjtu.edu.cn

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *24 June 2021*
Accepted: *07 February 2022*
Published: *28 February 2022*

Citation:
Zhang K, Qian Y, Quan X, Zhu T and Qian B (2022) A Novel Signature of Lipid Metabolism-Related Gene Predicts Prognosis and Response to Immunotherapy in Lung Adenocarcinoma. Front. Cell Dev. Biol. 10:730132. doi: 10.3389/fcell.2022.730132

Abbreviations: LUAD, lung adenocarcinoma; NSCLC, non-small-cell lung cancer; SCLC, small cell lung cancer; LMRGs, lipid metabolism-related genes; TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; TIDE, tumor immune dysfunction and exclusion; TMB, tumor mutation burden; CNV, copy number variation; OS, overall survival; LASSO, least absolute shrinkage and selection operator; DEG, differentially expressed gene; PCA, principal component analysis; ssGSEA, single-sample gene set enrichment analysis; GSEA, gene set enrichment analysis; ROC, receiver-operating characteristic; AUC, area under the curve; ESTIMATE, Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data; ICI, immune checkpoint inhibitor; DCA, decision curve analysis; TIMER, Tumor IMmune Estimation Resource; MSigDB, Molecular Signatures Database; FDR, false discovery rate; TIIC, tumor-infiltrating immune cell.

INTRODUCTION

Lung cancer, composed of approximately 85% non-small-cell lung cancer (NSCLC) and 15% small cell lung cancer (SCLC), is one of the most prevalent malignant cancers worldwide, with over 1.4 million deaths each year (Zarogoulidis et al., 2013). To date, lung adenocarcinoma (LUAD) has exceeded lung squamous cell carcinoma (LUSC) in the morbidity and become the most common pathological type of lung cancer (Bray et al., 2018). As we all know, diagnosed at an advanced stage remains a major cause of the high mortality of LUAD. Traditional prognostic prediction still relies on histopathologic diagnosis and tumor stage. However, these models fail to identify high-risk population and predict LUAD patients who are more likely to benefit from immunotherapies. Hence, it is imperative to explore accurate and effective prognostic biomarkers and models to assist clinical individualized treatment.

Metabolic reprogramming, a well-established hallmark of cancer, is critical for tumor initiation and progression (Hanahan and Weinberg, 2011). Lipids are composed of thousands of different molecules, including phospholipids, fatty acids, triglycerides, and sphingolipids. Fatty acids and cholesterol are the basic structure of cell membrane, which play a significant role in tumor cell proliferation, invasion, and metastasis (Orita et al., 2008). Lipids also can act as second messengers to transmit signals in tumor cells and participate in energy supply to tumor progression (Orita et al., 2007). In recent years, plenty of studies have focused on revealing the molecular mechanisms and signal pathways of LUAD caused by alteration of lipid composition. Hall et al. confirmed that MYC expression drives aberrant lipid metabolism in LUAD using the transgenic mouse model (Hall et al., 2016). Masri et al. found that the STAT3–SOCS3 pathway mediated the changes of insulin, glucose, and lipid metabolism in lung adenocarcinoma-bearing mice (Masri et al., 2016). These findings confirm the role of lipid metabolism in LUAD and indicate that lipid metabolism-related genes (LMRGs) may have a great prospect to act as prognostic markers of LUAD.

Therefore, we integrated gene expression and survival information of LUAD patients from TCGA and GEO databases to construct and validate a gene signature related to lipid metabolism to assess the risk of adverse outcomes. We then used this gene signature to divide LUAD patients into high- and low-risk groups and analyze the differences of signaling pathways and tumor microenvironment. On this basis, we identified the potential molecule affecting the response to immunotherapy in LUAD patients. Taken together, our study aimed to construct a robust prognosis prediction model based on LMRGs and provide an effective tool for immunotherapy of LUAD patients.

MATERIALS AND METHODS

Patient Data Collection and Procession

The level 3 RNA sequencing profile and matching clinical data of 500 LUAD tumor tissues and 59 adjacent non-tumorous tissues in The Cancer Genome Atlas (TCGA) were downloaded using the R package "TCGAbiolinks," whose expression profiles were normalized using the R package "Deseq2." RNA-seq expression matrix and clinical data of GSE31210 (226 lung adenocarcinoma samples; GPL570 platform) and GSE3141 (111 lung adenocarcinoma samples; GPL570 platform) were obtained from the Gene Expression Omnibus (GEO) database using the R package "GEOquery," whose expression profiles were normalized using the R package "limma" and served as the validation set. The data from TCGA and GEO are both publicly available and were used following the TCGA and GEO data access policies and publication guidelines. Eleven cases, with matched non-tumor and LUAD tumor tissue samples from the Shanghai Tongren Hospital of Shanghai Jiaotong University, were enrolled in this study. The CD79A mRNA expression profiling of GSE11969 (90 tumor samples and 5 non-tumor samples of lung adenocarcinoma; GPL7015 platform), GSE30219 (85 tumor samples and 14 non-tumor samples of lung adenocarcinoma; GPL570 platform), GSE31210 (226 tumor samples and 20 non-tumor samples of lung adenocarcinoma; GPL570 platform), GSE32683 (53 lung adenocarcinomas and 53 matched adjacent non-malignant lung samples; GPL570 platform), GSE75037 (83 lung adenocarcinomas and 83 matched adjacent non-malignant lung samples; GPL6884 platform), and GSE81089 (108 tumor samples and 19 non-tumor samples of lung adenocarcinoma; GPL570 platform) was downloaded from the GEO database and compared between tumor and non-tumor tissues of LUAD patients with "limma" R package. A total of 500 samples from the TCGA-LUAD cohort were divided into high- and low-risk groups by the risk prediction model. Then, the "limma" R package was utilized to analyze the mRNA expression profile between two groups to screen out DEGs with a threshold of FDR < 0.05.

Identification of LMRGs for Prognostic Prediction

Five lipid metabolism-related gene sets, including Hallmark fatty acid metabolism, KEGG glycerophospholipid metabolism, lipid raft, Reactome metabolism of lipids and lipoproteins, and Reactome phospholipid metabolism, were selected and extracted from Molecular Signatures Database v 7.2 (MSigDB) (Subramanian et al., 2005). After removing the overlapped genes, 664 lipid metabolism-related genes were acquired.

Differentially expressed genes between tumor samples and normal samples in TCGA-LUAD cohort were analyzed with the R package "DESeq2" based on the mRNA expression data of raw counts. A threshold of FDR less than 0.05 (FDR < 0.05) and the absolute $\log_2$ fold-change greater than 1 ($|\log_2 FC| > 1$) were set to define differentially expressed genes (DEGs). Finally, 217 LMRGs were obtained by intersecting DEGs and genes involved in five lipid metabolism-related gene sets.

Construction and Evaluation of the Prognostic Signature

First, univariate Cox regression analysis of overall survival (OS) was conducted to prefilter lipid metabolism-related genes (LRGs) with prognostic values. The least absolute shrinkage and selection

operator (LASSO) Cox regression analysis was used to narrow the range of variables with the R package "glmnet." Multivariate Cox stepwise regression analysis was applied to ultimately determine the target variable. Then, an LRG-related prognostic gene signature was constructed based on a linear combination of the regression coefficient derived from the multivariate Cox stepwise regression model coefficients (β) multiplied with its mRNA expression level. The detailed calculation formula was presented as follows: risk score = $\sum_{i=1}^{n} coef\,i\,x\,Genei$, where "coefi" and "Genei" denote the coefficient and mRNA expression levels of each lipid metabolism-related gene. All patients were stratified into high-risk and low-risk groups based on the median value of the calculated risk score. According to the expression profile of selected genes in the signature, PCA and t-SNE were conducted to explore the distribution of high- and low-groups with R packages "ggfortify" and "Rtsne," respectively. The Kaplan–Meier survival curve combined with a two-tailed log-rank test was used to compare the survival probability between the high- and low-risk groups by R packages "survival" and "survminer." The time-dependent receiver-operating characteristic (ROC) curve was applied to evaluate the predictive power of the gene signature for OS by the R package "survivalROC." The prognostic model was also validated in two independent cohorts from GSE31210 and GSE3141.

Establishment and Evaluation of the Nomograms for LUAD Survival Prediction

Univariate and multivariate Cox regression analyses were utilized to identify whether the risk score and other clinicopathological variables (including age, gender, pathological stage, tumor stage, and risk score) could be independent of this prognostic model for LUAD patients. Subsequently, we employed all independent clinical prognostic factors selected from previous analysis to construct a nomogram which can assess the OS probability of 1, 2, and 3 years in LUAD patients. The 1-year, 2-year, and 3-year calibration curves were drawn to compare the observed prediction probability with the actual OS probability to verify the accuracy of the nomogram. Overlapping the reference line showed the great accuracy of the model. Decision curve analysis (DCA) curves were drawn to visually evaluate the net benefit for different independent clinical prognostic factors of the model.

Function Enrichment Analysis

Risk score-related DEGs ($|\log_2 FC| \geq 1$, FDR < 0.05) were identified between high- and low-risk groups. Then, Gene Ontology (GO) and gene set enrichment analysis (GSEA) were conducted to explore enriched terms predicted to have an association with the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway in all hallmarks with R packages "clusterProfiler" and "GSEABase," respectively.

Immunohistochemical (IHC) Analysis

Immunohistochemically stained images were downloaded from the Human Protein Atlas (HPA) (https://www.proteinatlas.org/) and compared between tumor and non-tumor tissues of LUAD patients.

Mutant Gene Analysis

TCGA-LUAD mutation data with the non-synonymous mutations and non-sense mutations included were downloaded with "TCGAbiolinks" package and analyzed and visualized using "maftools" package. The corresponding TMB value of each sample was obtained by calculating the mutations per million bases.

Tumor Immune Dysfunction and Exclusion (TIDE) Analysis

TIDE is an effective algorithm that integrates two mechanisms of tumor immune escape—T-cell dysfunction and T-cell exclusion, which could be useful to predict immune checkpoint inhibitor (ICI) response in cancer treatment. The TIDE online web (http://tide.dfci.harvard.edu) was used to calculate the TIDE score, MSI score, T-cell exclusion score, and T-cell dysfunction score with the normalized RNA-seq data.

Copy Number Variation (CNV) Analysis

After TCGA-LUAD CNV data were downloaded with "TCGAbiolinks" package and combined with the mRNA expression data, the local Perl script was used to calculate the amplification and deletion frequency of each sample. The location of 11 hub genes was identified and visualized with "RCircos" R package.

Evaluation of Tumor-Infiltrating Immune Cells (TIICs) and Immune Checkpoint Inhibitors (ICIs) Using ssGSEA, CIBERSORT, and ESTIMATE Algorithms and TIMER and TISIDB Databases

Single-sample gene set enrichment (ssGSEA) analysis was conducted to calculated the TIIC abundance profiles in LUAD tumor samples with "GSVA" package. The CIBERSORT algorithm was used to quantify the abundance of 22 immune cell type signature (LM22) subtypes, including seven T-cell types, two NK-cell types, three macrophage cell types, two B-cell types, two dendritic cell types, two mast cell types, plasma cells, monocytes, eosinophils, and neutrophils, in both high- and low-risk groups. The ESTIMATE algorithm was utilized to estimate the ratio of the immune–stromal components in the TME for each sample calculated with the immune score, stromal score, and ESTIMATE score with "estimate" R package. The "Gene" module of TIMER (https://cistrome.shinyapps.io/timer/) and TISIDB (http://cis.hku.hk/TISIDB/index.php) was used to explore the association of abundance of tumor-infiltrating lymphocytes (TILs) with gene expression. In addition, the mRNA level of 25 critical immune checkpoints obtained from the literature between high- and low-risk groups was compared in TCGA-LUAD cohort.

Statistical Analysis

All data were presented as mean ± SD and analyzed with R studio software (Version 1.2.5001). An unpaired two-tailed Student's t-test between two groups was performed. The

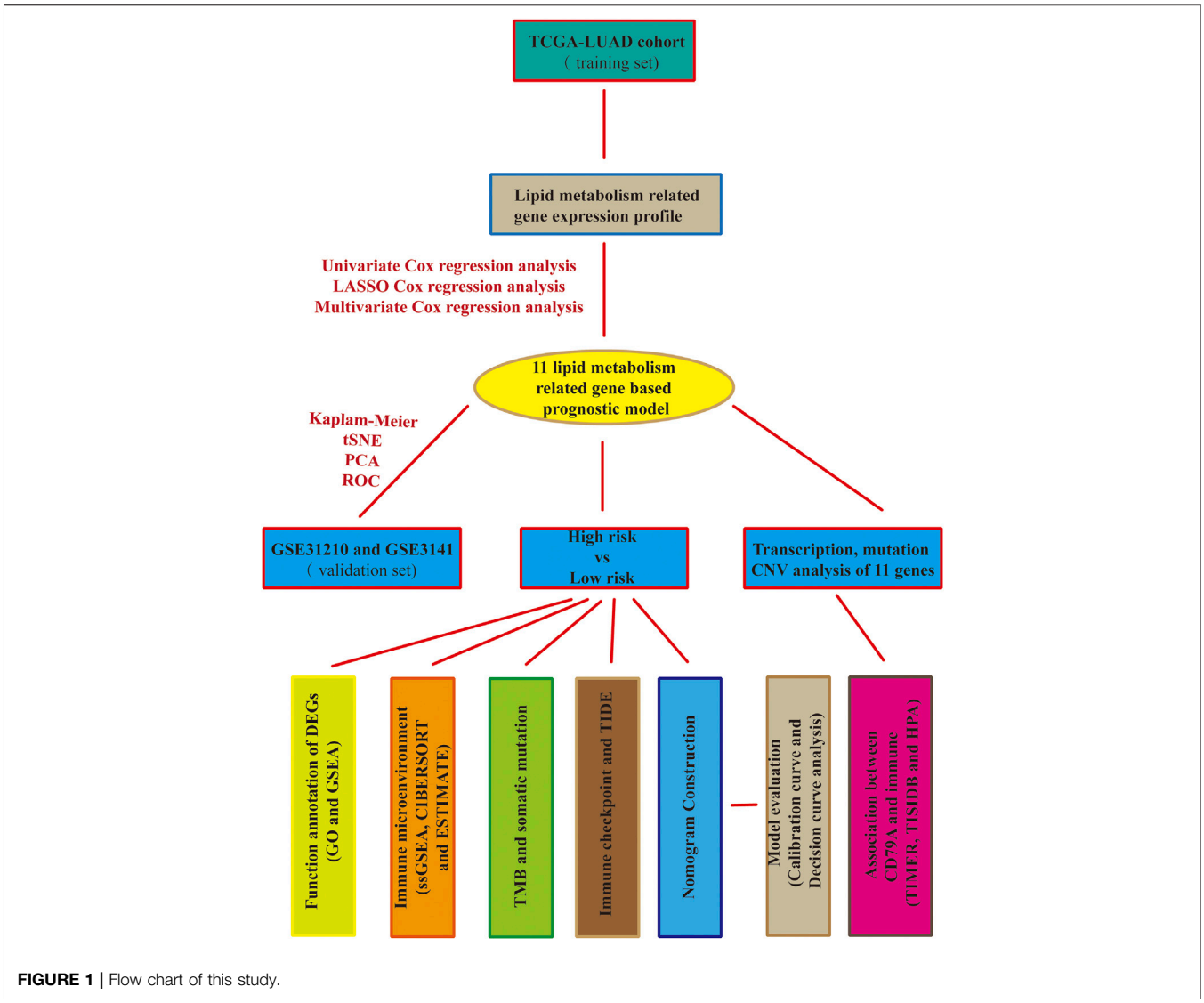


FIGURE 1 | Flow chart of this study.

Kaplan–Meier plot and log-rank test were applied to show the survival difference between high- and low-risk groups with R package "Survival." Harrell's concordance index (C-index) was computed for the predictive accuracy of the prognostic model. A p value <0.05 was considered statistically significant.

RESULTS

Identification of Prognostic LMRGs in the TCGA-LUAD Cohort

The entire work of this study was conducted as described in the flow chart (**Figure 1**). A total of 217 (32.7%) LMRGs (**Figure 2A**) were differentially expressed between tumor tissues (n = 500) and adjacent tissues (n = 59), 61 of which were remarkably associated with overall survival (OS) with the univariate Cox regression analysis (**Figure 2B**).

LASSO Cox regression analysis was employed to narrow the genes to 20 from 61 genes mentioned above (**Figure 2C**). Finally, an 11-LRG-based signature including ANGPTL4, CD79A, CIDEC, CYP17A1, ELOVL2, HACD1, LDHA, MTMR10, NPAS2, PLEKHA6, and SLCO1B3 was identified with multivariate stepwise regression analysis and subsequently used to construct a clinical prognostic model. The risk score of each LUAD patient was calculated with the coefficients (**Figure 2D**) achieved from the multivariate stepwise regression analysis. The detailed calculation formula was presented as follows: 0.086 × expression level of ANGPTL4 + (−0.128) × expression level of CD79A + 0.247 × expression level of CIDEC + (−0.584) × expression level of CYP17A1 + 0.398 × expression level of ELOVL2 + 0.132 × expression level of HACD1 + 0.221 × expression level of LDHA + (−0.412) × expression level of MTMR10 + 0.165 × expression level of NPAS2 + 0.112 × expression level of PLEKHA6 + 0.111 × expression level of SLCO1B3. Compared with normal tissues, three genes

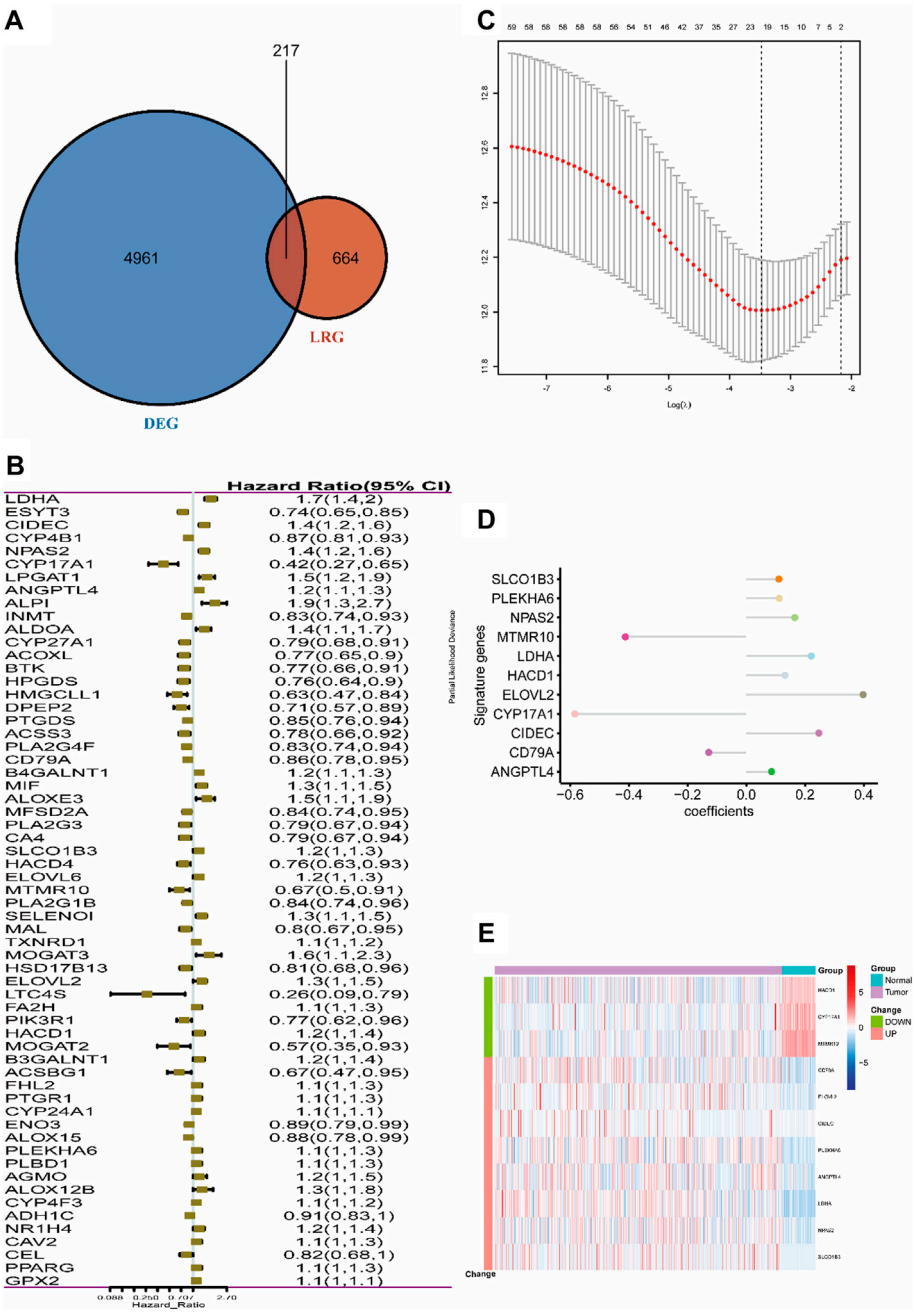


FIGURE 2 | Selection of LMRGs associated with the survival of lung adenocarcinoma. **(A)** Venn diagram shows the intersection of DEGs and LRGs. **(B)** Forest plot of LMRGs associated with TCGA-LUAD survival. **(C)** Selection of the optimal parameter (lambda) in the LASSO Cox model for TCGA-LUAD. **(D)** LASSO coefficients of the 11 hub genes in TCGA-LUAD. **(E)** Heatmap of top 11 hub genes in TCGA-LUAD. "Group" means the tumor or adjacent tissues of TCGA-LUAD samples, while "Change" represents the expression trend of 11 lipid metabolism-related genes in tumor tissues when compared with adjacent tissues. DEG, differentially expressed gene; LMRG, lipid metabolism-related gene. Construction of the 11-LRG-based prognostic model in the TCGA-LUAD cohort.

(CYP17A1, HACD1, and MTMR10) were down-regulated in tumor tissues, while the others were up-regulated (**Figure 2E**). Then, all TCGA-LUAD patients were stratified into a high-risk group (n = 250) and a low-risk group (n = 250) according to the median cut-off value (**Figure 3A**). Patients in the high-risk group had a higher probability of premature death compared with those in the low-risk group. t-SNE and PCA showed that the 11-gene signature expression of patients in the high- and low-risk groups was clearly divided into two clusters (**Figure 3B**). Consistently, dramatically worse OS was observed in the high-risk group than the low-risk group through the Kaplan–Meier plot ($p < 0.0001$) (**Figure 3C**). After that, the time-dependent ROC curve was

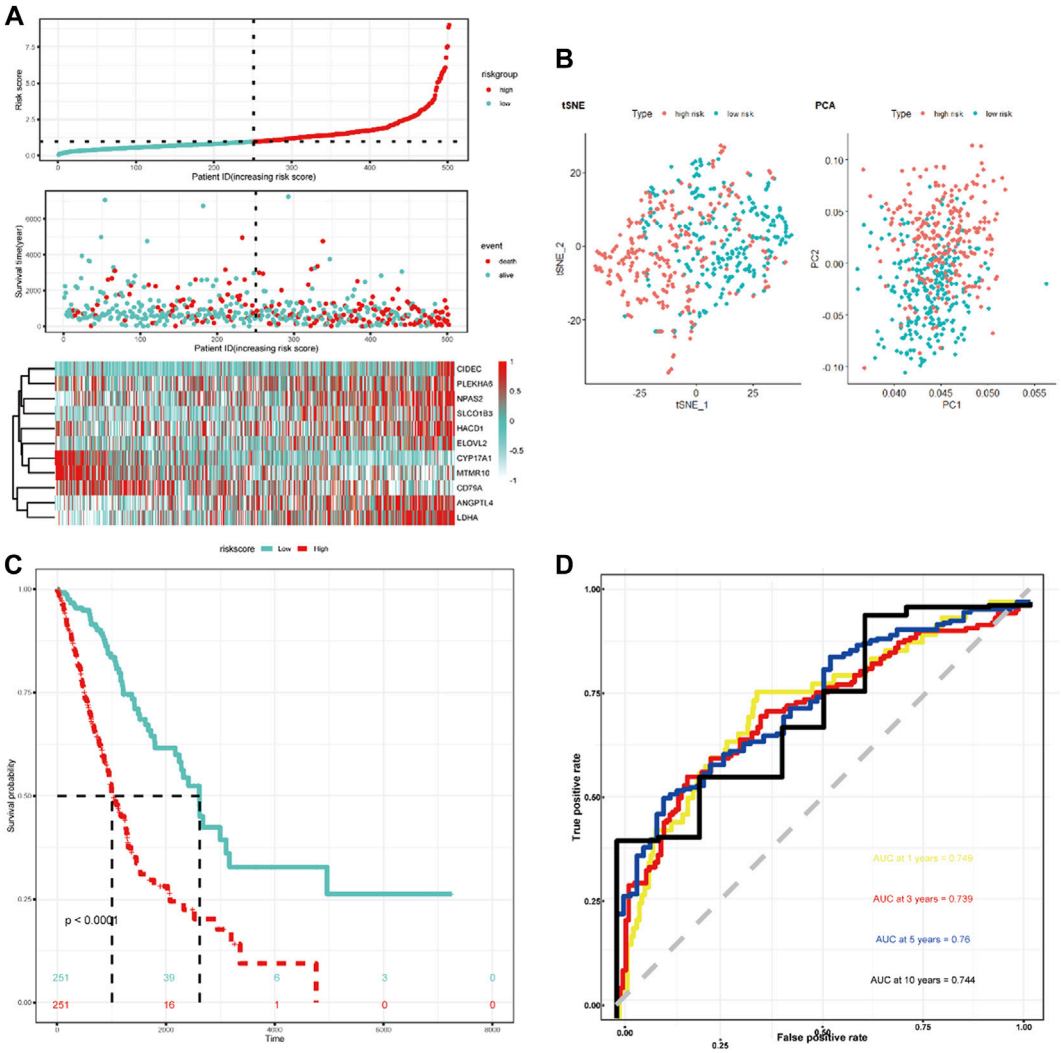


FIGURE 3 | Characteristics of the LRG-related prognostic gene signature. **(A)** Heatmap of the 11-LRG-related gene expression profiles and the distribution of risk score and survival time/status for TCGA-LUAD. **(B)** t-SNE and PCA plot of 11-LRG-related gene expression profiles. **(C)** The Kaplan–Meier curve was plotted to estimate the overall survival probabilities of the low- and high-risk groups. **(D)** The ROC curve was plotted to predict the 1-, 3-, 5-, and 10-year prognosis.

utilized to evaluate the prediction performance of the risk score. It was found the predicted area under the curve (AUC) of this model for 1-year, 3-year, 5-year, and 10-year overall survival rates reached 0.749, 0.739, 0.76, and 0.744, respectively (**Figure 3D**). The C-index for this clinical prognostic model reached 0.735.

The GSE31210 and GSE3141 external validation cohort was utilized to verify the robustness of the 11-LRG-based risk model. We calculated the risk score of two validation datasets with the above formula, based on the median using which patients were divided into high-risk and low-risk groups, respectively. The results showed patients with high risk score had worse OS in both GSE31210 and GSE3141, which was in line with that of TCGA-LUAD cohort (**Figures 4A,D**). In addition, t-SNE dimensionality reduction analysis found the mRNA profiles of 11 genes of the high- and low-risk groups were distinctly divided into two clusters in both GSE31210 and GSE3141 (**Figures 4B,E**). Furthermore, KM curves demonstrated patients in the low-risk group had higher survival probability than those in the high-risk group in both GSE31210 (p = 0.019) and GSE3141 (p = 0.039) (**Figures 4C,F**).

Differences in tumor immune microenvironment were explored to distinguish high- and low-risk LUAD patients. The ssGSEA results showed that the infiltrating levels of neutrophils, aDCs, iDCs, pDCs, macrophages, γδT, Tfh, Th1 cells, Tregs, NK cells, and B cells were notably elevated with the increased risk score (**Figure 5A**), indicating the activation of the body's immune response. In order to further clarify the immune cell infiltration status between the high- and low-risk groups, the infiltration of 22 kinds of tumor-infiltrating immune cells (TIICs) was analyzed with the CIBERSORT algorithm. Compared with TIICs of LUAD patients in the low-risk group, B-cell memory (p = 0.0018), T-cell CD8 (p = 0.0072), T-cell CD4 memory

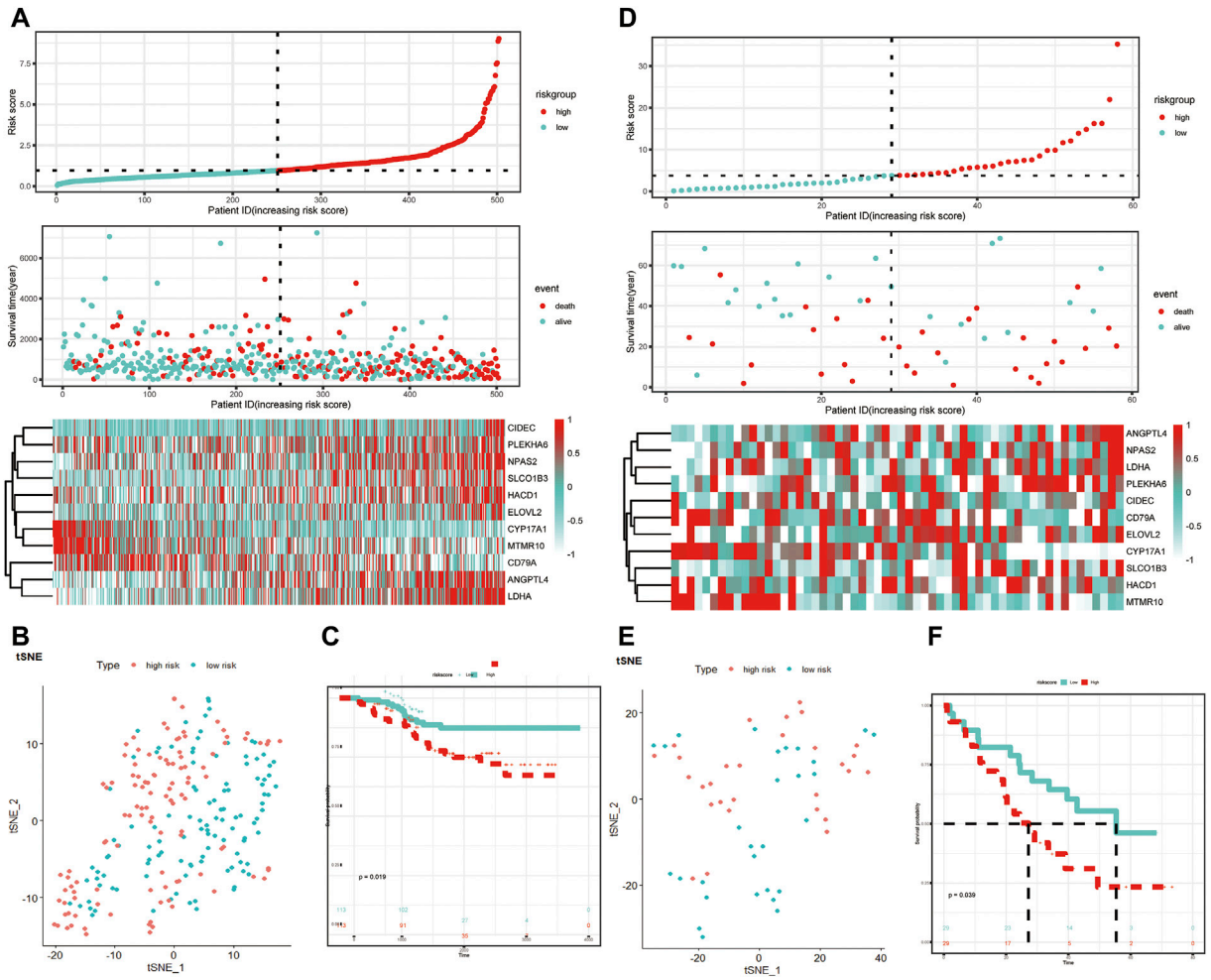


FIGURE 4 | Validation of the LRG-related prognostic model. **(A,D)** Heatmap of the 11-LRG-related gene expression profiles and the distribution of risk score and survival time/status for GSE31210 and GSE3141. **(B,E)** t-SNE and PCA plot of 11-LRG-related gene expression profiles for GSE31210 and GSE3141. **(C,F)** The Kaplan–Meier curve was plotted to estimate the overall survival probabilities of the low- and high-risk groups for GSE31210 and GSE3141. Diverse tumor immune microenvironment between high- and low-risk LUAD patients.

resting ($p < 0.0001$), monocytes ($p = 0.0015$), and mast cell resting ($p = 0.020$) remarkably decreased, while T-cell CD4 naïve ($p = 0.046$), NK cell activated ($p = 0.019$), macrophage M0 ($p < 0.0001$), and eosinophils ($p = 0.0093$) significantly increased in the high-risk group (**Figure 5B**). The stacked bar chart of the proportions of 22 TIICs in the patients from both groups is exhibited in **Supplementary Figure S1B**. Among all TIICs with significant differences in abundance, T-cell CD4 memory resting, macrophage M0, and T-cell CD8 accounted for the highest proportions, demonstrating a key immunological function of them in the development and progression of LUAD. In addition, the ESTIMATE results revealed that LUAD patients with low risk had a higher stromal score, immune score, and ESTIMATE score but lower tumor purity than those in the high-risk group (**Figures 5C–F**).

Antigens produced by tumor mutations were closely related to immunotherapy. We found that the tumor mutation burden (TMB) of LUAD patients in the high-risk group was significantly higher than that in the low-risk group (**Figure 6A**). In detail, the mutation frequency of the top 20 genes in the high-risk group was higher than that in the low-risk group (**Figure 6B**). Moreover, TP53, TTN, MUC16, CSMD3, and RYR2 ranked as the top five genes with the highest mutation frequency in both groups. Among them, P53 mutations are mainly missense mutation and non-sense mutation, while TTN, MUC16, CSMD3, and RYR2 mutations are mainly missense mutation and multi-hit.

The up-regulated genes were mainly enriched in the biological processes such as the ncRNA metabolic process, post-translational protein modification, and the ATP metabolic process (**Supplementary Figure S1A**), while the down-regulated genes were involved in the process of T-cell activation, regulation of T-cell activation, leukocyte proliferation, and regulation of lymphocyte-mediated immunity (**Supplementary Figure S1B**). Overall, all DEGs were related to several biological processes including the

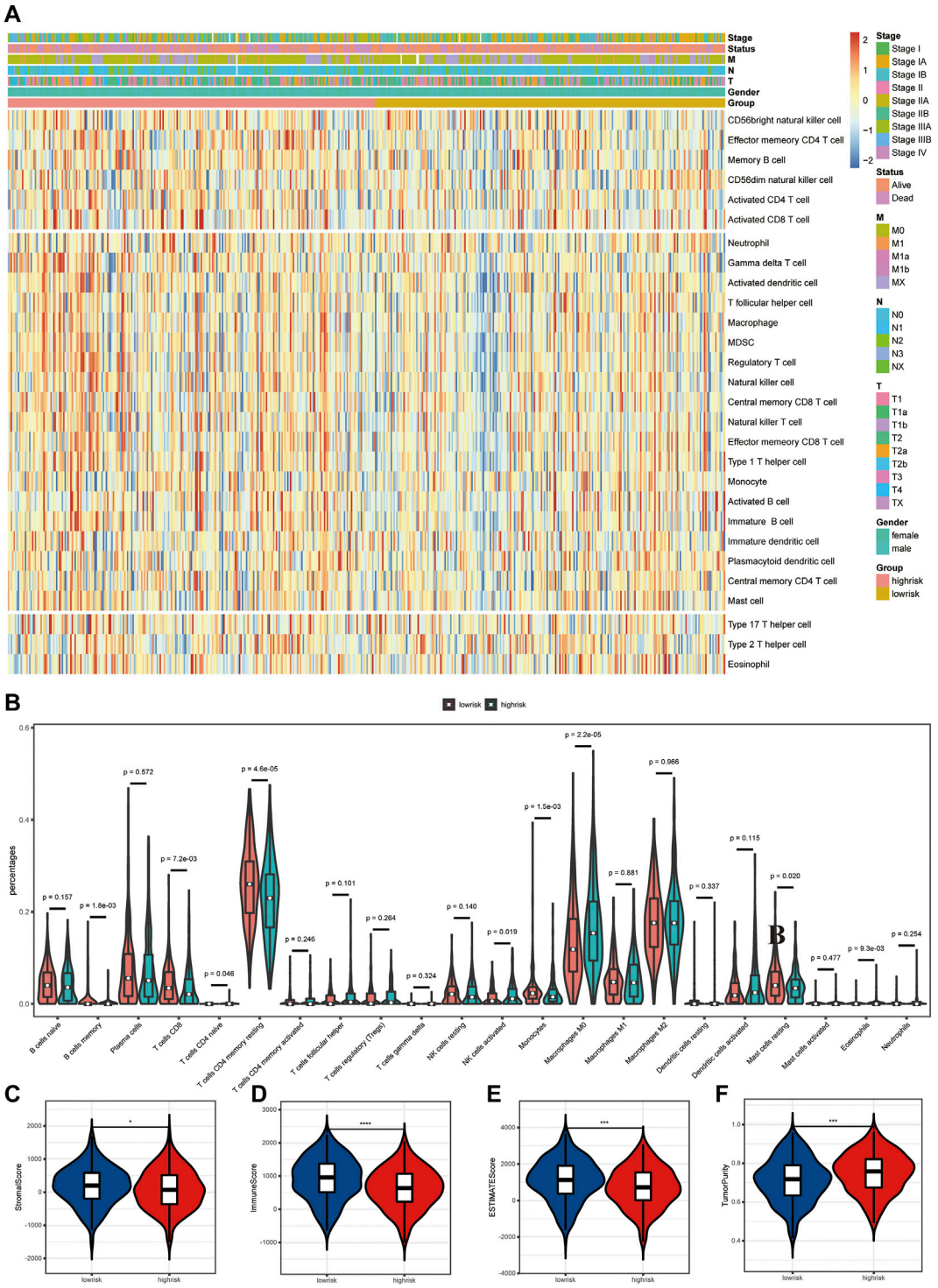


FIGURE 5 | Differences in tumor immune microenvironment (TME) between high- and low-risk LUAD patients. **(A)** Heatmap of 28 infiltrating immune cell levels between high and low risks with ssGSEA. **(B)** Comparisons of 22 important immune fractions between high and low risks with CIBERSORT. **(C–F)** Comparison of the ESTIMATE score, stromal score, immune score, and tumor purity between high and low risks with ESTIMATE.

T-cell receptor signaling pathway, antigen processing and presentation of exogenous antigen, leukocyte proliferation, and response to hypoxia (**Figure 7A**). Based on these DEGs, GSEA and several dominantly enriched pathways, including immune-related pathways such as TNFα signaling via NF-κB, and other pathways including angiogenesis,

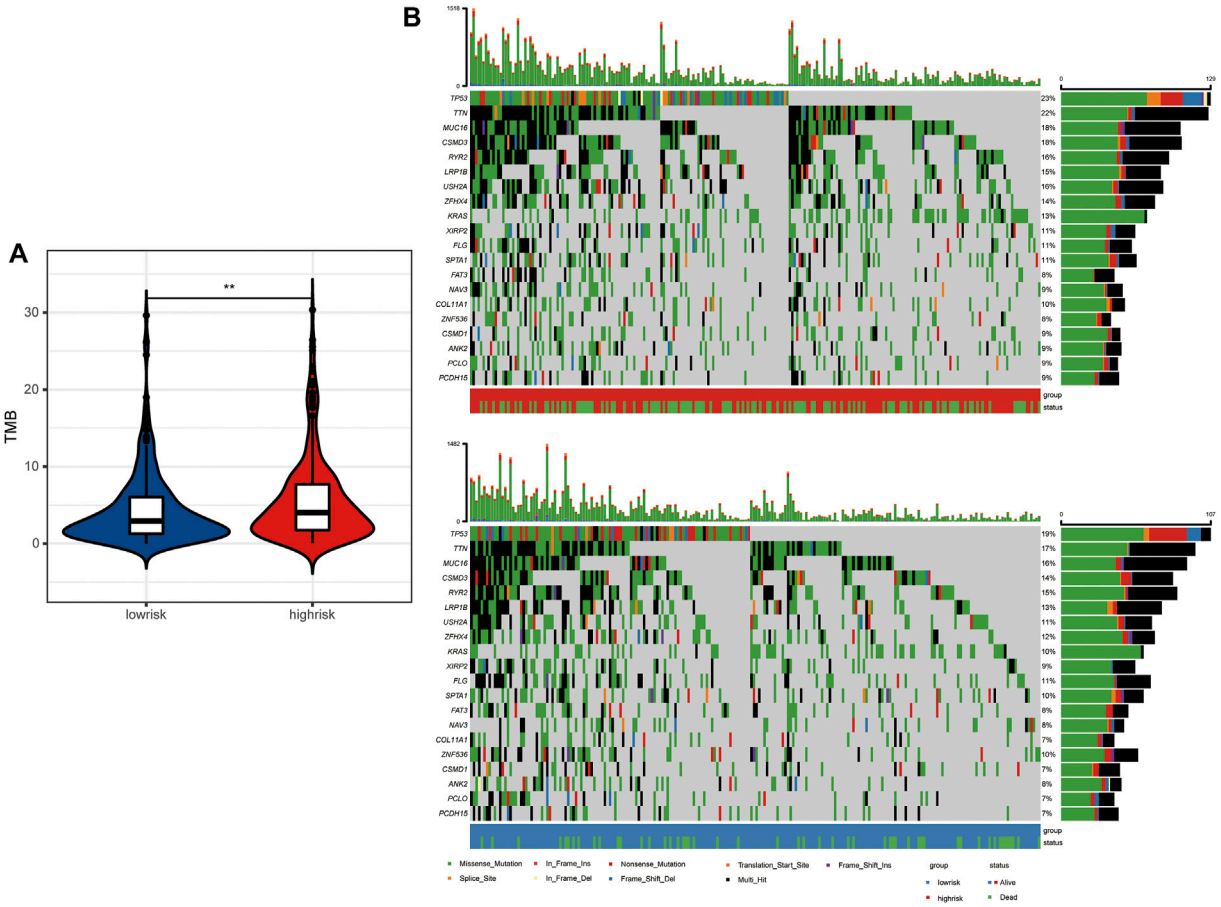


FIGURE 6 | Tumor mutation landscape between high- and low-risk LUAD patients. **(A)** Tumor mutation burden between high- and low-risk groups. **(B)** Tumor mutation landscape of high- and low-risk groups.

epithelial–mesenchymal transition, estrogen response early, estrogen response late, glycolysis, and hypoxia were also performed (**Supplementary Figure S1C**).

As we all know, an immune checkpoint inhibitor is a promising lung cancer treatment strategy that has emerged in recent years. Compared with the high-risk group, 25 emerging immune checkpoints including PD1, CTLA4, and LAG3 were highly expressed, while CD276 and TNFRSF14 showed low expression level in the low-risk group ($p < 0.05$) (**Figure 7B**). Then, we utilized the tumor immune dysfunction and exclusion (TIDE) algorithm to predict the efficacy of response to immunotherapy. Interestingly, we found that the TIDE score of the low-risk group was significantly lower than that of the high-risk group, indicating more likely effectiveness for immunotherapy in the low-risk group (**Figure 7C**). In addition, the remarkably higher MSI score and T cell exclusion score and lower T cell dysfunction score were observed in the high-risk group (**Figures 7D–F**). These data further suggest that patients in the low-risk subtype may be more likely to benefit from immunotherapy.

Nomogram Construction and Model Evaluation

The 11-LRG-based signature and main clinical factors were incorporated into univariate and multivariate Cox regression analyses to determine whether the signature could independently predict the survival of LUAD patients. First, univariate Cox analysis confirmed the T stage, tumor stage, and risk score of signature were associated with poor prognosis in patients with LUAD (**Figure 8A**). Then, multivariate Cox regression analysis further indicated that both the risk score of signature and tumor stage were significantly correlated with the OS of LUAD patients, which proved that the 11-LRG-based signature could independently predict the OS for the TCGA-LUAD patients (**Figure 8B**). Subsequently, a nomogram integrating the signature and tumor stage was constructed (**Figure 8C**). To evaluate the predictive performance of the nomogram, calibration curves and decision curve analysis (DCA) were conducted, which showed good performance in predicting the survival probability of LUAD patients. The results confirmed that calibration curves of 1-, 2-, 3-, and 5-year OS showed superior

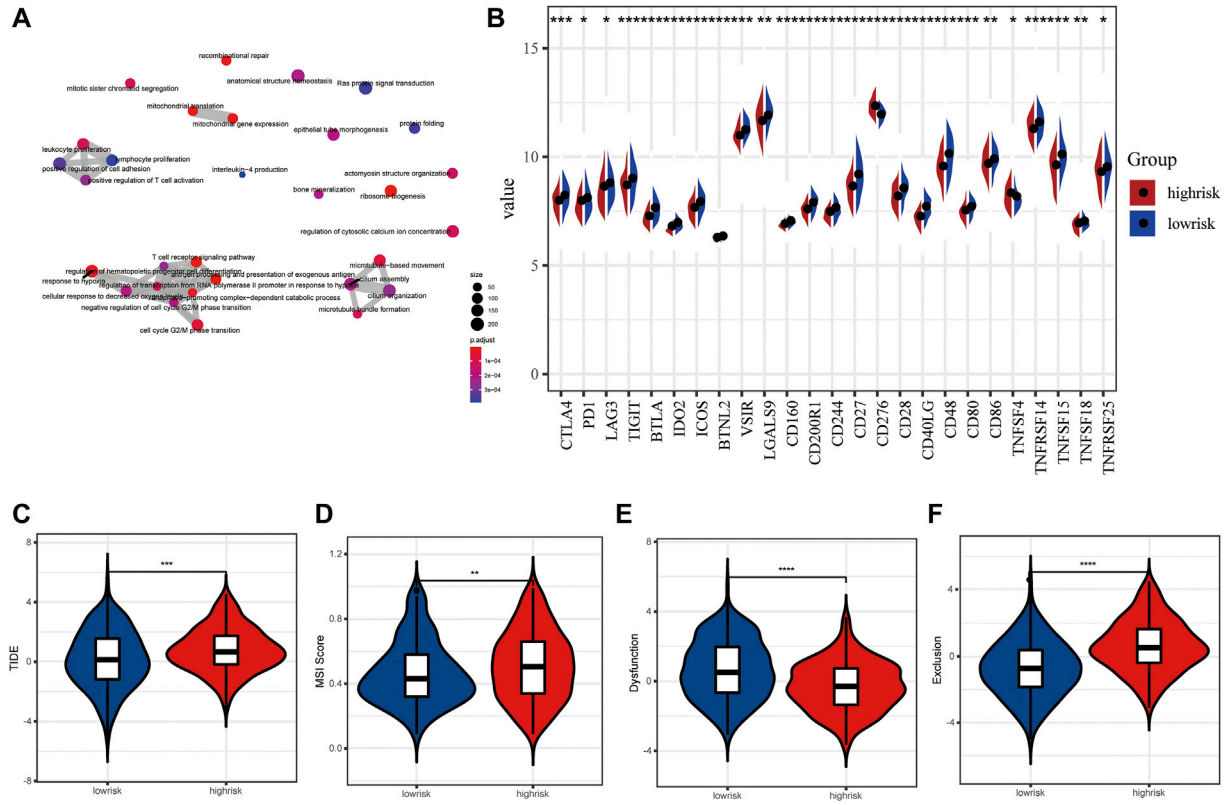


FIGURE 7 | Immune checkpoint inhibitors and immunotherapy between high- and low-risk LUAD patients. **(A)** Tumor mutation burden between high- and low-risk groups. **(B)** Tumor mutation landscape of high- and low-risk groups. **(C–F)** TIDE score, MSI score, dysfunction score, and exclusion score between high- and low-risk groups.

agreement between the predicted and actual survival LUAD cohorts (**Figure 8D**). Furthermore, DCA of the risk model showed the best net benefit for predicting 1-, 2-, 3-, 5-, and 10-year overall survival, especially for 5-year survival (**Figure 8E**). Thus, these results indicated that the nomogram was an optimal model in predicting the survival probability of LUAD patients.

We first summarized the transcriptional regulation, frequency of copy number variations (CNVs), and somatic mutations of 11 hub lipid metabolism-related genes in LUAD. The comprehensive landscape of 11 hub gene interactions, prognostic significance, and association with immune cells for LUAD patients was exhibited with the regulatory network (**Figure 9A**). We found that the six hubs including ANGPTL4, LDHA, CYP17A1, HACD1, SLCO1B3, and CD79A not only presented an extensive correlation in expression but also markedly correlated with B cells. RNA-seq data from 12 paired normal tissues and LUAD tissues collected from Tongren Hospital showed that SLCO1B3 (logFC = 4.23, p = 0.01), CD79A (logFC = 2.80, p = 0.00), LDHA (logFC = 1.41, p = 0.00), NPAS2 (logFC = 1.39, p = 0.01), PLEKHA6 (logFC = 1.23, p = 0.00), ANGPTL4 (logFC = 0.85, p = 0.04), and CIDEC (logFC = 1.82, p = 0.00) increased, while HACD1 (logFC = −1.67, p = 0.00), CYP17A1 (logFC = −1.91, p = 1.00), MTMR10 (logFC = −1.03, p = 0.08), and ELOVL2 (logFC = −0.16, p = 0.89) decreased in tumor (**Figure 9C**). Among 567 samples, 55 experienced mutations of 11 hub regulators, with frequency 9.7%. It was found that SLCO1B3 showed the highest mutation frequency followed by PLEKHA6 and CYP17A1, while ANGPTL4, CD79A, HACD1, and CIDEC did not show any mutations in LUAD samples (**Figure 9B**). The investigation of CNV alteration frequency showed several CNV alteration in 11 hub regulators, 5 of which including PLEKHA6, ELOVL2, NPAS2, SLCO1B3, and HACD1 focused on the amplification in copy number, while the rest including ANGPTL4, MTMR10, CIDEC, CYP17A1, LDHA, and CD79A had more popular frequency of CNV deletion (**Figure 9D**). In addition, the location of CNV alteration of 11 hub regulators on chromosomes was exhibited (**Figure 9E**).

In order to investigate the molecule most relevant to the immune microenvironment, we conducted the correlation analysis between 11 lipid metabolism-related genes and six kinds of immune cells using the TIMER database (**Figure 10**, **Supplementary Figure S4**). Among the 11 genes, we found that eight genes were closely related to the activation of B cells, four genes were related to the activation of $CD8^+$ T cells, seven genes were related to the activation of $CD4^+$ T cells, four genes were related to the activation of macrophages, seven genes were related to the activation of neutrophils, and four genes were related to the activation of dendritic cells ($p < 0.05$). Interestingly, CD79A showed a close relationship with B cells (r = 0.616, p = 5.73e-52), $CD4^+$ T cells (r = 0.4, p = 4.84e-20), $CD8^+$ T cells (r = 0.199, p = 9.43e-06), dendritic cells (r = 0.225, p = 5.41e-07), and

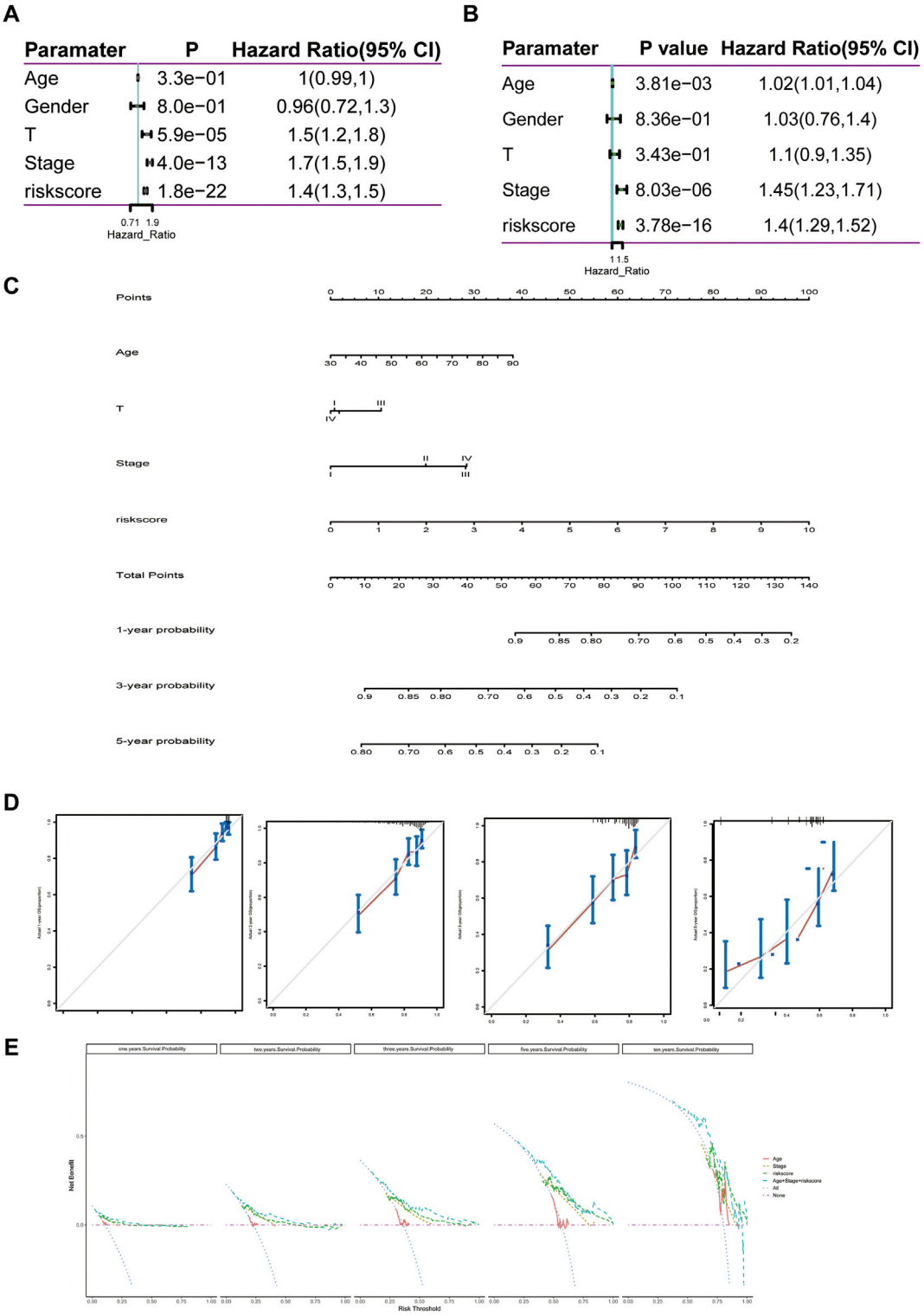


FIGURE 8 | Construction and evaluation of the nomogram predicting the OS for TCGA-LUAD patients. **(A,B)** Univariate and multivariate regression analyses of clinical parameters. **(C)** Nomogram for predicting 1-, 3-, and 5-year OS of LUAD patients based on four independent prognostic factors. **(D)** One-, 2-, 3-, and 5-year calibration curves of TCGA-LUAD. **(E)** One-, 2-, 3-, 5-, and 10-year decision curve analysis of TCGA-LUAD. Comprehensive analysis of transcription, mutation, and copy number variation among 11 lipid metabolism-related genes.

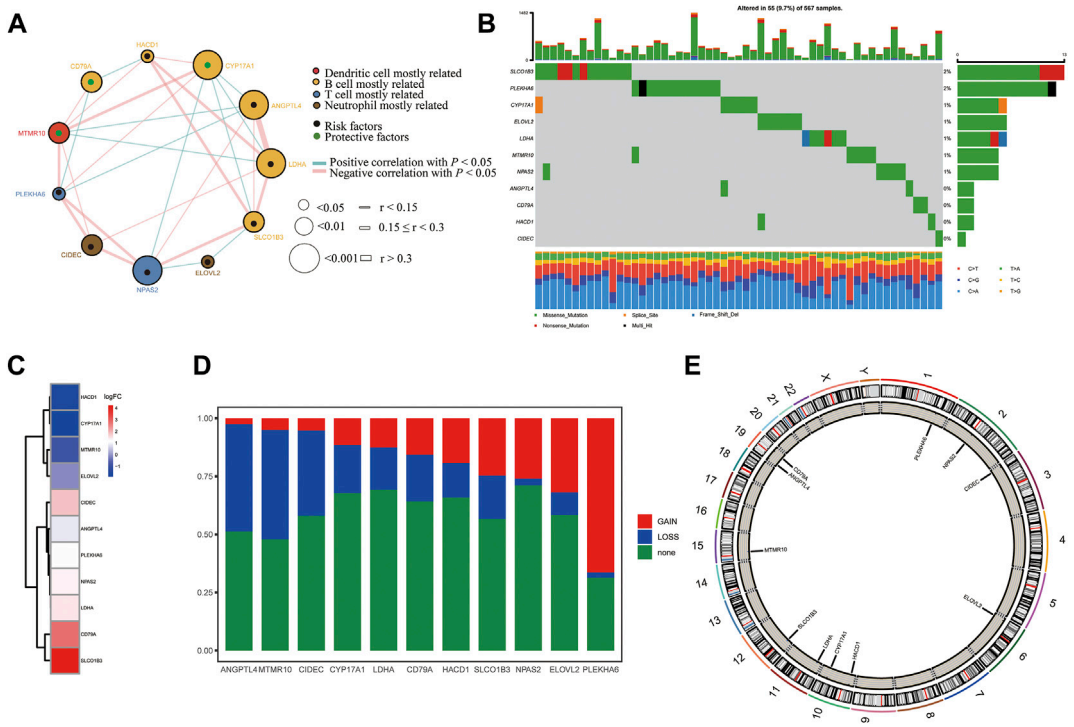


FIGURE 9 | Landscape of the transcriptional regulation, incidence of copy number variations (CNVs), and somatic mutations of 11 hub lipid metabolism-related genes in LUAD. **(A)** Interaction between 11 hub genes in LUAD. The circle size represents the *p* value of each gene on the prognosis, and the range of values calculated by the log-rank test was $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively. Green dots in the circle mean risk factors of prognosis; black dots in the circle mean protective factors of prognosis. The lines linking genes show their interactions, and thickness shows the correlation strength calculated by the correlation coefficient was $r < 0.15$, $0.15 \le r < 0.3$, and $r > 0.3$. Significant negative correlation is marked in blue, while significant positive correlation is marked in red. The gene associated with mostly related immune cells, such as dendritic cells, B cells, T cells, and neutrophils, based on TIMER is marked in red, yellow, blue, and brown, respectively. **(B)** Waterfall plot of tumor somatic mutation for 11 hub lipid metabolism-related genes. **(C)** Log (fold-change) of 11 hub lipid metabolism-related genes between 12 paired normal and LUAD tissues collected from Shanghai Tongren Hospital. **(D,E)** CNV alteration frequency and chromosome location of 11 hub lipid metabolism-related genes. CD79A with good prognosis was the molecule most relevant to the immune microenvironment.

neutrophils (r = 0.237, p = 1.36e-07), which may mediate the linkage between lipid metabolism and immunotherapy. Therefore, we further analyzed the expression, prognosis, and relationship with the risk score and immune microenvironment of CD79A in TCGA and TISIDB databases.

In the TCGA-LUAD cohort, the mRNA level of CD79A showed a negative correlation with risk score based on the 11-LMRG signature (r = −0.28, p = 3.8e-10), macrophage M0 (r = −0.21, p = 3.52e−06), and mast cell resting (r = -0.21, p = 2.32e−06), but a positive correlation with B-cell memory (r = 0.32, p = 5.34e−13), T-cell CD8 (r = 0.21, p = 1.44e−06), and plasma cells (r = 0.45, p = 2.52e−25) (**Supplementary Figure S2A**). Moreover, CD79A exhibited a strong correlation with several immune cells, immunoinhibitors, and immunostimulators (**Supplementary Figure S2B**), such as tumor-infiltrating lymphocytes including immature B cells (r = 0.697, p = 2.2−16) and activated B cells (r = 0.848, p = 2.2e−16), immunoinhibitors including BTLA (r = 0.659, p = 2.2−16) and TIGIT (r = 0.614, p = 2.2e−16), and immunostimulators including CD27 (r = 0.907, p = 2.2−16) and TNFRSF17 (r = 0.883, p = 2.2e−16) in the TISIDB database (**Supplementary Figure S2C**).

The mRNA expression level of CD79A was higher in LUAD tissues than in the non-tumor tissues in six datasets including GSE31210 (p = 3.4e-07), GSE75037 (p = 1e-05), GSE32683 (p = 0.00043), GSE11969 (p = 0.02), GSE30219 (p = 9.9e-06), and GSE81089 (p = 0.0093) (**Supplementary Figure S3A**), and it was negatively correlated with tumor stages (rho = −0.207, p = 2.59e-06) (**Supplementary Figure S3C**). In addition, the CD79A protein expression level increased in LUAD tissues than in the normal lung tissues in the HPA database (**Supplementary Figure S3B**). Survival analysis showed that the OS of the CD79A high expression group was significantly higher than that of the low expression group (p < 0.01) in the TISIDB database (**Supplementary Figure S3D**).

DISCUSSION

Lipid metabolism disorder is a hallmark of malignant tumor. Convincing evidence has suggested that abnormal metabolic activities could promote cancer development, due to its crucial role in angiogenesis, energy regulation, and cell proliferation (Liu et al., 2020a; Valença and Ferreira, 2020; Vazquez et al., 2020). Several lipid metabolism-based clinical prognostic models have been constructed for colorectal cancer, glioblastoma, and oral

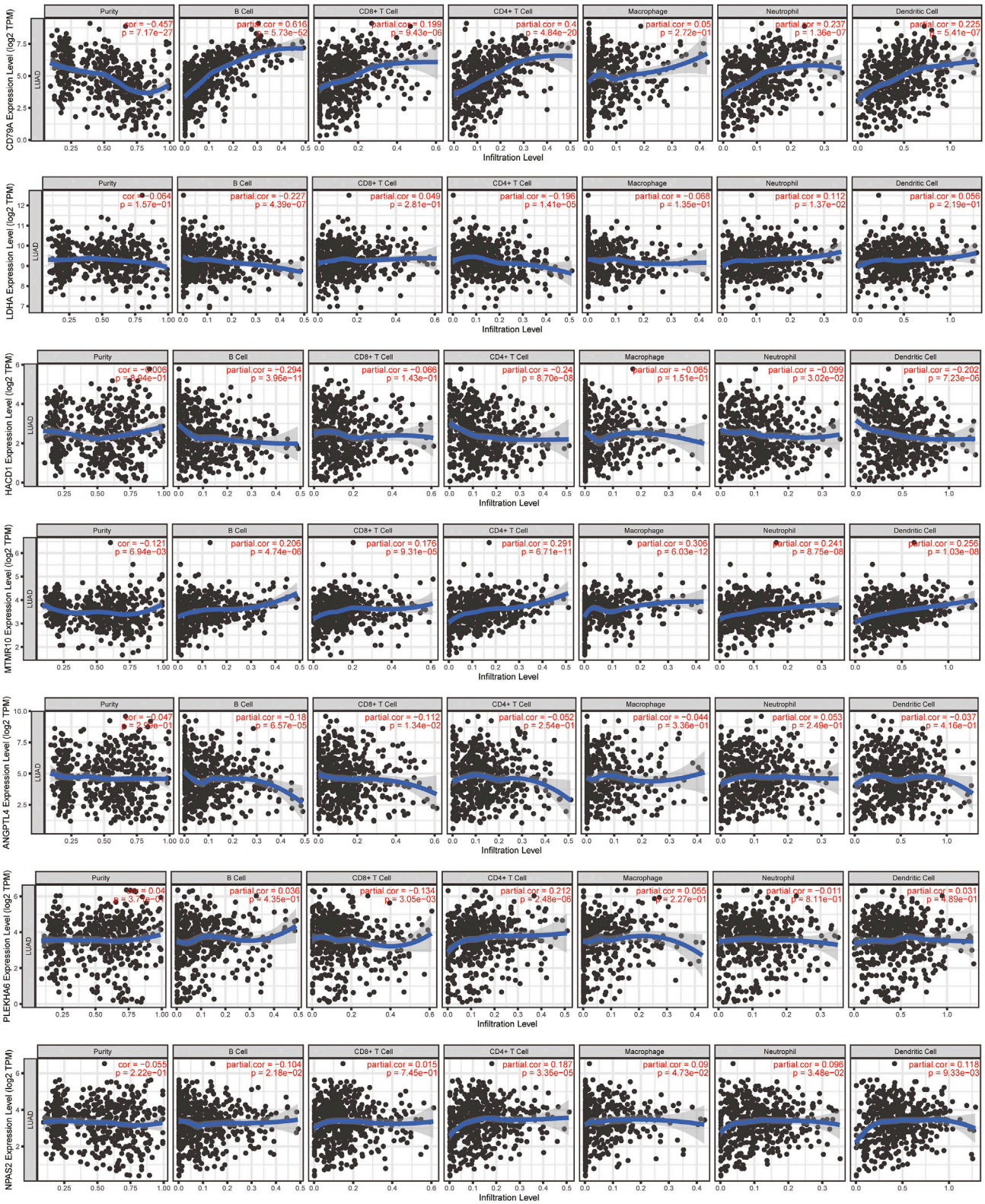


FIGURE 10 | Relationship between 11 lipid metabolism-related genes and 6 kinds of immune cells using the TIMER database. The correlation scatter plot showed the relationship between 11 genes and infiltration of B cells, $CD8^+$ T cells, $CD4^+$ T cells, macrophages, neutrophils, and dendritic cells in LUAD.

squamous cell carcinoma (Hu et al., 2019; Wu et al., 2019; Gharib et al., 2020). However, to our knowledge, the prognostic model of LMRG signature for LUAD patients has not been reported. In this case, our research focused on five lipid metabolism pathways strongly related to cancers, including fatty acid metabolism, glycerophospholipid metabolism, lipid raft, metabolism of lipids and lipoproteins, and phospholipid metabolism.

In this research, we attempted to construct an efficient prognostic model for LUAD patients based on 11 LMRGs screened by univariate Cox regression analysis, LASSO Cox regression analysis, and multivariate Cox stepwise regression analysis and calculated the risk score. Then, a nomogram was constructed to better predict accurate clinical outcomes. The nomogram integrated multiple prediction indicators based on multivariate regression analysis and then used line segments with scale to show the relationship between variables in the prediction model. A nomogram based on the autophagy gene signature had a concordance index of 0.71 to predict the survival possibility of NSCLC patients (Liu et al., 2019). Our nomogram generated a graphical statistical prediction model that assigns scores to each level of all independent factors, including age, T stage, and clinical stage and risk score. Calibration curves exhibited that the nomogram could accurately predict 1-, 2-, 3-, and 5-year survival probabilities. DCA curves showed that the nomogram had good benefit in predicting 1-year, 2-year, 3-year, 5-year, and 10-year survival, especially 5-year survival, meaning a potential translational value of the nomogram in LUAD patients.

Next, we conducted the lipid metabolism phenotype and differentially expressed gene analysis between high-risk and low-risk groups. Our results found up-regulated genes exhibited enrichment of post-translational protein modification, ATP metabolic progress, and ncRNA metabolic progress, whereas down-regulated genes mainly displayed enrichment of some immune progress, including regulation of T-cell activation, T-cell activation, and regulation of lymphocyte-mediated immunity, which hinted that different lipid metabolism status might be associated with immune microenvironment. A study found that cholesterol in the tumor environment can increase the CD36 expression of $CD8^+$ T cells and then ingest too many fatty acids, leading to lipid oxidative damage and iron death, further resulting in the loss of its lethal function and promoting the growth of tumors (Bray et al., 2018). At present, it is believed that the relationship between lipids and immunity is that cancer cells use lipids to support their aggressive behavior and allow immune escape, while metabolic reprogramming of cancer cells destroys the balance between lipid synthesis and catabolism, resulting in lipid accumulation in the tumor microenvironment (Hanahan and Weinberg, 2011). Therefore, lipid metabolism and immunity are closely related in cancer, which is worthy of in-depth exploration.

Immune checkpoints have become a promising strategy for the treatment of many cancers. Whether the immune microenvironment and response to immunotherapy between high- and low-risk groups differed was further explored. CIBERSORT is a powerful tool to deconvolute the expression matrix of 22 human immune cell subtypes based on the principle of linear support vector regression, which has been widely used in plenty of oncology research studies (Huang et al., 2019; Liu et al., 2020b; Deng et al., 2020). Immune cell infiltration analysis found a distinctly different immune microenvironment especially the abundance of T cells and B cells between high- and low-risk groups. Interestingly, a large number of clinical cohort studies also showed that there was a positive correlation between B-cell infiltration and response to immunotherapy in a variety of different tumor types, emphasizing the important role of B cells in anti-tumor immunity (Helmink et al., 2020; Petitprez et al., 2020). T-cell failure has been identified as an important mechanism for cancer cells to escape host immunity (Marijt et al., 2019). Our results showed that, of all tumor-infiltrating immune cells (TIICs), the abundance of B-cell memory, T-cell CD8 memory, and T-cell CD4 memory resting significantly decreased in the high-risk group, which was consistent with the GO enrichment result that the immune process was inhibited, indicating there existed immune escape caused by T-cell depletion in the high-risk group. Higher tumor TIDE prediction scores were associated not only with poor efficacy of immune checkpoint inhibition therapy but also with poor survival of patients treated with anti-PD1 and anti-CTLA4 (Zarogoulidis et al., 2013). We then analyzed the expression of 25 common immune checkpoint molecules (including CTLA4, PD1, PD-L1, TIGIT, and LAG3). The results showed that 23 of 25 immune checkpoint molecules were lowly expressed in the high-risk group, suggesting the low-risk patients may be more likely to benefit from the updated immunotherapy. However, the lower TIDE score in the low-risk group suggested that the immune checkpoint blocking (ICB) therapy effect may be worse in the high-risk group, whose overall survival time may be shortened. In general, these data further support that the prognosis of low-risk subtypes is better, which may have a better prospect for immunotherapy.

Subsequently, we tried to explore the relationship between the genomic changes of 11 LMRGs and the susceptibility of LUAD. Further analysis confirmed that six genes (ANGPTL4, LDHA, CYP17A1, HACD1, SLCO1B3, and CD79A) related to B-cell activation were highly co-expressed, three of which (SLCO1B3, CYP17A1, and LDHA) had higher mutation frequency and higher levels of copy number variation, as well as relating to the pathogenesis of LUAD. Therefore, we concluded gene mutation and copy number variation may affect the susceptibility of LUAD patients genotyped by the risk score through influencing B-cell function.

To find the correlation between lipid metabolism and TME, we next conducted further analysis with TIMER and TISIDB databases. TISIDB is a kind of tumor immune-related repository, including 4,176 records from 2,530 publications and 988 genes related to anti-tumor immunity (Ru et al., 2019). CD79A (IgA), a part of BCR complex, was one of the specific antibodies on the surface of B cells (Müller-Winkler et al., 2021). However, when the antigen binds to the BCR complex receptor, CD79A is phosphorylated and internalized to activate B cells into plasma cells or memory B cells. At the same time, the number of lipid raft-related receptors increased, which participated in the regulation of lipid metabolism (Eleftheriadis et al., 2020). A study also confirmed that the increased expression of CD79A

protein was associated with good prognosis of LUAD patients (Enfield et al., 2019). Moreover, the strong correlation between CD79A and risk score, B cells, T cells, TIGIT, BTLA, CD27, and TNFRSF17 revealed the crucial role of CD79A in regulating the tumor microenvironment (TME) in LUAD patients. High-risk genes in the model, including ANGPTL4, LDHA, and NPAS2, have been reported to promote angiogenesis, glycolysis, drug tolerance, the invasive and migratory potential, and cell cycle and apoptosis in lung cancer patients or cell lines (Galaup et al., 2006; Park et al., 2016; Gu et al., 2017; Li et al., 2018). CIDEC, CYP17A1, ELOVL2, and HACD1 were also associated with a variety of cancers including clear cell renal cell carcinoma, prostate cancer, neuroblastoma, and uveal melanoma and mainly participated in the metabolism of lipid storage droplets, promotion of *de novo* androgen biosynthesis, and regulation of fatty acid and inflammatory response (Yu et al., 2013; Xiao et al., 2018; Ding et al., 2019; Xu et al., 2019). The molecular function of PLEKHA6 was related to the hormonal receptor, thus leading to the poor prognosis of breast cancer (Aushev et al., 2019). Overall, we believed that CD79A may be the crucial molecule connecting the two mechanisms due to its dual role in regulation of lipid metabolism *via* participating in B-cell activation and have the potential to influence the response to immunotherapy in LUAD patients. Specifically, when B cells are stimulated by antigens to differentiate into plasma cells and memory B cells, frequent BCR signal transmission leads to the increase of lipid raft quantity and energy demand. At this time, lipid droplets act as regulators to maintain energy metabolism and provide material preparation (such as cholesterol and phospholipids) and energy for lipid rafts and their biological processes. However, once the lipid decomposition process is abnormal, lipid droplets would accumulate in the cells and further lead to lipid metabolism disorder and inflammation. Interestingly, CD79A may also play a role in enhancing the body's immune capacity to repair lipid abnormalities and anti-inflammatory effect in this circumstance.

CONCLUSION

In summary, our study constructed a high efficiency LMRG prognosis predictive model to separate LUAD patients into high- and low-risk groups, which may be helpful to promote the individualized immunotherapy for LUAD patients.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

This article was approved by the Medical Ethics Committee of Shanghai Tongren Hospital, Shanghai Jiao Tong University School of Medicine.

AUTHOR CONTRIBUTIONS

KZ, TZ, and BQ conceptualized and designed the study and wrote, reviewed, and revised the manuscript. KZ, YQ, XQ, and TZ were involved in the acquisition and analysis of data (provided RNA-seq data, statistical analysis, biostatistics, etc.). TZ and BQ supervised the study. All of the authors have read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 81973135).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.730132/full#supplementary-material

REFERENCES

Aushev, V. N., Gopalakrishnan, K., Teitelbaum, S. L., Parada Jr, H., Jr., Santella, R. M., Gammon, M. D., et al. (2019). Tumor Expression of Environmental Chemical-Responsive Genes and Breast Cancer Mortality. *Endocrine-related cancer* 26 (12), 843–851. doi:10.1530/erc-19-0357

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a Cancer J. clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Deng, C., Xu, Y., Fu, J., Zhu, X., Chen, H., Xu, H., et al. (2020). Reprograming the Tumor Immunologic Microenvironment Using Neoadjuvant Chemotherapy in Osteosarcoma. *Cancer Sci.* 111 (6), 1899–1909. doi:10.1111/cas.14398

Ding, Y., Yang, J., Ma, Y., Yao, T., Chen, X., Ge, S., et al. (2019). MYCN and PRC1 Cooperatively Repress Docosahexaenoic Acid Synthesis in Neuroblastoma via ELOVL2. *J. Exp. Clin. Cancer Res.* 38 (1), 498. doi:10.1186/s13046-019-1492-5

Eleftheriadis, T., Pissas, G., Antoniadi, G., Filippidis, G., Liakopoulos, V., and Stefanidis, I. (2020). Urate Crystals Trigger B-Cell Receptor Signal Transduction and Induce B-Cell Proliferation. *J. Basic Clin. Physiol. Pharmacol.* 31 (4). doi:10.1515/jbcpp-2019-0054

Enfield, K. S. S., Martin, S. D., Marshall, E. A., Kung, S. H. Y., Gallagher, P., Milne, K., et al. (2019). Hyperspectral Cell Sociology Reveals Spatial Tumor-Immune Cell Interactions Associated with Lung Cancer Recurrence. *J. Immunotherapy Cancer* 7 (1), 13. doi:10.1186/s40425-018-0488-6

Galaup, A., Cazes, A., Le Jan, S., Philippe, J., Connault, E., Le Coz, E., et al. (2006). Angiopoietin-like 4 Prevents Metastasis through Inhibition of Vascular Permeability and Tumor Cell Motility and Invasiveness. *Proc. Natl. Acad. Sci.* 103 (49), 18721–18726. doi:10.1073/pnas.0609025103

Gharib, E., Nasrinasrabadi, P., and Zali, M. R. (2020). Development and Validation of a Lipogenic Genes Panel for Diagnosis and Recurrence of Colorectal Cancer. *PLoS ONE* 15 (3), e0229864. doi:10.1371/journal.pone.0229864

Gu, F., Zhang, H., Hyland, P. L., Berndt, S., Gapstur, S. M., Wheeler, W., et al. (2017). Inherited Variation in Circadian Rhythm Genes and Risks of Prostate Cancer and Three Other Cancer Sites in Combined Cancer Consortia. *Int. J. Cancer* 141 (9), 1794–1802. doi:10.1002/ijc.30883

Hall, Z., Ament, Z., Wilson, C. H., Burkhart, D. L., Ashmore, T., Koulman, A., et al. (2016). Myc Expression Drives Aberrant Lipid Metabolism in Lung Cancer. *Cancer Res.* 76 (16), 4608–4618. doi:10.1158/0008-5472.can-15-3403

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: the Next Generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Helmink, B. A., Reddy, S. M., Gao, J., Zhang, S., Basar, R., Thakur, R., et al. (2020). B Cells and Tertiary Lymphoid Structures Promote Immunotherapy Response. *Nature* 577 (7791), 549–555. doi:10.1038/s41586-019-1922-8

Hu, Q., Peng, J., Chen, X., Li, H., Song, M., Cheng, B., et al. (2019). Obesity and Genes Related to Lipid Metabolism Predict Poor Survival in Oral Squamous Cell Carcinoma. *Oral Oncol.* 89, 14–22. doi:10.1016/j.oraloncology.2018.12.006

Huang, R., Meng, T., Chen, R., Yan, P., Zhang, J., Hu, P., et al. (2019). The Construction and Analysis of Tumor-Infiltrating Immune Cell and ceRNA Networks in Recurrent Soft Tissue Sarcoma. *Aging* 11 (22), 10116–10143. doi:10.18632/aging.102424

Li, X., Zhang, Z., Zhang, Y., Cao, Y., Wei, H., and Wu, Z. (2018). Upregulation of Lactate-Inducible Snail Protein Suppresses Oncogene-Mediated Senescence through p16INK4a Inactivation. *J. Exp. Clin. Cancer Res.* 37 (1), 39. doi:10.1186/s13046-018-0701-y

Liu, R., Hu, R., Zeng, Y., Zhang, W., and Zhou, H.-H. (2020). Tumour Immune Cell Infiltration and Survival after Platinum-Based Chemotherapy in High-Grade Serous Ovarian Cancer Subtypes: A Gene Expression-Based Computational Study. *EBioMedicine* 51, 102602. doi:10.1016/j.ebiom.2019.102602

Liu, R. Z., Choi, W. S., Jain, S., Dinakaran, D., Xu, X., Han, W. H., et al. (2020). The FABP12/PPARγ Pathway Promotes Metastatic Transformation by Inducing Epithelial-To-Mesenchymal Transition and Lipid-Derived Energy Production in Prostate Cancer Cells. *Mol. Oncol.* 14, 3100–3120. doi:10.1002/1878-0261.12818

Liu, Y., Wu, L., Ao, H., Zhao, M., Leng, X., Liu, M., et al. (2019). Prognostic Implications of Autophagy-Associated Gene Signatures in Non-small Cell Lung Cancer. *Aging* 11 (23), 11440–11462. doi:10.18632/aging.102544

Marijt, K. A., Sluijter, M., Blijleven, L., Tolmeijer, S. H., Scheeren, F. A., van der Burg, S. H., et al. (2019). Metabolic Stress in Cancer Cells Induces Immune Escape through a PI3K-dependent Blockade of IFNγ Receptor Signaling. *J. Immunotherapy Cancer* 7 (1), 152. doi:10.1186/s40425-019-0627-8

Masri, S., Papagiannakopoulos, T., Kinouchi, K., Liu, Y., Cervantes, M., Baldi, P., et al. (2016). Lung Adenocarcinoma Distally Rewires Hepatic Circadian Homeostasis. *Cell* 165 (4), 896–909. doi:10.1016/j.cell.2016.04.039

Müller-Winkler, J., Mitter, R., Rappe, J. C. F., Vanes, L., Schweighoffer, E., Mohammadi, H., et al. (2021). Critical Requirement for BCR, BAFF, and BAFFR in Memory B Cell Survival. *J. Exp. Med.* 218 (2). doi:10.1084/jem.20191393

Orita, H., Coulter, J., Lemmon, C., Tully, E., Vadlamudi, A., Medghalchi, S. M., et al. (2007). Selective Inhibition of Fatty Acid Synthase for Lung Cancer Treatment. *Clin. Cancer Res.* 13 (23), 7139–7145. doi:10.1158/1078-0432.ccr-07-1186

Orita, H., Coulter, J., Tully, E., Kuhajda, F. P., and Gabrielson, E. (2008). Inhibiting Fatty Acid Synthase for Chemoprevention of Chemically Induced Lung Tumors. *Clin. Cancer Res.* 14 (8), 2458–2464. doi:10.1158/1078-0432.ccr-07-4177

Park, H. S., Lim, S. M., Shin, H. J., Cho, A., Shin, J.-G., Lee, M. G., et al. (2016). Pharmacogenetic Analysis of Advanced Non-small-cell Lung Cancer Patients Treated with First-Line Paclitaxel and Carboplatin Chemotherapy. *Pharmacogenetics and genomics* 26 (3), 116–125. doi:10.1097/fpc.0000000000000196

Petitprez, F., de Reyniès, A., Keung, E. Z., Chen, T. W.-W., Sun, C.-M., Calderaro, J., et al. (2020). B Cells Are Associated with Survival and Immunotherapy Response in Sarcoma. *Nature* 577 (7791), 556–560. doi:10.1038/s41586-019-1906-8

Ru, B., Wong, C. N., Tong, Y., Zhong, J. Y., Zhong, S. S. W., Wu, W. C., et al. (2019). TISIDB: an Integrated Repository portal for Tumor-Immune System Interactions. *Bioinformatics (Oxford, England)* 35 (20), 4200–4202. doi:10.1093/bioinformatics/btz210

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Valença, I., and Ferreira, A. R. (2020). Prostate Cancer Proliferation Is Affected by the Subcellular Localization of MCT2 and Accompanied by Significant Peroxisomal Alterations. *Cancers*, 12(11)3152. doi:10.3390/cancers12113152

Vazquez, M. M., Gutierrez, M. V., Salvatore, S. R., Puiatti, M., Dato, V. A., Chiabrando, G. A., et al. (2020). Nitro-oleic Acid, a Ligand of CD36, Reduces Cholesterol Accumulation by Modulating Oxidized-LDL Uptake and Cholesterol Efflux in RAW264.7 Macrophages. *Redox Biol.* 36, 101591. doi:10.1016/j.redox.2020.101591

Wu, F., Zhao, Z., Chai, R. C., Liu, Y. Q., Li, G. Z., Jiang, H. Y., et al. (2019). Prognostic Power of a Lipid Metabolism Gene Panel for Diffuse Gliomas. *J. Cel Mol Med* 23 (11), 7741–7748. doi:10.1111/jcmm.14647

Xiao, L., Wang, Y., Xu, K., Hu, H., Xu, Z., Wu, D., et al. (2018). Nuclear Receptor LRH-1 Functions to Promote Castration-Resistant Growth of Prostate Cancer via its Promotion of Intratumoral Androgen Biosynthesis. *Cancer Res.* 78 (9), 2205–2218. doi:10.1158/0008-5472.can-17-2341

Xu, Y., Han, W., Xu, W. H., Wang, Y., Yang, X. L., Nie, H. L., et al. (2019). Identification of Differentially Expressed Genes and Functional Annotations Associated with Metastases of the Uveal Melanoma. *J. Cel Biochem* 120 (11), 19202–19214. doi:10.1002/jcb.29250

Yu, M., Wang, H., Zhao, J., Yuan, Y., Wang, C., Li, J., et al. (2013). Expression of CIDE Proteins in clear Cell Renal Cell Carcinoma and Their Prognostic Significance. *Mol. Cel Biochem* 378 (1-2), 145–151. doi:10.1007/s11010-013-1605-y

Zarogoulidis, K., Zarogoulidis, P., Darwiche, K., Boutsikou, E., Machairiotis, N., Tsakiridis, K., et al. (2013). Treatment of Non-small Cell Lung Cancer (NSCLC). *J. Thorac. Dis.* 5 (Suppl. 4), S389–S396. doi:10.3978/j.issn.2072-1439.2013.07.10

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhang, Qian, Quan, Zhu and Qian. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 04 March 2022
doi: 10.3389/fcell.2022.794413

A Method of Optimizing Weight Allocation in Data Integration Based on Q-Learning for Drug-Target Interaction Prediction

Jiacheng Sun[1,2,3], *You Lu*[1,2,3]*, *Linqian Cui*[1,2,3], *Qiming Fu*[1,2,3], *Hongjie Wu*[1] and *Jianping Chen*[2,4]*

[1]School of Electronic and Information Engineering, SuZhou University of Science and Technology, Suzhou, China, [2]Jiangsu Province Key Laboratory of Intelligent Building Energy Efficiency, Suzhou University of Science and Technology, Suzhou, China, [3]Suzhou Key Laboratory of Mobile Network Technology and Application, Suzhou University of Science and Technology, Suzhou, China, [4]School of Architecture and Urban Planning, Suzhou University of Science and Technology, Suzhou, China

Calculating and predicting drug-target interactions (DTIs) is a crucial step in the field of novel drug discovery. Nowadays, many models have improved the prediction performance of DTIs by fusing heterogeneous information, such as drug chemical structure and target protein sequence and so on. However, in the process of fusion, how to allocate the weight of heterogeneous information reasonably is a huge challenge. In this paper, we propose a model based on Q-learning algorithm and Neighborhood Regularized Logistic Matrix Factorization (QLNRLMF) to predict DTIs. First, we obtain three different drug-drug similarity matrices and three different target-target similarity matrices by using different similarity calculation methods based on heterogeneous data, including drug chemical structure, target protein sequence and drug-target interactions. Then, we initialize a set of weights for the drug-drug similarity matrices and target-target similarity matrices respectively, and optimize them through Q-learning algorithm. When the optimal weights are obtained, a new drug-drug similarity matrix and a new drug-drug similarity matrix are obtained by linear combination. Finally, the drug target interaction matrix, the new drug-drug similarity matrices and the target-target similarity matrices are used as inputs to the Neighborhood Regularized Logistic Matrix Factorization (NRLMF) model for DTIs. Compared with the existing six methods of NetLapRLS, BLM-NII, WNN-GIP, KBMF2K, CMF, and NRLMF, our proposed method has achieved better effect in the four benchmark datasets, including enzymes(E), nuclear receptors (NR), ion channels (IC) and G protein coupled receptors (GPCR).

Keywords: drug-target interactions, heterogeneous information fusion, q-learning, weight distribution, drug similarity, target similarity

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Wenzheng Bao,
Xuzhou University of Technology, China
Yi Xiong,
Shanghai Jiao Tong University, China

***Correspondence:**
You Lu
luyou@usts.edu.cn
Jianping Chen
alanjpchen@aliyun.com

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *13 October 2021*
Accepted: *14 February 2022*
Published: *04 March 2022*

Citation:
Sun J, Lu Y, Cui L, Fu Q, Wu H and Chen J (2022) A Method of Optimizing Weight Allocation in Data Integration Based on Q-Learning for Drug-Target Interaction Prediction. Front. Cell Dev. Biol. 10:794413. doi: 10.3389/fcell.2022.794413

INTRODUCTION

Diseases are usually caused by defective proteins in the body or the functional structure of viral proteins. Effective drugs can combine well with these proteins, removing the original function and achieving the therapeutic effect (Ding et al., 2021). Previous research and development of novel drugs mainly relied on biochemical experiments, which are risky, expensive and time-consuming. In

addition, protein-protein interactions play a key role in many biological processes (Guo et al., 2015), leading to the emergence of large-scale experimental data on genes and proteins, making drug discovery and repositioning in biomedical research more difficult (Ding et al., 2019a). The purpose of DTIs prediction is to identify potential novel drugs or novel targets for existing drugs, and provide a list of candidate drugs for drug discovery, thus greatly improving the efficiency of research and development and reducing the cost of experiments. In recent years, the research methods of DTIs prediction have been proposed successively. So far, a large number of associations known to date have been proved by previous experiments and stored in some public databases (Wang et al., 2021). Therefore, existing methods predict DTIs mainly based on a small number of experimentally validated interactions in existing databases, such as DrugBank (Wishart et al., 2008), KEGG DRUG (Kanehisa et al., 2012), and SuperTarget (Günther et al., 2008). Previous studies have shown that DTIs prediction based on experimental verification can effectively predict some novel interactions between drugs and targets, and the computational methods used to predict DTIs can significantly improve the efficiency of drug discovery.

In general, traditional computational methods proposed for DTIs prediction are mainly based on ligands and target-based (Liu et al., 2021). The ligand-based method usually compares the candidate ligand with the known interacting ligand of the target to determine the binding between them. However, the disadvantage is that the ligand-based method cannot be used for targets which have no or only a small amount of known binding ligands. The target-based (or docking simulation) method uses docking techniques to predict interactions between drug candidates and targets. Nevertheless, this method requires detailed structures. Not all proteins have structural information (Ding et al., 2020). For example, most targets of GPCRs (G protein coupled receptors) are unknown (Yan et al., 2019).

In order to solve the difficulties of traditional methods, chemical genomics methods have been successfully used for large-scale drug discovery and repositioning (Chen et al., 2018). The purpose of chemical genomics research is to integrate drug and target information into a unified framework, in order to identify potentially useful compounds (Ezzat et al., 2019). Chemical genomics methods are usually divided into ligand-based, target-based, and target-ligand, all of which are based on similarity between member proteins and targets. In fact, this significant similarity-based view of chemical genomics has made machine learning methods widely used in DTIs prediction tasks (Ding et al., 2017; Lo et al., 2018; Ding et al., 2019b; Ding et al., 2019c; Qu et al., 2019; Liu et al., 2020; Bagherian et al., 2021), where DTIs prediction is regarded as a binary classification problem, in which drug-target pairs are taken as examples, the internal chemical structure of the drug and the amino acid subsequence of the target are considered as features. Then, we can build a binary classification model through some classic classification methods, such as support vector machines (SVM), neural networks and nearest neighbors. Yamanishi et al. (Yamanishi et al., 2010) proposed a framework for supervised bipartite graph reasoning, and this framework could predict unknown DTIs based on chemical, genomic and pharmacological data. Tabei et al. (Tabei et al., 2012) used complex protein pairs represented by eigenvectors to correspond to chemical substructures and protein domains, and used logistic regression and SVM to establish a prediction model. Mei et al. (Mei et al., 2013a) improved the bipartite local model by considering novel drug candidates through the interaction profile of neighbors. In short, DTIs prediction based on machine learning comes down to a process of data collection, feature representation, similarity calculation and machine learning modeling. Although machine learning has achieved good results in DTIs prediction, there are still bottlenecks. For example, most drug relocation methods based on machine learning only use a single metric to evaluate the similarity between diseases and between drugs. In fact, the similarity of drugs/diseases is not only noisy, but also multimodal, which can be measured from different aspects (Yang et al., 2021). The fusion of multiple similarity measures can avoid the noise and extract effective features in individual similarity calculation, thus effectively improving the accuracy of DTIs prediction.

In order to overcome the bottleneck encountered by machine learning in DTIs prediction, improving the prediction performance of DTIs by integrating heterogeneous information related to drugs and targets has been extensively studied by the academic community. Among them, the challenge is how to extract and integrate such information (Wang et al., 2011). Wang et al. (Wang et al., 2011) proposed a kernel function method based on SVM predictor to integrate heterogeneous information sources to improve the accuracy of DTIs prediction. Yan et al. (Yan et al., 2019) proposed a method based on multi-kernel learning and clustering to integrate heterogeneous information sources related to multiple drugs and targets to improve the accuracy of DTIs prediction, in which the weight of linear weighting was obtained through interior point optimization algorithm (Byrd et al., 1999). (Ding et al., 2020) used six different features to characterize protein sequences in the experimental process of predicting the subcellular localization of proteins. These features are constructed into corresponding kernels and combined by optimizing the weight of these kernels by Kernel Target Alignment-based Multiple Kernel Learning (KTA-MKL) (Ding et al., 2020). (Zhou et al., 2021) built a multi-kernel learning model with Hilbert-Schmidt independence criterion (HSIC) to obtain optimal weights for vairous features, and identified ncRNA subcellular localization by using the graph regularization k local hyperplane distance nearest neighbor model. Although the above methods have achieved good results, there are still some problems. For example, Wang et al. (Wang et al., 2011) found that the experimental effect of fusion of three information sources was inferior to that of fusion of two information sources in the process of fusion, and the reason may be that drug targets for the contributions to the similarity

TABLE 1 | Information about the four datasets.

Dataset	E	IC	GPCR	NR
Drugs(n)	445	210	223	54
Targets(m)	664	204	95	26
Interactions	2926	1476	635	90

of DTIs prediction is different, and the integration of equal-weight information cannot reflect the difference of contribution ratio of heterogeneous information, DTIs prediction accuracy improved range is limited. To solve the above problems, we used Q-learning algorithm to reasonably allocate weights in the process of heterogeneous information integration, and realized DTIs prediction based on the NRLMF model proposed by Liu et al. (Liu et al., 2016). Finally, our method QLNRLMF achieved good results on four benchmark datasets.

The main contributions of this paper are summarized as follows:

- According to three heterogeneous data sources, including drug chemical structure, target protein sequence and drug-target interaction, we obtain three drug-drug similarity matrices and three target-target similarity matrices using different similarity calculation methods.
- We use Q-learning algorithm to reasonably allocate weight to the similarity matrices obtained above, and then perform linear weighted integration after obtaining the optimal weight.
- We perform DTIs prediction experiments on four benchmark datasets. Experimental results show that the DTIs prediction accuracy of QLNRLMF model is better to several advanced comparison models

MATERIALS AND METHODS

Problem Description

In this paper, the drug set is defined as $D = \{d_i\}_{i=1}^{m}$, and the target set is defined as $T = \{t_j\}_{j=1}^{n}$, where m and n respectively represent the number of drugs and the number of targets. The interaction between the drug and the target is defined as $Y \in R^{m \times n}$, which is composed of drug d_i $(1 \le i \le m)$ and target t_j $(1 \le j \le n)$, the matrix $Y \in R^{m \times n}$ is defined as **Eq. 1**.

$$Y_{ij} = \begin{cases} 1 & Drug\ d_i\ is\ related\ to\ Target\ t_j \\ 0 & Drug\ d_i\ is\ not\ related\ to\ Target\ t_j \end{cases} \quad (1)$$

We take each drug-target pair in Y as a sample, and divide equally all drug-target pairs in Y into ten pieces by random seed shuffled. Then we select one sample as the test set in turn, and the remaining nine samples as the training set. At the same time, we obtain all the samples in the test set corresponding to the subscripts in Y and assign all the corresponding positions to 0, and a new drug and target interaction $Y_{train} \in R^{m \times n}$ is obtained. The purpose of our experiment is to predict these unknown pairs of interactions labeled 0.

TABLE 2 | Summary of similarity matrix of two feature spaces.

Space	Similarity matrix	Description
Drug	$SIM_{strdrug}$	Chemical structure
	$SIM_{cosdrug}$	Cosine similarity of drugs
	$SIM_{Jacdrug}$	Jaccard similarity of drugs
Target	SIM_{strtar}	Target sequence information
	SIM_{costar}	Cosine similarity of target
	SIM_{Jactar}	Jaccard similarity of the target

Data Preparation

Chemical Data

Yamanishi et al. (Yamanishi et al., 2008) obtained the chemical structure of the compound from the drugs and compounds section of the KEGG ligand database (Kanehisa et al., 2007). The chemical structure similarity between drugs is obtained by the SIMCOMP algorithm (Hattori et al., 2003), and the similarity score is calculated according to the number of common substructures between two compounds. In this paper, the chemical structure similarity between two compounds d_i and d_j can be calculated according to the Tanimoto coefficient as follows **Eq. 2**, and the chemical structure similarity matrix of the drug compound is expressed as $SIM_{strdrug}$.

$$SIM_{strdrug_{ij}} = \frac{|d_i \cap d_j|}{|d_i \cup d_j|} \quad (2)$$

Genetic Data

Due to the rapid development of sequencing technology, a large amount of data has been accumulated, we use amino acid sequence data to measure the similarity of proteins. In this paper, normalized Smith-Waterman score (Smith and Waterman, 1981) is used to calculate the sequence similarity between protein t_i and protein t_j as shown in **Eq. 3**, where $SW(t_i, t_j)$ represents the Smith-Waterman score between the target protein t_i and t_j, and the sequence similarity matrix of the protein is expressed as SIM_{strtar}.

$$SIM_{strtar_{ij}} = \frac{SW(t_i, t_j)}{\sqrt{SW(t_i, t_i) SW(t_j, t_j)}} \quad (3)$$

Drug-Target Interactions Data:

As for the performance evaluation of the DTIs prediction algorithm, we will verify it on four benchmark datasets, including E, IC, GPCR, and NR. Get the address of datasets: http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. **Table 1** counts the relevant information of the four datasets, including the number of drugs (n), the number of targets (m), the number of interactions.

According to the above $Y_{train} \in R^{m \times n}$, we use cosine similarity and Jaccard similarity coefficient to calculate the drug-drug similarity matrices $SIM_{cosdrug}$, $SIM_{Jacdrug}$, target-target similarity matrices SIM_{costar}, SIM_{Jactar}.

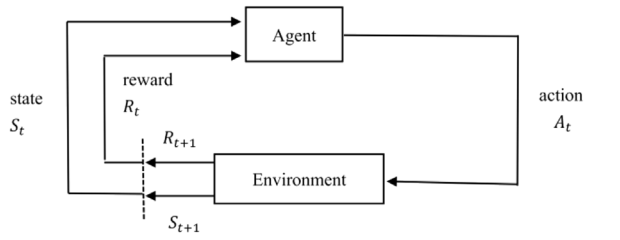


FIGURE 1 | Reinforcement learning.

$$SIM_{cosdrug_{ij}} = \frac{\left|d_i \cap d_j\right|}{\sqrt{\left|d_i\right|\left|d_j\right|}} \tag{4}$$

$$SIM_{Jacdrug_{ij}} = \frac{\left|d_i \cap d_j\right|}{\left|d_i \cup d_j\right|} \tag{5}$$

$$SIM_{costar_{ij}} = \frac{\left|t_i \cap t_j\right|}{\sqrt{\left|t_i\right|\left|t_j\right|}} \tag{6}$$

$$SIM_{Jactar_{ij}} = \frac{\left|t_i \cap t_j\right|}{\left|t_i \cup t_j\right|} \tag{7}$$

Table 2 lists the above similarity matrices. There are three similarity matrices in the drug space and the target space respectively.

NRLMF

NRLMF is the method proposed by Lin et al. (Liu et al., 2016), which combines Logistic matrix decomposition (LMF) and domain regularization to predict associations. Some of the equations that will be used in the model are as follows:

$$\begin{aligned} \underset{\mathbf{U},\mathbf{V}}{min} \sum_{i=1}^{m}\sum_{j=1}^{n}\left(1 + cy_{ij} - y_{ij}\right)\ln\left[1 + \exp\left(\mathbf{u}_i\mathbf{V}_j^{\mathrm{T}}\right)\right] - cy_{ij}\mathbf{u}_i\mathbf{V}_j^{\mathrm{T}} \\ + \frac{1}{2}\mathrm{tr}\left[\mathbf{U}^{\mathrm{T}}\left(\lambda_d\mathbf{I} + \alpha\mathbf{L}^d\right)\mathbf{U}\right] + \frac{1}{2}\mathrm{tr}\left[\mathbf{V}^{\mathrm{T}}\left(\lambda_t\mathbf{I} + \beta\mathbf{L}^t\right)\mathbf{V}\right] \end{aligned} \tag{8}$$

Denoting the objective function in **Eq. 8** by L, the partial gradients with respect to U and V are as follows:

$$\frac{\partial L}{\partial \mathbf{U}} = \mathbf{PV} + (c-1)(\mathbf{Y} \odot \mathbf{P})\mathbf{V} - c\mathbf{YV} + \left(\lambda_d\mathbf{I} + \alpha\mathbf{L}^d\right)\mathbf{U} \tag{9}$$

$$\frac{\partial L}{\partial \mathbf{V}} = \mathbf{P}^{\mathrm{T}}\mathbf{U} + (c-1)\left(\mathbf{Y}^{\mathrm{T}} \odot \mathbf{P}^{\mathrm{T}}\right)\mathrm{U} - c\mathbf{Y}^{\mathrm{T}}\mathbf{U} + \left(\lambda_t\mathbf{I} + \beta\mathbf{L}^t\right)\mathbf{V} \tag{10}$$

where $\mathbf{P} \in R^{m \times n}$, in which the (I, j) element is P_{ij}, $\odot$ denotes the Hadamard product of two matrices. View reference Lin et al. (Liu et al., 2016), where we can see more details about above equation.

Q-Learning Algorithm

Reinforcement learning is an algorithm model that can make the best strategy through self-study in a particular situation. As shown in **Figure 1**, for real-world problems, reinforcement learning can be modeled by abstracting them as the interaction process between agent and environment. For each time step, the agent receives the state of the environment and chooses to perform the corresponding action, and then in the next time step, the agent obtains a reward and a new state based on the feedback of the environment. In other words, reinforcement learning refers to continuously learning to adapt to the environment according to the rewards obtained, and the goal of the agent is to maximize the expected cumulative reward.

Q-learning is a value-based algorithm among reinforcement learning algorithms. Q (*s*, *a*) is the expectation that taking action *a* (*a*∈A) can obtain benefits under state *s* (*s*∈S) at a certain moment, and the environment will feedback corresponding return *r* according to the action taken by the agent. Therefore, the main idea of this algorithm is to construct a Q-Table (as shown in **Table 3**) between state S and action A to store Q value, and then select the action that can obtain the maximum benefit according to Q value.

The updated formula for Q value calculation in Q-Table is as follows:

$$\boldsymbol{Q}(s,a) \leftarrow \boldsymbol{Q}(s,a) + \boldsymbol{\alpha}\left[\boldsymbol{r} + \boldsymbol{\gamma} max_{a'}\boldsymbol{Q}\left(s',a'\right) - \boldsymbol{Q}(s,a)\right] \tag{11}$$

Where α is learning rate and γ is discount rate.

MDP Modeling

State Space Design

Before the experiment begins, we initialize a set of weights for the drug similarity matrix and the target similarity matrix respectively, and then find a set of optimal weight assignments through the Q-learning algorithm. The state space at time step t is defined as $S_t = (\alpha_t, \beta_t, \gamma_t, x_t, y_t, z_t)$, where α_t, β_t, γ_t, x_t, y_t and z_t represent respectively the weights in front of the three drug similarity matrices and the three target similarity matrices at the time step t, and $\alpha_t + \beta_t + \gamma_t = 1$, $x_t + y_t + z_t = 1$.

$$SIM_{drug} = \alpha SIM_{strdrug} + \beta SIM_{cosdrug} + \gamma SIM_{Jacdrug} \tag{12}$$

$$SIM_{tar} = xSIM_{strtar} + ySIM_{costar} + zSIM_{Jactar} \tag{13}$$

Action Space Design

In order to optimize the weight of the drug similarity matrix and the target similarity matrix, we define the action space as A = (a, b, c, -(b + c)), where a∈{0, 1}, 0 means that the current action only acts on the weight of the drug similarity matrix, and 1 means that it only acts on the weight of the target similarity matrix; b, c∈{−0.1, 0, 0.1}, for example, the current state is (0.3, 0.3, 0.4, 0.3, 0.3, 0.4) and the executed action is (0, 0.1, 0.1, −0.2), so the next state is (0.4, 0.4, 0.2, 0.3, 0.3, 0.4).

TABLE 3 | Q-Table.

Q-Table	a_1	a_2
s_1	q (s_1, a_1)	q (s_1, a_2)
s_2	q (s_2, a_1)	q (s_2, a_2)
s_3	q (s_3, a_1)	q (s_3, a_2)

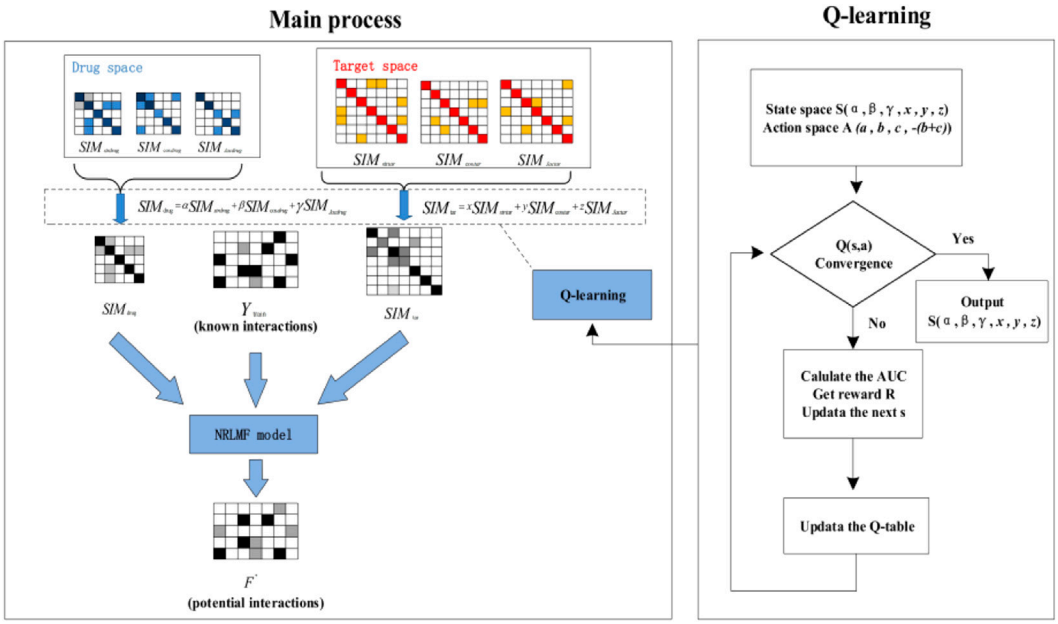


FIGURE 2 | Algorithm flow chart.

Design of Reward Function

After each action is performed, environment will respond with an instant reward R in return. In this paper, we will use AUC as a reward. When we perform an action a at each time step, we will get a new state s that is a new weight. Then, data integration is carried out according to the new weights, and the integration results and drug-target interactions data are taken as the input of the NRLMF model, and the output result AUC is used as the instant reward after performing the corresponding action. If the state does not match after the action is taken, that is, there is a situation that is less than 0 or greater than 1 in the state, we will feedback 0 as an instant reward.

$$R(S_t) = \begin{cases} AUC & If\ all\ weights\ are\ between\ 0\ and\ 1 \\ 0 & Otherwise \end{cases} \quad (14)$$

QLNRLMF

In our experiment, we divide the dataset into training set and test set in a ratio of 9:1, and validate the performance with ten-fold-cross-validation. The main ideas of our method are as follows: First, we use different similarity calculation methods to calculate heterogeneous data sources and obtain drug Space and target Space. Then, we initialize weights for the matrices in Drug space and target space respectively and optimize them through Q-learning algorithm. After obtaining the optimal weight, linear weighted integration is carried out to obtain SIM_{drug} and SIM_{tar}. Finally, we take SIM_{drug}, SIM_{tar} and Y_{train} as the inputs of NRLMF model for DTIs prediction.

The pseudocode of QLNRLMF algorithm is shown in Algorithm 1. The overall flow chart for QLNRLM is shown in **Figure 2**.

Algorithm 1. Qlnrlmf

Require: drugs set **D**, targets set **T**, drug-target interactions **Y**, state space **S**, action space **A**, reward **R**, learning_rate α, discount rate γ
Ensure: The prediction of $S(\alpha, \beta, \gamma, x, y, z)$
1. Divide the dataset and get the Y_{train}
2. Calculate the similarity matrix $SIM_{strdrug}$、$SIM_{cosdrug}$、$SIM_{jacdrug}$、SIM_{strtar}、SIM_{costar}、SIM_{jactar}
3. Initialize $Q(s, a)$
4. repeat (for each episode):
5. Initialize s
6. repeat (for each step of episode):
7. choose a from s using policy derived from Q(e.g., e_greedy)
8. take action a
9. $s' \leftarrow s, a$
10. $SIM_{drug} = \alpha SIM_{strdrug} + \beta SIM_{cosdrug} + \gamma SIM_{jacdrug}$
11. $SIM_{tar} = xSIM_{strtar} + ySIM_{costar} + zSIM_{jactar}$
12. $r \leftarrow NRLMF(SIM_{drug}, SIM_{tar}, Y_{train})$
13. $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma max_{a'} Q(s', a') - Q(s, a)]$
14. $s \leftarrow s'$
15. until s is terminal
16. until Q(s, a) convergence

RESULTS

Evaluation Measurements

In order to verify the performance of our proposed method, we evaluate it from the feasibility, efficiency and accuracy of the algorithm. First of all, we prove the feasibility of our algorithm through the convergence graph of the average reward for each episode. Secondly, we compare the Q-learning algorithm with the brute force method in terms of the time required to find the optimal weights, to prove the effectiveness of our algorithm.

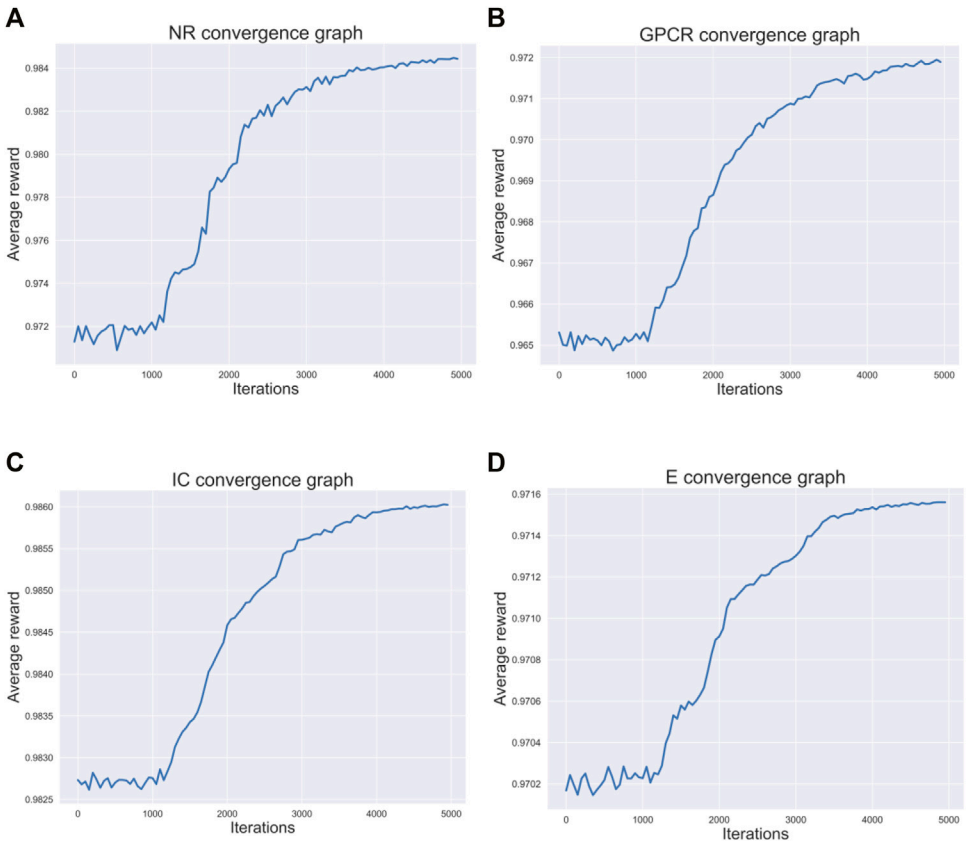


FIGURE 3 | Four datasets convergence graphs. **(A)** The convergence graph of NR dataset. **(B)** The convergence graph of GPCR dataset. **(C)** The convergence graph of IC dataset. **(D)** The convergence graph of E dataset.

TABLE 4 | Time contrast between Q-learning algorithm and Brute Force algorithm.

Dataset	1	2	3	4	5	6	7	8	9	10	Avg	Avg/ S_{num}
NR	1023	1001	1032	1027	1029	987	1008	1016	1035	1005	1016	0.784
GPCR	1033	1069	1001	1050	1015	1055	1050	1015	1050	1024	1036	0.799
IC	997	1021	1030	1034	1045	1034	1027	1012	1006	1017	1022	0.789
E	1045	994	1036	1039	1031	1001	1032	1064	982	1026	1025	0.791

Finally, we compare our algorithm with other algorithms on AUC and AUPR to prove the accuracy of our algorithm.

Average Reward Convergence Graph

Figure 3 respectively describes the average reward convergence graph of four benchmark datasets under q-Learning algorithm, where the abscissa represents the number of iterations and the ordinate represents the average reward. As shown in the **Figure 3**, the average reward does not show an obvious upward trend in the first 1000 iterations, because the agent is constantly exploring trial and error at the beginning. After that, with the continuous accumulation of experience, the strategies learned by agents are getting better and better, and the average reward also presents an upward trend. When the number of iterations reaches 4000, the average reward basically tends to be stable, indicating that agent has learned an optimal strategy.

Comparisons with Brute Force Algorithm

In this experiment, the number S_{num} of the state space is 1296. If the Brute Force algorithm is adopted, the optimal state can be found by traversing 1296 states, that is, the optimal weight allocation can be obtained. If we use the Q-learning algorithm, we separately record the number of different states visited when each fold finds the optimal weight. As shown in **Table 4**, we can conclude that the ratio of the time required by the four datasets to the time required by the Brute Force algorithm are: 78.4, 79.9,

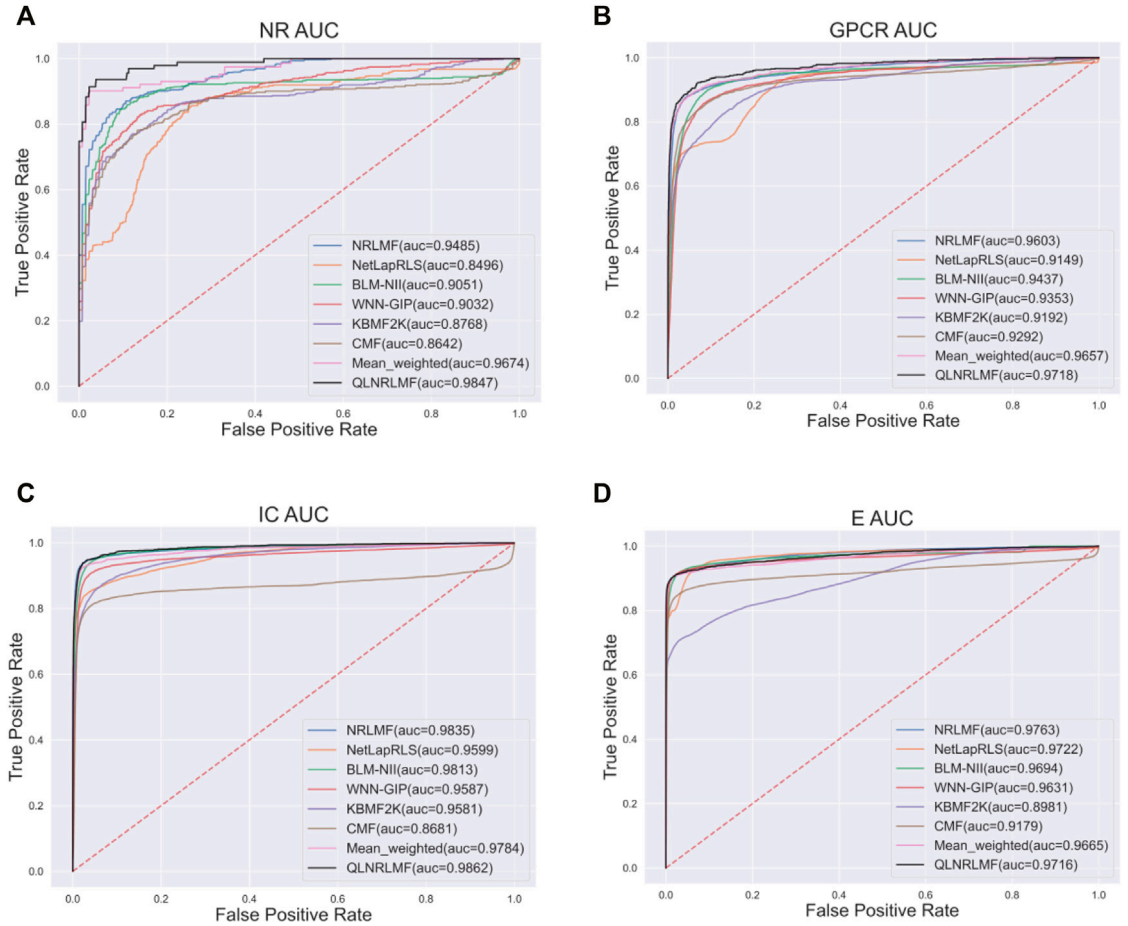


FIGURE 4 | The AUC of QLNRLMF and Other Methods on Benchmark Datasets. **(A)** The AUC of QLNRLMF and Other Methods on NR Dataset. **(B)** The AUC of QLNRLMF and Other Methods on GPCR Dataset. **(C)** The AUC of QLNRLMF and Other Methods on IC Dataset. **(D)** The AUC of QLNRLMF and Other Methods on E Dataset.

78.9, 79.1%, and the average time ratio is 79.1%. All in all, the running time of Q-learning algorithm has been reduced by nearly 20.9% on average.

Comparisons with Other Methods

There are many evaluation indexes for DTIs prediction methods. Among them, AUC and AUPR are widely used. AUC is the area under the receiver operating characteristic (ROC) curve, which can also be understood as the probability that a positive sample and a negative sample are randomly given, and the prediction probability of the positive sample is greater than the probability of the prediction probability of the negative sample. The value of AUC can directly evaluate the performance of the DTIs prediction method, and the greater the value, the better. AUPR is the area under the curve of precision and recall rate, and is a quantitative measurement method that can determine the average separation degree between the predicted fraction of real interactions and the predicted fraction of real non-interactions (Peng et al., 2017). Relatively few interactions are known for DTIs prediction. Therefore, AUPR is a more effective evaluation indicator than AUC, because AUPR takes some measures to reduce the impact of predicted fake DTIs data on the highest ranking scores (Davis and Goadrich, 2006). Therefore, we use these two indicators to evaluate the performance of our proposed method.

Figure 4; **Figure 5** respectively describe the comparison of AUC and AUPR between our method and other seven methods under four benchmark datasets. As shown in the figures, the comparison between our QLNRLMF method and NRLMF method shows that in four datasets, our method improves 3.62, 1.15, 0.27% and -0.47% respectively in AUC, and 18.08, 11.23, 3.25% and −0.48% respectively in AUPR. In general, our method is superior to NRLMF method. In order to prove the advantages of Q-learning algorithm, we also perform the integration experiment of linear equal-weight strategy. The experimental results show that the linear integration strategy based on Q-learning algorithm is obviously better than the linear equal-weight integration strategy. In addition, we also compare

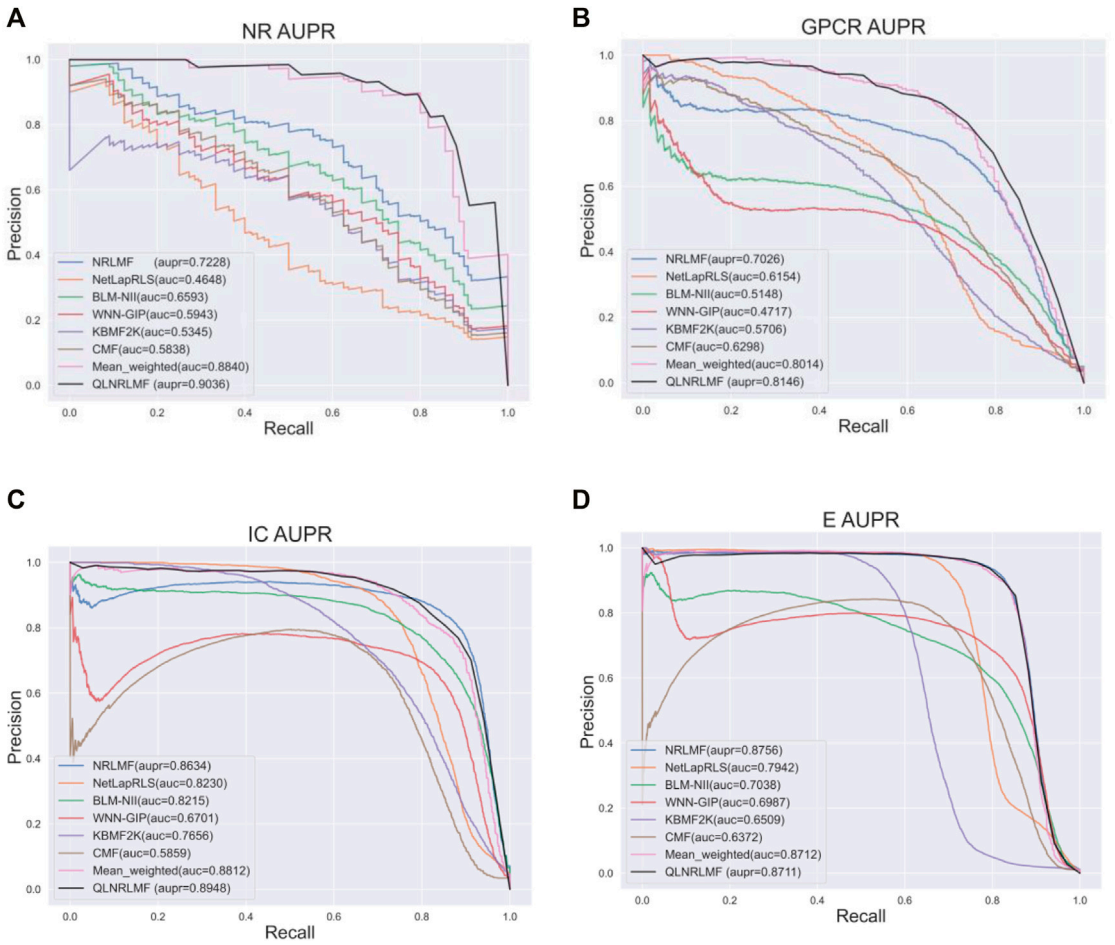


FIGURE 5 | The AUPR of QLNRLMF and Other Methods on Benchmark Datasets. **(A)** The AUPR of QLNRLMF and Other Methods on NR Dataset. **(B)** The AUPR of QLNRLMF and Other Methods on GPCR Dataset. **(C)** The AUPR of QLNRLMF and Other Methods on IC Dataset. **(D)** The AUPR of QLNRLMF and Other Methods on E Dataset.

TABLE 5 | Comparsion with the other seven methods.

AUC								
Dataset	**NetLapRLS**	**BLM-NII**	**WNN-GIP**	**KBMF2K**	**CMF**	**Mean-weighted**	**NRLMF**	**QLNRLMF**
NR	0.849	0.905	0.903	0.876	0.864	0.967	0.948	**0.984**
GPCR	0.914	0.943	0.935	0.919	0.929	0.965	0.960	**0.971**
IC	0.959	0.981	0.958	0.958	0.868	0.978	0.983	**0.986**
E	0.972	0.969	0.963	0.898	0.917	0.966	**0.976**	0.971
Avg	0.923	0.949	0.939	0.912	0.894	0.969	0.966	**0.978**
AUPR								
NR	0.464	0.659	0.594	0.534	0.583	0.884	0.722	**0.903**
GPCR	0.615	0.514	0.471	0.570	0.629	0.801	0.702	**0.814**
IC	0.823	0.821	0.670	0.765	0.585	0.881	0.863	**0.894**
E	0.794	0.703	0.698	0.650	0.637	0.871	**0.875**	0.871
Avg	0.674	0.674	0.608	0.629	0.608	0.859	0.790	**0.870**

The values in bold mean the best result for each line.

our algorithm with other advanced five methods, which contain NetLapRLS (Xia et al., 2010), BLM-NII (Mei et al., 2013b), WNN-GIP (van Laarhoven and Marchiori, 2013), KBMF2K (Gönen, 2012) and CMF (Zheng et al., 2013), we can find that our proposed method has certain advantages from the figures. Finally, we summarize the experimental results and data of all

the above methods into **Table 5**, from which we find that our QLNRLMF method achieves better results in DTIs prediction.

CONCLUSION AND DISCUSSION

In this paper, we propose a model for optimizing weight allocation of heterogeneous data based on Q-learning algorithm to improve the accuracy of DTIs prediction. We obtain multiple drug-drug similarity matrices and target-target similarity matrices from heterogeneous data through different similarity calculation methods, and then optimize the linear weighted weights of these similarity matrices based on Q-learning algorithm. Finally, we perform the experiment of DTIs prediction based on NRLMF model. To evaluate the performance and advantages of our proposed QLNRLMF method, we perform a series of experiments on four benchmark datasets to demonstrate the feasibility, efficiency, and accuracy of our proposed method. There are two main advantages to our approach. In our study, we use AUC and AUPR as evaluation indicators to evaluate the performance of our proposed method. On the one hand, it can be seen from the experimental results that our method achieves better results on the four benchmark datasets compared with other methods. On the other hand, we add reinforcement learning method, which speeds up the acquisition of optimal weight and enables us to predict DTIs more effectively.

Through this experiment, we can find that the integration of multiple information can improve the prediction accuracy of DTIs to some extent, and the rational allocation of the weight of these information also plays a key role in the prediction performance of DTIs. One of our future work directions is to further propose better optimization algorithms for weight distribution. For example, we can change the action space from discrete to continuous to make the weight distribution more accurate, so as to further improve the accuracy of DTIs prediction.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

JS, YL, and HW: conception. JS: experiment and analysis of data. JS, LC and QF:preparation of the manuscript. JC:supervision. All authors contributed to the article and approved the submitted version.

FUNDING

The research was funded by the National Key Research and Development Program of China (2020YFC2006602), the National Natural Science Foundation of China (62176175, 62073231, 62072324, 61902272, 61902271, 61876217, 61876121, 61772357); The Key Research and Development Program of Jiangsu Province (BE2020026).

REFERENCES

Bagherian, M., Sabeti, E., Wang, K., Sartor, M. A., Nikolovska-Coleska, Z., and Najarian, K. (2021). Machine Learning Approaches and Databases for Prediction of Drug-Target Interaction: a Survey Paper. *Brief. Bioinform.* 22, 247–269. doi:10.1093/bib/bbz157

Byrd, R. H., Hribar, M. E., and Nocedal, J. (1999). An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM J. Optim.* 9 (4), 877–900. doi:10.1137/S1052623497325107

Chen, R., Liu, X., Jin, S., Lin, J. W., and Liu, J. (2018). Machine Learning for Drug-Target Interaction Prediction. *Mol. : A J. Synth. Chem. Nat. Product. Chem.* 23, n. pag. doi:10.3390/molecules23092208

Davis, J., and Goadrich, M. H. (2006). "The Relationship between Precision-Recall and ROC Curves," in Proceedings of the 23rd international conference on Machine learning, n. pag. doi:10.1145/1143844.1143874

Ding, Y., Tang, J., and Guo, F. (2021). Identification of Drug-Target Interactions via Multi-View Graph Regularized Link Propagation Model. *Neurocomputing* 2021, n.p-ag. doi:10.1016/j.neucom.2021.05.100

Ding, Y., Tang, J., and Guo, F. (2019). Identification of Drug–Target Interactions via Fuzzy Bipartite Local Model. *Neural Comput. Appl.* 32, 10303–10319. doi:10.1007/s00521-019-04569-z

Ding, Y., Tang, J., and Guo, F. (2019). Identification of Drug-Side Effect Association via Multiple Information Integration with Centered Kernel Alignment. *Neurocomputing* 325, 211–224. doi:10.1016/j.neucom.2018.10.028

Ding, Y., Tang, J., and Guo, F. (2019). Identification of Drug-Side Effect Association via Semisupervised Model and Multiple Kernel Learning. *IEEE J. Biomed. Health Inform.* 23, 2619–2632. doi:10.1109/jbhi.2018.2883834

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. *Knowledge-Based Syst.* 204, 106254. doi:10.1016/j.knosys.2020.106254

Ding, Y., Tang, J., and Guo, F. (2020). Human Protein Subcellular Localization Identification via Fuzzy Model on Kernelized Neighborhood Representation. *Appl. Soft Comput.* doi:10.1016/j.asoc.2020.106596

Ding, Y., Tang, J., and Guo, F. (2017). Identification of Drug-Target Interactions via Multiple Information Integration. *Inf. Sci.* 418-419, 546–560. doi:10.1016/j.ins.2017.08.045

Ezzat, A., Wu, M., Li, X., and Kwoh, C. (2019). Computational Prediction of Drug-Target Interactions Using Chemogenomic Approaches: an Empirical Survey. *Brief. Bioinformatics* 2019, n. pag. doi:10.1093/bib/bby002

Gönen, M. (2012). Predicting Drug-Target Interactions from Chemical and Genomic Kernels Using Bayesian Matrix Factorization. *Bioinformatics* 28, 2304–2310. doi:10.1093/bioinformatics/bts360

Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Sen-ger, C., Petsalaki, E., et al. (2008). "SuperTarget and Matador: Resources for Exploring Drug-Target Relationships. *NucleicAcids Res.* 36, D919–D922. doi:10.1093/nar/gkm863

Guo, F., Ding, Y., Li, Z., and Tang, J. (2015). Identification of Protein-Protein Interactions by Detecting Correlated Mutation at the Interface. *J. Chem. Inf. Model.* 55 (9), 2042–2049. doi:10.1021/acs.jcim.5b00320

Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. (2003). Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways. *J. Am. Chem. Soc.* 125, 11853–11865. doi:10.1021/JA036030U

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2007). KEGG for Linking Genomes to Life and the Environment. *Nucleic Acids Res.* 36, D480–D484. doi:10.1093/nar/gkm882

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets. *NucleicAcids Res.* 40, D109–D114. doi:10.1093/nar/gkr988

Liu, B., Pliakos, K., Vens, C., and Tsoumakas, G. (2021). Drug-Target Interaction Prediction via an Ensemble of Weighted Nearest Neighbors with Interaction Recovery. *ArXiv abs/2012.12325*, n. pag. doi:10.1007/S10489-021-02495-Z

Liu, H., Ren, G., Chen, H., Liu, Q., Yang, Y., and Zhao, Q. (2020). "Predicting lncRNA-miRNA Interactions Based on Logistic Matrix Factorization with Neighborhood Regularized. *Knowl. Based Syst.* 191, 105261. doi:10.1016/j.knosys.2019.105261

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X. (2016). Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. *PLoS Comput. Biol.* 12, n. pag. doi:10.1371/journal.pcbi.1004760

Lo, Y.-C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* 23, 1538–1546. doi:10.1016/j.drudis.2018.05.010

Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., and Zheng, J. (2013, Drug-target Interaction Prediction by Learning from Local Information and Neighbors, 29): 238–245. doi:10.1093/bioinformatics/bts670

Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., and Zheng, J. (2013). Drug-target Interaction Prediction by Learning from Local Information and Neighbors. *Bioinformatics* 29, 238–245. doi:10.1093/bioinformatics/bts670

Peng, L., Liao, B., Zhu, W., Li, Z., and Li, K. (2017). Predicting Drug–Target Interactions with Multi-Information Fusion. *IEEE J. Biomed. Health Inform.* 21, 561–572. doi:10.1109/JBHI.2015.2513200

Qu, J., Chen, X., Yin, J., Zhao, Y., and Li, Z. (2019). Prediction of Potential miRNA-Disease Associations Using Matrix Decomposition and Label Propagation. *Knowl. Based Syst.* 186, n. pag. doi:10.1016/J.KNOSYS.2019.104963

Smith, T. F., and Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *J. Mol. Biol.* 147, 195–197. doi:10.1016/0022-2836(81)90087-5

Tabei, Y., Pauwels, E., Stoven, V., Takemoto, K., and Yamanishi, Y. (2012). Identification of Chemogenomic Features from Drug-Target Interaction Networks Using Interpretable Classifiers. *Bioinformatics* 28, i487–i494. doi:10.1093/bioinformatics/bts412

van Laarhoven, T., and Marchiori, E. (2013). Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile. *PLoS ONE* 8, n. pag. doi:10.1371/journal.pone.0066952

Wang, Y.-C., Zhang, C.-H., Deng, N.-Y., and Wang, Y. (2011). Kernel-based Data Fusion Improves the Drug-Protein Interaction Prediction. *Comput. Biol. Chem.* 35, 353–362. doi:10.1016/j.compbiolchem.2011.10.003

Wang, H., Tang, J., Ding, Y., and Guo, F. (2021). Exploring Associations of Non-coding RNAs in Human Diseases via Three-Matrix Factorization with Hypergraph-Regular Terms on center Kernel Alignment. *Brief. Bioinform.* doi:10.1093/bib/bbaa409

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a Knowledgebase for Drugs, Drug Actions and Drug Targets: a Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* 36, D901–D906. doi:10.1093/nar/gkm958

Xia, Z., Wu, L.-Y., Zhou, X., and Wong, S. T. (2010). Semi-supervised Drug-Protein Interaction Prediction from Heterogeneous Biological Spaces. *BMC Syst. Biol.* 4, S6. doi:10.1186/1752-0509-4-S2-S6

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of Drug-Target Interaction Networks from the Integration of Chemical and Genomic Spaces. *Bioinformatics* 24, i232–i240. doi:10.1093/bioinformatics/btn162

Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. (2010). Drug-target Interaction Prediction from Chemical, Genomic and Pharmacological Data in an Integrated Framework. *Bioinformatics* 26, i246–i254. doi:10.1093/bioinformatics/btq176

Yan, X.-Y., Zhang, S.-W., and He, C.-R. (2019). Prediction of Drug-Target Interaction by Integrating Diverse Heterogeneous Information Source with Multiple Kernel Learning and Clustering Methods. *Comput. Biol. Chem.* 78, 460–467. doi:10.1016/j.compbiolchem.2018.11.028

Yang, M., Wu, G., Zhao, Q., Li, Y., and Wang, J. (2021). Computational Drug Repositioning Based on Multi-Similarities Bilinear Matrix Factorization. *Brief. Bioinformatics* 2021, n. pag. doi:10.1093/bib/bbaa267

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). "Collaborative Matrix Factorization with Multiple Similarities for Predicting Drug-Target Interactions," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, n. pag. doi:10.1145/2487575.2487670

Zhou, H., Wang, H., Tang, J., Ding, Y., and Guo, F. (2021). Identify ncRNA Subcellular Localization via Graph Regularized Klocal Hyperplane Distance Nearest Neighbor Model on Multikernel Learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi:10.1109/TCBB.2021.3107621

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Sun, Lu, Cui, Fu, Wu and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 30 March 2022
doi: 10.3389/fcell.2022.858633

Identification of Potential Biomarkers and Small Molecule Drugs for Cutaneous Melanoma Using Integrated Bioinformatic Analysis

Yong Liu[1,2†], *Jiayi Sun*[1†], *Dongran Han*[1], *Shengnan Cui*[3,4*] *and Xiaoning Yan*[2*]

[1]School of Life Science, Beijing University of Chinese Medicine, Beijing, China, [2]Department of Dermatology, Shaanxi Provincial Hospital of Traditional Chinese Medicine, Xi'an, China, [3]Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing, China, [4]Graduate School, China Academy of Chinese Medical Sciences, Beijing, China

Background: Cutaneous melanoma (CM) is a type of skin cancer with a high fatality rate, and its pathogenesis has not yet been fully elucidated.

Methods: We obtained the gene expression datasets of CM through the Gene Expression Omnibus (GEO) database. Subsequently, robust rank aggregation (RRA) method was used to identify differentially expressed genes (DEGs) between CM cases and normal skin controls. Gene functional annotation was performed to explore the potential function of the DEGs. We built the protein–protein interaction (PPI) network by the Interactive Gene database retrieval tool (STRING) and selected hub modules by Molecular Complexity Detection (MCODE). We furthered and validated our results using the TCGA-GTEX dataset. Finally, potential small molecule drugs were predicted by CMap database and verified by molecular docking method.

Results: A total of 135 DEGs were obtained by RRA synthesis analysis. GMPR, EMP3, SLC45A2, PDZD2, NPY1R, DLG5 and ADH1B were screened as potential targets for CM. Furazolidone was screened as a potential small molecule drug for the treatment of CM, and its mechanism may be related to the inhibition of CM cell proliferation by acting on GMPR.

Conclusion: We identified seven prognostic therapeutic targets associated with CM and furazolidone could be used as a potential drug for CM treatment, providing new prognostic markers, potential therapeutic targets and small molecule drugs for the treatment and prevention of CM.

Keywords: cutaneous melanoma, normal skin, prognosis, furazolidone, bioinformatics

OPEN ACCESS

Edited by:
Yan Zhang,
Harbin Institute of Technology, China

Reviewed by:
Cynthia M Simbulan-Rosenthal,
Georgetown University, United States
Olga Papadodima,
National Hellenic Research Foundation, Greece

***Correspondence:**
Shengnan Cui
csn_sxtcm@163.com
Xiaoning Yan
xny_sxtcm@126.com

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *20 January 2022*
Accepted: *16 March 2022*
Published: *30 March 2022*

Citation:
Liu Y, Sun J, Han D, Cui S and Yan X (2022) Identification of Potential Biomarkers and Small Molecule Drugs for Cutaneous Melanoma Using Integrated Bioinformatic Analysis. Front. Cell Dev. Biol. 10:858633. doi: 10.3389/fcell.2022.858633

INTRODUCTION

Global statistics in 2020 showed that CM accounted for 1.7% of global cancer. Melanoma is the fifth most common cancer diagnosis in the United States by 2021, accounting for 5.6% of all cancer diagnoses. (Saginala et al., 2021). It is widely known, genetic changes facilitate early diagnosis and individualized treatment in patients with melanoma. In the past few years, microarray technology has been widely used in the analysis of gene expression profiles in the skin tissues of melanoma patients or experimental animals. Identifying gene-specific expression patterns makes it possible for

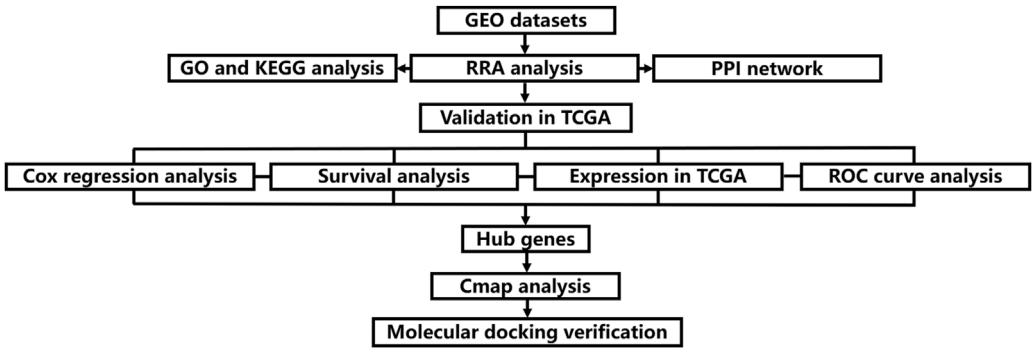


FIGURE 1 | Flow chart of the present study.

people to discover the key gene changes in melanoma tissues and cells, which helps to understand the pathogenic mechanism of the disease or evaluate the treatment. However, there are some inconsistencies in these microarray studies, such as different laboratory conditions, racial differences in clinical samples, and differences in chip platforms. Therefore, it is of great significance to find an effective method to evaluate the results of different gene expression profiles.

RRA is a method that uses probability models to integrate ranked lists which has four key features: strong robustness to noise, ability to deal with incomplete ranking, giving significant scores to each element in the result ranking, and high computational efficiency (Kolde et al., 2012). Some studies have used it to integrate multiple sets of gene chip data lists, and achieved good results (Jia and Zhai, 2019). To our knowledge, previous CM studies did not use the RRA method to identify differentially expressed genes (DEGs), which facilitated this study. Therefore, we adopted a comprehensive bioinformatics approach to conduct a meta-analysis of gene expression between CM tissue and normal skin controls. In addition, according to the results of this analysis, gene enrichment and pathway annotation analysis were also carried out. In addition, based on the results of this analysis, the prediction of small molecule drugs to treat CM was also carried out, and the molecular docking method was used to verify and study the possible mechanism of action. The flow chart of this research is displayed in **Figure 1**.

MATERIALS AND METHODS

Microarray Datasets of Cutaneous melanoma

The expression profiles of CM were retrieved from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/). The search strategy ["Cutaneous melanoma" (MeSH Terms) OR Cutaneous melanoma (All Fields)] and ["*Homo sapiens*" (Organism) and "Expression profiling by array" (Filter)] was adopted.

Datasets Analyses

The expression microarray datasets were all standardized by quantiles. A linear model was used to assess differential expression between CM cases and normal skin controls using R package named "limma." The |log2 fold change (FC)| > 1 and p-value < 0.05 were regarded as the cut-off criteria to determine DEGs (Liu et al., 2021).

Robust Rank Aggregation Analysis

We integrated all the obtained up-regulated and down-regulated gene lists of each dataset using the "Robust Rank Aggregation" R package. Genes with p-value <0.05 and the fold change >1 were considered as significant genes. The adjusted p-value in the RRA tool indicate the possibility of ranking high of each gene in the final gene list.

Functional and Pathway Enrichment Analysis

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of the significant genes in RRA analysis were performed via The Database for Annotation, Visualization, and Integrated Discovery (DAVID 6.8, https://david.ncifcrf.gov/). p-value <0.05 was regarded as the cut-off criteria.

Protein–Protein Interaction Network Establishment and Module Analysis

We uploaded the genes screened by the RRA method to the STRING database to obtain the interaction relationship information between genes, and the cutoff value was set to 0.4 to explore the interaction relationship between genes. Then, the interaction information was imported into Cytoscape, and a protein–protein interaction (PPI) network diagram was constructed, and sorted according to the degree value (Shannon et al., 2003). Modular analysis using Molecular Complex Detection (MCODE) plugin in Cytoscape with threshold nodes numbers >3, k-score = 2 and MCODE scores >3 (Bader and Hogue, 2003).

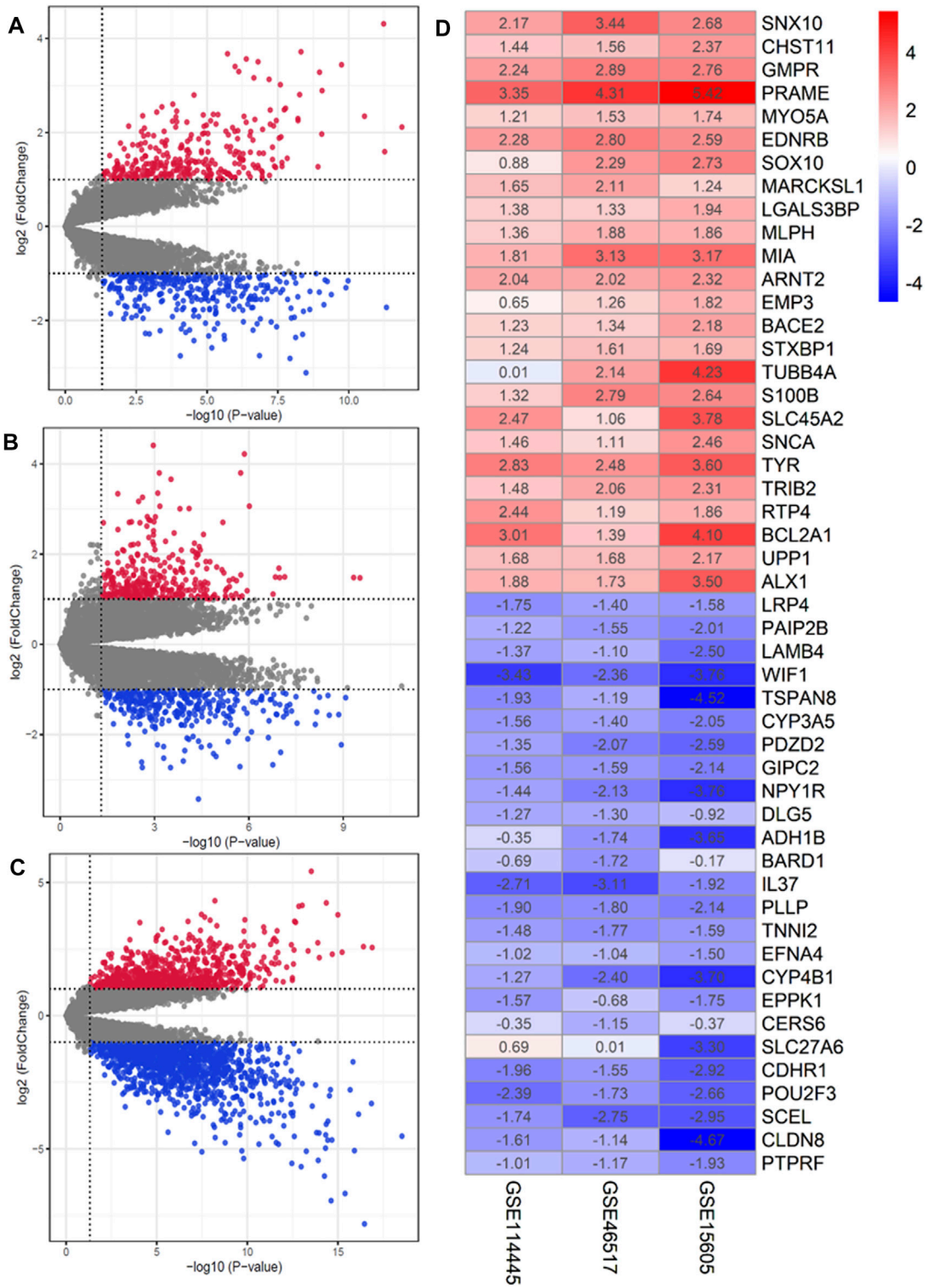


FIGURE 2 | Identification of DEGs in GEO database. **(A–C)** Volcano plot of the DEGs in GSE15605, GSE46517 and GSE114445. Red and green indicate upregulated and downregulated genes ($p < 0.05$, LogFC ≥1 or ≤ −1), respectively. **(D)** Heatmap of the top 25 up- and down-regulated genes in the RRA analysis. The abscissa is the geo ID, and the ordinate is the gene name. Red represents logFC>0, blue represents logFC<0, and the values in the box represent the logFC values.

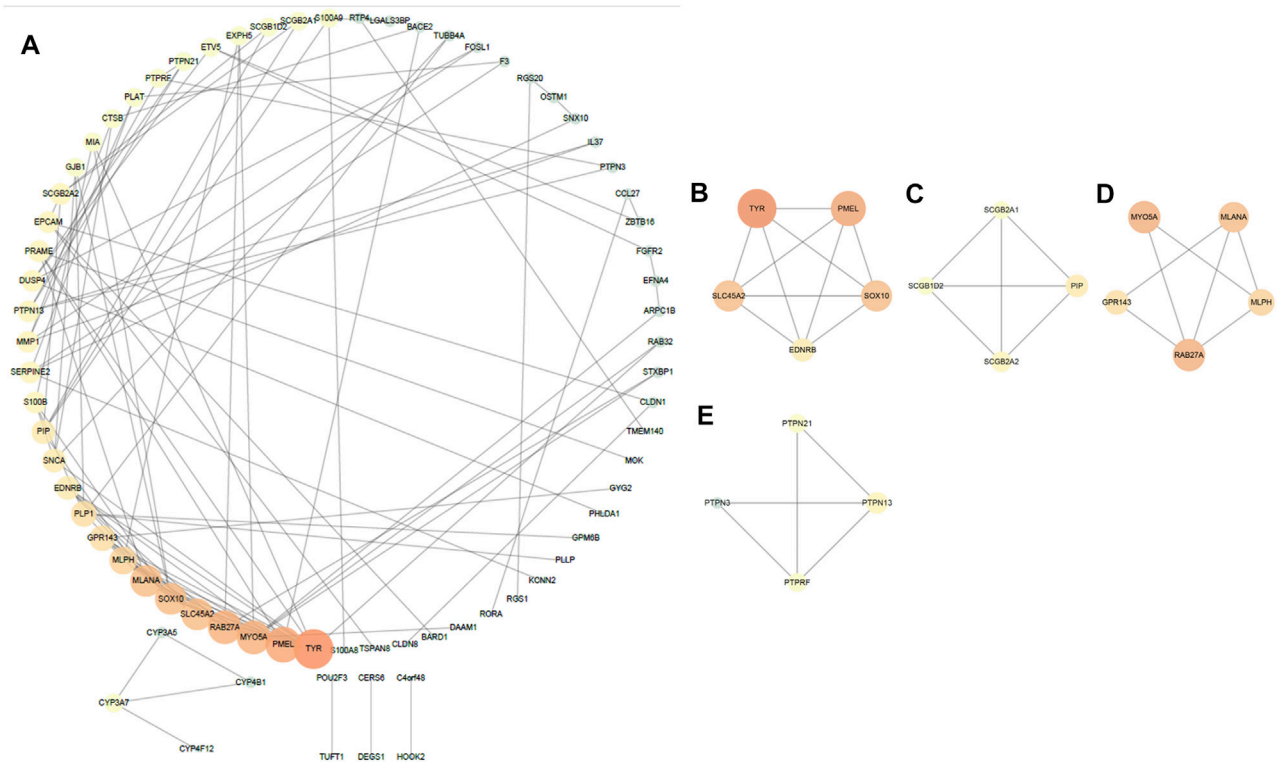


FIGURE 3 | Protein–protein interaction (PPI) network establishment and module analysis of the 135 DEGs. **(A)** PPI network with 75 nodes and 118 edges. **(B–E)** Four modules were identified by MCODE arithmetic.

Key Genes Validation Study

We use the CM RNA-seq dataset (461 CM cases) in the cancer genome atlas (TCGA) TCGA (https://portal.gdc.cancer.gov/) and 558 normal skin tissues RNA-seq dataset in GTEx (https://xenabrowser.net/datapages/) to verify the expression level of the selected key gene. Genes with p-value <0.05 were considered to be significant.

We utilized the CM dataset in TCGA to verify whether the DEGs analyzed by RRA had clinical prognostic significance. Through univariate Cox regression analysis, the Hazard Ratio (HR) value was calculated to judge the clinical prognostic significance of the variable, and the "Survminer" package was used to obtain the median value of the risk score, and the median value was used as the cut-off point to divide CM patients into high-risk groups and low-risk group, and Kaplan-Meier curve analysis was used to compare the survival time of the low-risk and high-risk groups. In addition, survival score and survival status curves and heatmaps were used to illustrate the distribution of CM patients in the two groups (high-risk and low-risk). We also performed ROC curve analysis to assess the predictive value of the results (the size of the area under the ROC curve was used to describe the predictive value). The above analysis was considered statistically significant at $p < 0.05$.

Identification of Candidate Small Molecules

CMap is a program for predicting potential drugs that may induce biological states encoded by specific gene expression signatures (Lamb, 2007). We divided the finally screened differential genes into up-regulated and down-regulated groups, and imported them into CMap database, in order to explore small molecule drugs that might treat CM. A negative mean score indicates that the drug reverses the desired biological property and has potential therapeutic value ($p < 0.05$ for statistical significance).

Molecular Docking Verification

We performed molecular docking verification between the small molecule drugs predicted in CMap and the potential target proteins of CM, and judged the reliability of drug treatment of CM by the size of the binding energy. The mol2 file format structures of the compounds were obtained from the PubChem database, and the crystal structures of the core targets were collected from the RCSB Protein Data Bank (PDB, http://www.rcsb.org/). First, target proteins were dehydrated and ligand-removed using PyMOL 2.3.2 software and stored in PDB format. The processed target protein was then imported into AutoDock Tools 1.5.6 software for hydrogenation, charge calculation and stored in PDBQT format. Mol2 files of small molecule drugs were imported into AutoDock Tools 1.5.6 software, total charge was detected, charge was assigned, flexible rotatable bonds were viewed and saved in PDBQT format. Grid box data for the protein of interest was obtained. Finally, run Autodock Vina 1.1.2 for molecular docking. Molecular docking results were visualized using PyMOL 2.3.2 software (Gaudreault et al., 2015).

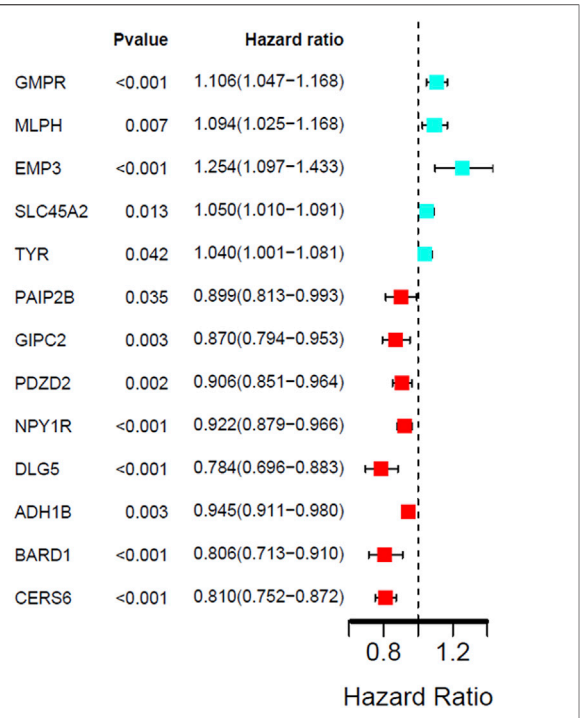


FIGURE 4 | Univariate Cox regression analysis showing the hazard ratios (HRs) with 95% confidence intervals (CIs) and *p* values for 13 DEGs.

RESULTS

Identification of Differentially Expressed Genes in Cutaneous Melanoma

In accordance with our search strategy, we downloaded and analyzed three microarray datasets from the Gene Expression Omnibus (GEO) database, including GSE46517, GSE114445, and GSE15605. GSE46517 was platform-based on GPL96 and contained a total of 39 samples, including 31 primary melanoma samples and 8 normal skin samples. GSE114445 was based on the GPL570 platform, and a total of 22 samples were collected, including 16 primary melanoma sample and six normal skin samples. GSE15605 was based on the GPL570 platform and included a total of 62 samples, including 46 primary melanoma samples and 16 normal skin samples. Primary melanoma sample and normal skin sample from the three datasets were included in this study. The results are shown in **Supplementary Table S1**. The volcano plots of the three microarrays are shown in **Figures 2A–C**. A total of 135 DEGs (70 up-regulated and 65 down-regulated) were obtained through RRA integrated Analysis (**Supplementary Table S1**). The heatmap of the top 25 up and down-regulated genes is shown in **Figure 2D**.

Functional and Network Analysis of Differentially Expressed Genes

GO functional enrichment analysis outcomes revealed that the most significant enrichment was positive regulation of cell proliferation, extracellular region and protein binding respectively among biological process (BP), cellular component (CC) and molecular function (MF). KEGG pathway enrichment analysis showed that pathways in cancer and transcriptional misregulation in cancer were significantly enriched, as shown in **Supplementary Table S1**.

By analyzing 135 DEGs, we got a network interaction graph with 75 nodes and 118 edges, where nodes represented genes, edges represented connections between two genes, and degree value represented the strength of association between genes. More precisely, the top 10 hub genes of DEGs were TYR, PMEL, RAB27A, MYO5A, MLANA, SOX10, SLC45A2, MLPH, GPR143 and PLP1 (**Figure 3A**). Four modules were identified by MCODE arithmetic (**Figures 3B–E**).

The Validation of Key Genes

The univariate cox regression analysis was used to calculate the Hazard Ratio (HR) of the top 25 up and down-regulated genes for CM patients. The results showed that among these top genes, the expression levels of GMPR, MLPH, EMP3, SLC45A2 TYR, PAIP2B, GIPC2, PDZD2, NPY1R, DLG5, ADH1B, BARD1 and CERS6 was closely related to the survival time of CM patients, with statistically significant differences ($p < 0.05$). HR values of less than one can be translated as PAIP2B, GIPC2, PDZD2, NPY1R, DLG5, ADH1B, BARD1 and CERS6 representing low-risk factors, while GMPR, MLPH, EMP3, SLC45A2 and TYR were high-risk factors (**Figure 4**). The genes screened by univariate cox regression analysis were used to draw survival curves from Kaplan-Meier estimations, and 8 genes (GMPR, EMP3, SLC45A2, PDZD2, NPY1R, DLG5, ADH1B, CERS6) finally met the requirements according to statistically significant differences ($p < 0.05$). (**Figures 5**, **6**). (**Supplementary Table S1**).

The expression of 8 hub genes associated with survival time in CM and normal tissues was analyzed in the TCGA- GTEx gene expression dataset and the validation results showed that a total of seven genes showed consistent expression trends in TCGA and GEO datasets, except for CERS6 (CERS6 was up-regulated in CM samples of TCGA dataset and down-regulated in CM samples of GEO dataset) (**Figure 7**).

ROC curve analysis was performed on seven Hub Genes using the package pROC. AUC >0.9 was taken as the cutoff value, and it was found that the AUC values of these seven genes were all greater than 0.9. The expression levels of these genes have high accuracy in distinguishing normal skin tissue from CM tissue, and could be regarded as potential "tumor biomarkers" for diagnosing CM (**Figure 8**). Moreover, the distribution of risk scores, survival status, and expression levels of three oncogenes and four tumor suppressor genes are shown in **Figure 9**.

Small Molecule Drugs Screening

CMap network was used to analyze 7 DEGs into two groups (3 in up-regulated group and four in downregulated group) (**Supplementary Table S1**). After the signature query, the 10 compounds with the highest negative enrichment score (furazolidone, ciclosporin, bisoprolol, rifampicin,

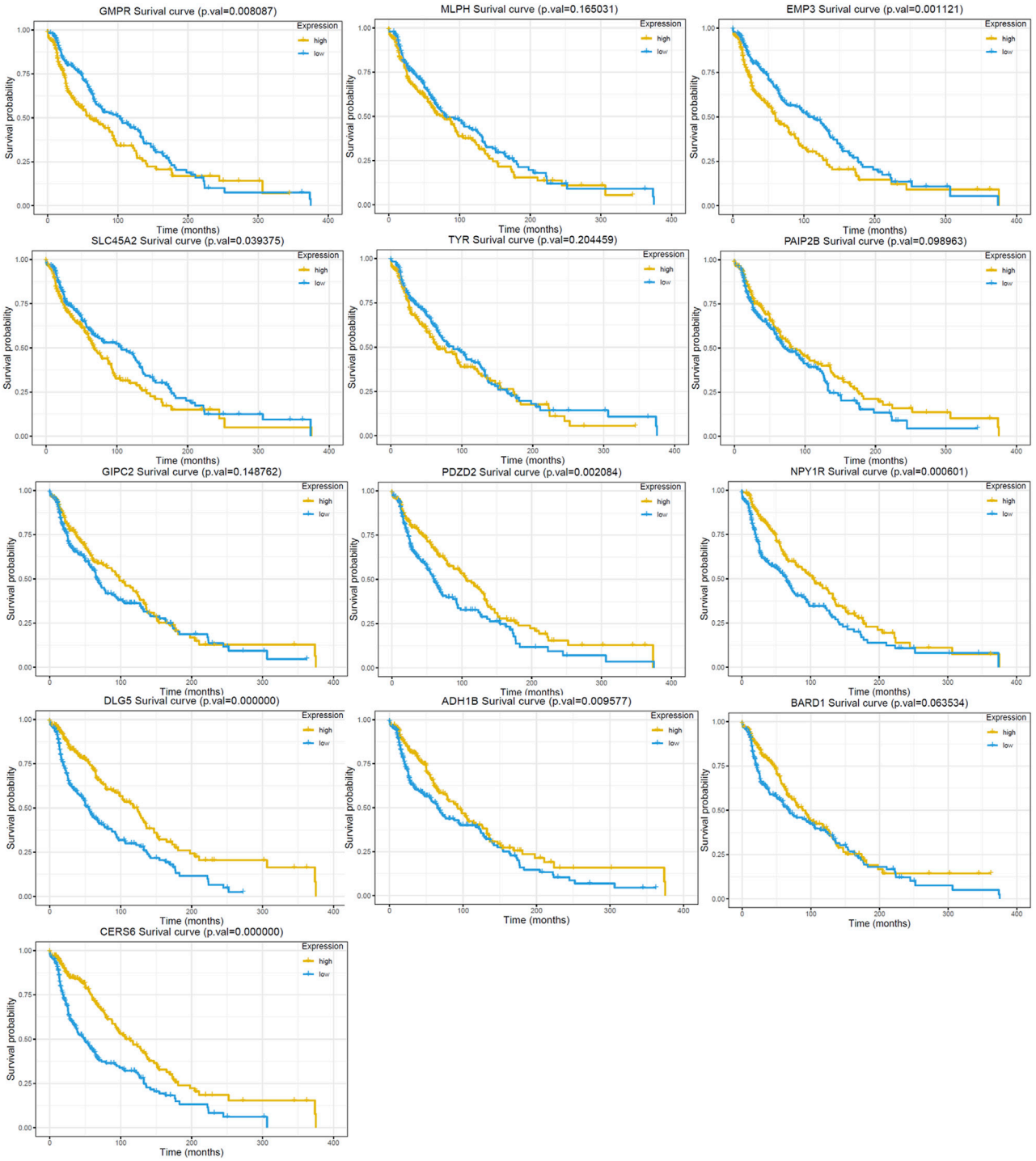


FIGURE 5 | Kaplan Meier survival curve of the 13 DEGs in the RRA analysis. The GMPR, MLPH, EMP3, SLC45A2 and TYR are up-regulated in GEO dataset. The PAIP2B, GIPC2, PDZD2, NPY1R, DLG5, ADH1B, BARD1 and CERS6 are down-regulated in GEO dataset.

pralidoxime, cinchonine, mevalolactone, nifenazone, doxycycline and chenodeoxycholic acid) were identified as potential therapeutic agents for CM (**Table 1**). The chemical structures of these ten compounds are shown in **Figure 10**.

Molecular Docking Verification

Using AutoDock Vina 1.1.2 software, the screened small molecule drugs were docked with six core targets (GMPR, EMP3, SLC45A2, NPY1R, DLG5, ADH1B). The crystal

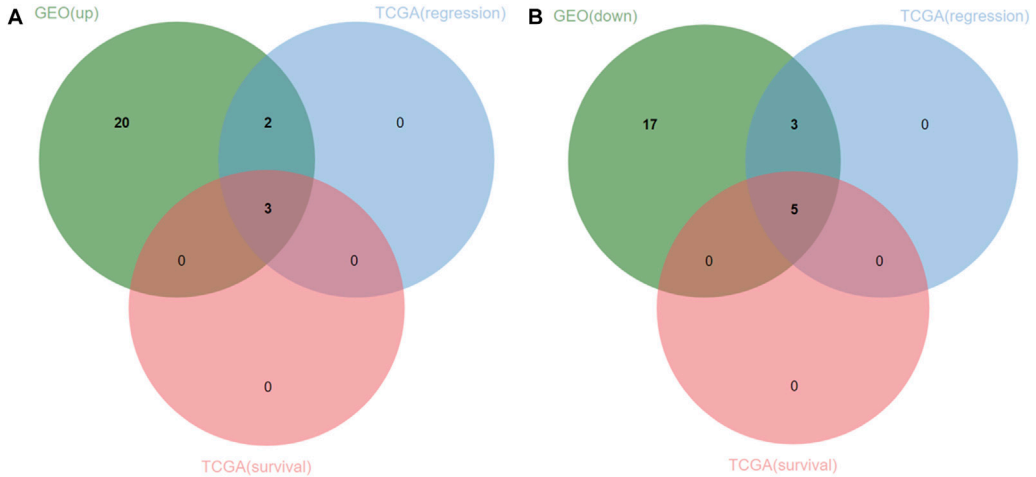


FIGURE 6 | Venn diagram of 25 upregulated and downregulated DEGs and TCGA datasets. **(A)** three upregulated genes (oncogenes) were identified. **(B)** five downregulated genes (tumor suppressor genes) were identified.

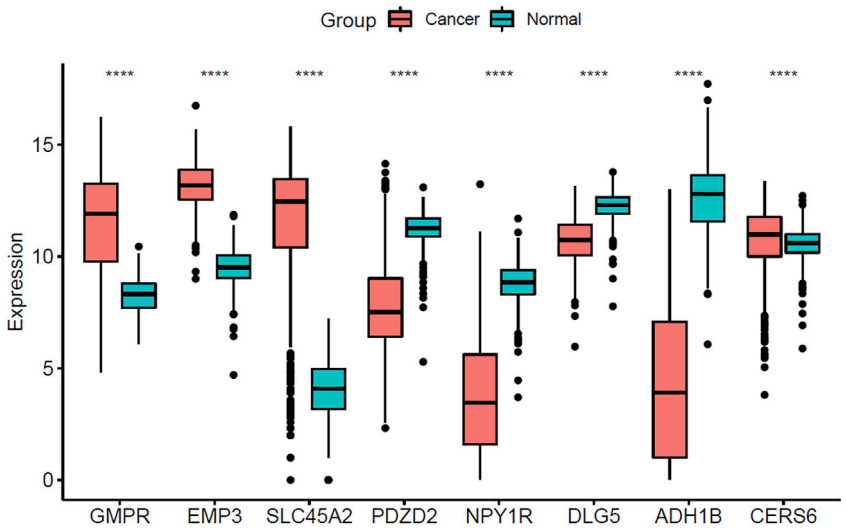


FIGURE 7 | Boxplot of the three screened up-regulated genes (GMPR, EMP3, SLC45A2) and five down-regulated genes (PDZD2, NPY1R, DLG5, ADH1B, CERS6) in TCGA-GTEx dataset. (****$p < 0.0001$).

structure of PDZD2 was not obtained, so it could not be docked. A binding energy less than 0 indicates spontaneous binding of the ligand and receptor. The lower the binding energy, the more stable the binding conformation and the greater the likelihood of action (Gao et al., 2016).

As can be seen from **Figure 11**, the minimum binding energy between the ligand and the receptor is mostly less than −7.0 kcal·mol-1, indicating that the target protein has a good affinity with the active ingredient, and small molecule drugs are likely to act on these targets. Small molecule drug docking targets with the lowest binding energy were selected for docking visualization (**Figure 12**). The dotted lines in the figure are hydrogen bonds. For example, furazolidone exerts its biological efficacy most likely by binding to GMPR and forming hydrogen bonds with the five amino acids GLY221, SER183, GLY242, GLY243 and MET269 near the active site.

DISCUSSION

CM is a tumor formed by malignant transformation of skin melanocytes. It has the characteristics of high degree of malignancy, strong invasiveness, and can affect all ages. If not actively treated, it is easy to spread and metastasize

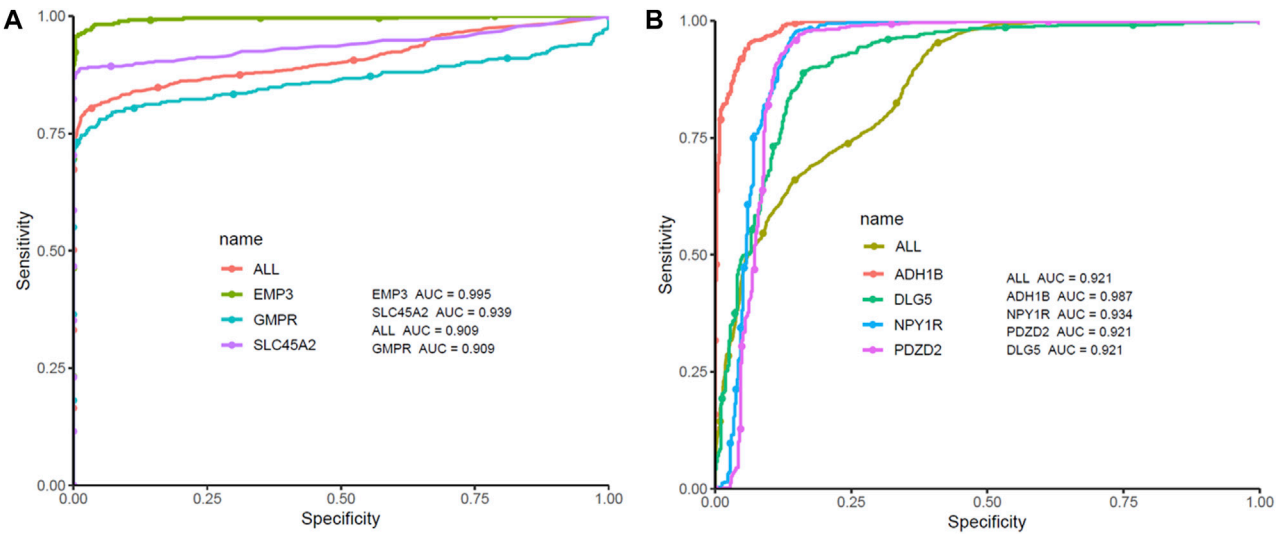


FIGURE 8 | ROC curve analysis and AUC analysis were implemented to evaluate the capacity of seven genes to distinguish CM tissue from normal tissue in TCGA-GTEx dataset. **(A)** ROC curves analysis of three screened up-regulated genes (GMPR, EMP3, SLC45A2). **(B)** ROC curves analysis of four screened down-regulated genes (PDZD2, NPY1R, DLG5, ADH1B).

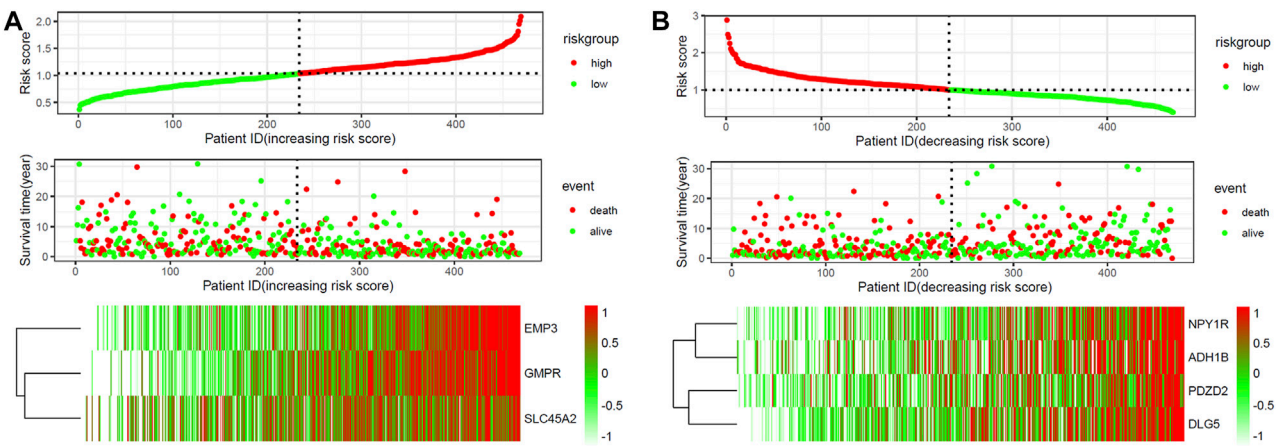


FIGURE 9 | The risk scores for all patients in TCGA cohort are plotted and marked as low risk (blue) or high risk (red), as divided by the threshold (vertical black line). Expression profiles of the DEGs of patients in TCGA cohort, with red indicating higher expression and light blue indicating lower expression. **(A)** three screened up-regulated genes (GMPR, EMP3, SLC45A2). **(B)** four down-regulated genes (PDZD2, NPY1R, DLG5, ADH1B).

through the dermis. Therefore, CM patients have poor prognosis and high mortality (Böhme and Bosserhoff, 2016). Although a series of treatment methods such as radiotherapy, chemotherapy, immunotherapy, and targeted therapy have been used to improve the survival rate of patients, problems such as drug resistance, low drug sensitivity and poor prognosis have to be paid attention to (Domingues et al., 2018; Lebbé et al., 2019; Pelster and Amaria, 2019). Therefore, it is urgent to identify more therapeutic targets, prognostic biomarkers, and potential drugs that can treat CM.

We performed an integrated analysis of the 3 CM DEGs using the RRA method. A total of 135 differential genes were identified, including 65 downregulated and 70 upregulated. GO analysis indicated that the DEGs were associated with melanin biosynthetic process, melanosome and RAGE receptor binding. KEGG analysis showed that these DEGs were primarily enriched in Pathways in cancer. These results are consistent with the existing research results of CM, reflecting the close correlation between the DEGs and CM.

PPI analysis shows that Tyrosinase (TYR) is the most associated and core gene in DEGs. TYR is a copper-

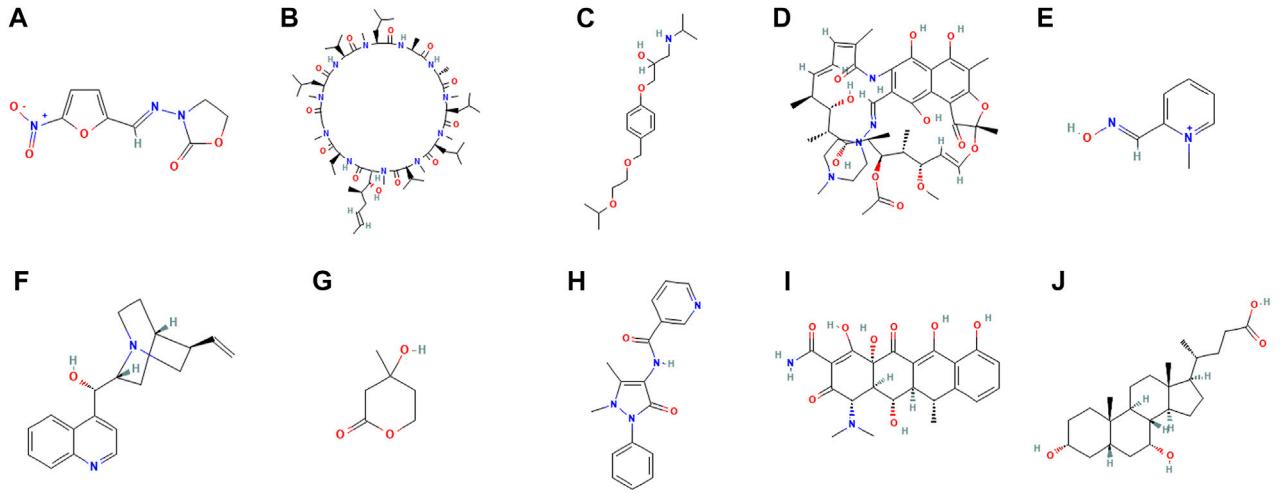


FIGURE 10 | Chemical structure depiction of the top ten most significant drugs. **(A)** Furazolidone **(B)** Ciclosporin **(C)** Bisoprolol **(D)** Rifampicin **(E)** Pralidoxime **(F)** Cinchonine **(G)** Mevalolactone **(H)** Nifenazone **(I)** doxycycline **(J)** chenodeoxycholic acid.

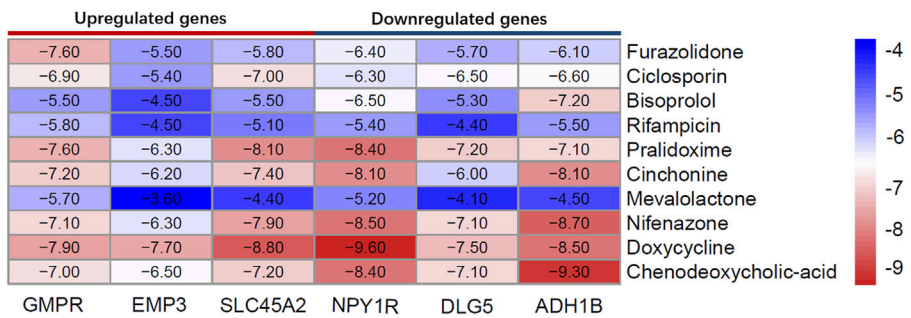


FIGURE 11 | Heat map of the lowest binding energy for molecular docking.

TABLE 1 | Results of Cmap analysis.

Rank	Cmap name	Mean	N	Enrichment	*p*-value
1	Furazolidone	−0.614	4	−0.841	0.00117
2	ciclosporin	−0.418	6	−0.673	0.00332
3	bisoprolol	−0.294	4	−0.784	0.00438
4	rifampicin	−0.294	4	−0.759	0.00688
5	pralidoxime	−0.403	4	−0.757	0.00712
6	cinchonine	−0.241	4	−0.75	0.00778
7	mevalolactone	−0.543	3	−0.836	0.00871
8	nifenazone	−0.488	5	−0.659	0.01103
9	doxycycline	−0.356	5	−0.646	0.01376
10	chenodeoxycholic acid	−0.545	4	−0.712	0.01387

containing oxidase that regulates melanin synthesis (Chang, 2009). Appropriate expression of tyrosinase is beneficial to human beings, as it is a key biochemical catalyst for biosynthesis of natural melanin to protect skin from photocarcinogenesis. Overexpression of tyrosinase can induce melanoma, and TYR can mediate melanoma cell apoptosis under the regulation of transcription factor MITF (Steingrímsson et al., 2004; Shirasugi et al., 2010; Lee et al., 2015; Liu et al., 2020).

Seven genes (GMPR, EMP3, SLC45A2, NPY1R, DLG5, PDZD2 and ADH1B) were finally screened out by combined analysis of GEO and TCGA databases. Cox regression analysis showed that these genes were independent prognostic indicators of CM patients. Among them, only PDZD2, NPY1R and ADH1B have not been reported in melanoma studies, and EMP3 has only been reported in uveal melanoma (Kaochar et al., 2018).

PDZD2 (PDZ domain containing 2) is a multi-PDZ protein expressed in many tissues (Yeung et al., 2003). We found that PDZD2 was poorly expressed in cutaneous melanoma. Previous studies have suggested that PDZD2 is an oncogene, which is over-expressed in the cell lines of osteosarcoma, human primary prostate tumor and prostate tumor (He et al., 2019). Recently, it has been shown that human secreted PDZD2 (sPDZD2) has anti-tumor properties, and it is also down-regulated in lung

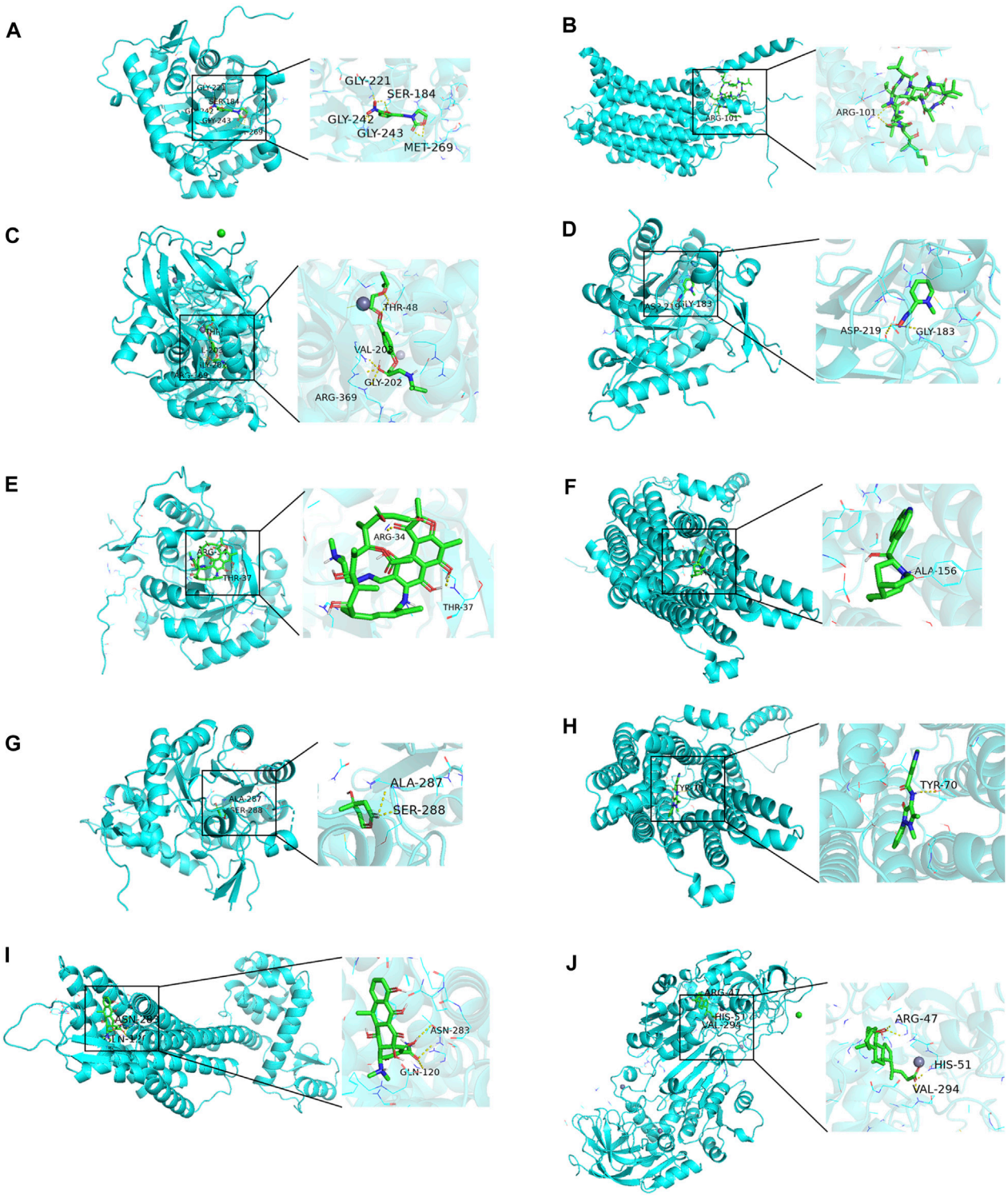


FIGURE 12 | Docking diagram of small molecule drugs with targets. **(A)** Furazolidone-GMPR **(B)** Ciclosporin-SLC45A2 **(C)** bisoprolol-ADH1B **(D)** rifampicin-GMPR **(E)** pralidoxime-NPY1R **(F)** cinchonine-NPY1R **(G)** mevalolactone-GMPR **(H)** nifenazone-SLC45A2 **(I)** doxycycline-NPY1R **(J)** chenodeoxycholic acid-ADH1B.

adenocarcinoma (Cui et al., 2020). Spdzd2 induces the aging of prostate cancer cells through mutation or transcriptional activation of wild-type p53, and makes cancer cells more sensitive to apoptosis through genotoxic stress (Tam et al., 2006). SPDZD2 plays an antiproliferative role in human cancer cells by affecting cell cycle arrest in S phase (Tam et al., 2008). The difference of PDZD2 expression in different tumor tissues may be related to tissue specificity.

NPY is an important regulator of tumor progression of nerve or endocrine related cancers. The role of NPY may be mediated by many NPY receptor subtypes. Many evidences show that the expression of NPY1R gene is up-regulated in various types of nerve or endocrine related cancers (including breast cancer, prostate cancer, adrenal tumor, renal cell cancer and ovarian cancer, etc.), (Kitlinska et al., 2005; Ruscica et al., 2006; Körner et al., 2008; Lv et al., 2016). Activation of Y1R and Y2R by NPY leads to tumor cell proliferation, angiogenesis and metastasis (Liu et al., 2015). In non-neuroendocrine related tumors such as in hepatocellular carcinoma cells, npy1r inhibit cell proliferation by activating mitogen-activated protein kinase signal pathway, promoting tumor growth and increasing tumorigenicity of cells (Lv et al., 2016).

ADH1B, a member of the alcohol dehydrogenase (ADH) family, is involved in the metabolism of acetaldehyde, a carcinogen (Seitz and Stickel, 2010; Galinsky et al., 2016). ADH1B (no. 1) has been inhibited in almost all cancer types, and the ability of ADH1B to inhibit cancer cells has been confirmed *in vitro* experiments (Li et al., 2017). ADH1B is involved in the metabolism of several anti-tumor drugs, including ifosfamide and cyclophosphamide (Polimanti et al., 2016). Differential expression of ADH1B in ovarian cancer can make cells secrete MMP-7 CD-26 and cathepsin to promote cancer progression (Gharpure et al., 2018).

New use of old drugs has become an important strategy for the development of anti-tumor drugs, which has the advantages of saving development time, cost and improving drug safety. According to CMap database analysis, furazolidone (ranked first), ciclosporin and bisoprolol can be used to treat CM.

Furazolidone (FZD) is a kind of synthetic nitrofuran derivative, which can sterilize or inhibit gram-positive and gram-negative bacteria (Karamanakos, 2013). FZD has also been shown to have antitumor activity in various other cancers (Jiang et al., 2013). FZD has significant antitumor activity against acute myeloid leukemia (AML) (Enzenauer et al., 1990). FZD promotes the apoptosis of AML cells and induces the differentiation of myeloid cells by stabilizing the tumor suppressor protein p53 (Tang et al., 2006). Another paper also showed that FZD has inhibitory activity on hepatoma cells. FZD induces oxidative DNA damage by increasing reactive oxygen species (ROS), thereby inducing HCC cell cycle arrest. Targeting nuclear factor -κB (NFκB) signaling pathway may serve as a predictor of immunotherapy response in melanoma patients (Poźniak et al., 2019). FZD can inhibit NF-κB signaling pathway and induce apoptosis of small cell lung cancer cells (Yu et al., 2020). NF-κB is a melanoma pathogenic factor (Pozniak et al.), which can regulate the transcription of genes involved in cell survival, and inhibition of NF-κB activation has been considered as a strategy for the treatment of melanoma. Therefore, FDZ may also treat melanoma by inhibiting the activation of NF-κB.

At the same time, through molecular connection, it was found that FZD was most likely to treat CM by acting on guanosine monophosphate reductase (GMPR). In this study, GMPR is considered to be an oncogene, and some studies have found that GMPR is a new melanoma invasion inhibitor. This seemingly contradictory conclusion happened simultaneously in MITF characterized as both a melanoma oncogene (McGill et al., 2002; Wellbrock et al., 2008) and an invasion suppressor (Carreira et al., 2006; Arozarena et al., 2011; Cheli et al., 2011; Javelaud et al., 2011; Cheli et al., 2012), which may be a kind of "Rheostat model" (Wawrzyniak et al., 2013). GMPR is the downstream target of MITF (Bianchi-Smiraglia et al., 2017). MITF can promote proliferation at low and medium levels and differentiation at high levels, and such dynamic changes are related to the dynamic changes of tumor microenvironment after tumor cell metastasis and epigenetic changes related to melanoma metastasis caused by tumor microenvironment changes after tumor cell metastasis (Carreira et al., 2006). Therefore, FZD may achieve its therapeutic purpose by acting on GMPR to inhibit CM cell proliferation rather than invasion.

In summary, RRA method was used in this study to systematically analyze three groups of CM gene chip data, and then the TCGA data set was combined to verify and screen, and finally the key DEGs, such as GMPR, EMP3, SLC45A2, NPY1R, DLG5, PDZD2 and ADH1B, were screened out. In this study, furazolidone was predicted and validated as a potential small-molecule drug for the treatment of CM, providing reference for the selection of CM markers, therapeutic targets and therapeutic drugs. In future studies, we will need to verify the target genes and small molecule drugs we have found through experiments.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

All of the data used in this study were obtained from public databases. This study does not contain any studies associated with animals or humans.

AUTHOR CONTRIBUTIONS

SC conceived and designed the workflow. YL did the analysis work and wrote the manuscript. JS and YL gave analysis

advice, did figure creation and modification. All authors approved the manuscript.

FUNDING

This study was funded by including the National Key R and D Program of China (Grant Nos. 2018YFC1705303), Innovation Talent Promotion Program—Key Science and Technology Innovation Team Program (Grant Nos.2017KCT-27), Shaanxi province key industry innovation chain (Grant Nos.2019ZDLSF04-08) and National Natural Science Foundation of China (Grant Nos.82174386).

ACKNOWLEDGMENTS

We appreciate GEO database, The Cancer Genome Atlas (TCGA) database, GTEx database and CMap database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.858633/full#supplementary-material

REFERENCES

Arozarena, I., Bischof, H., Gilby, D., Belloni, B., Dummer, R., and Wellbrock, C. (2011). In Melanoma, Beta-Catenin Is a Suppressor of Invasion. *Oncogene* 30, 4531–4543. doi:10.1038/onc.2011.162

Bader, G. D., and Hogue, C. W. (2003). An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC Bioinformatics* 4, 2. doi:10.1186/1471-2105-4-2

Bianchi-Smiraglia, A., Bagati, A., Fink, E. E., Moparthy, S., Wawrzyniak, J. A., Marvin, E. K., et al. (2017). Microphthalmia-associated Transcription Factor Suppresses Invasion by Reducing Intracellular GTP Pools. *Oncogene* 36, 84–96. doi:10.1038/onc.2016.178

Böhme, I., and Bosserhoff, A. K. (2016). Acidic Tumor Microenvironment in Human Melanoma. *Pigment Cel Melanoma Res.* 29, 508–523. doi:10.1111/pcmr.12495

Carreira, S., Goodall, J., Denat, L., Rodriguez, M., Nuciforo, P., Hoek, K. S., et al. (2006). Mitf Regulation of Dia1 Controls Melanoma Proliferation and Invasiveness. *Genes Dev.* 20, 3426–3439. doi:10.1101/gad.406406

Chang, T.-S. (2009). An Updated Review of Tyrosinase Inhibitors. *Int. J. Mol. Sci.* 10, 2440–2475. doi:10.3390/ijms10062440

Cheli, Y., Guiliano, S., Botton, T., Rocchi, S., Hofman, V., Hofman, P., et al. (2011). Mitf Is the Key Molecular Switch between Mouse or Human Melanoma Initiating Cells and Their Differentiated Progeny. *Oncogene* 30, 2307–2318. doi:10.1038/onc.2010.598

Cheli, Y., Giuliano, S., Fenouille, N., Allegra, M., Hofman, V., Hofman, P., et al. (2012). Hypoxia and MITF Control Metastatic Behaviour in Mouse and Human Melanoma Cells. *Oncogene* 31, 2461–2470. doi:10.1038/onc.2011.425

Cui, S., Lou, S., Guo, W., Jian, S., Wu, Y., Liu, X., et al. (2020). Prediction of MiR-21-5p in Promoting the Development of Lung Adenocarcinoma via PDZD2 Regulation. *Med. Sci. Monit.* 26, e923366. doi:10.12659/MSM.923366

Domingues, B., Lopes, J., Soares, P., and Pópulo, H. (2018). Melanoma Treatment in Review. *Immunotargets Ther.* 7, 35–49. doi:10.2147/ITT.S134842

Enzenauer, R. J., Stock, J. G., Enzenauer, R. W., Pope, J., and West, S. G. (1990). Retinal Vasculopathy Associated with Systemic Light Chain Deposition Disease. *Retina* 10, 115–118. doi:10.1097/00006982-199004000-00005

Galinsky, K. J., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N. J., et al. (2016). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* 98, 456–472. doi:10.1016/j.ajhg.2015.12.022

Gao, J., Liang, L., Zhu, Y., Qiu, S., Wang, T., and Zhang, L. (2016). Ligand and Structure-Based Approaches for the Identification of Peptide Deformylase Inhibitors as Antibacterial Drugs. *Int. J. Mol. Sci.* 17, 1141. doi:10.3390/ijms17071141

Gaudreault, F., Morency, L.-P., and Najmanovich, R. J. (2015). NRGsuite: a PyMOL Plugin to Perform Docking Simulations in Real Time Using FlexAID. *Bioinformatics* 31, btv458–3858. doi:10.1093/bioinformatics/btv458

Gharpure, K. M., Lara, O. D., Wen, Y., Pradeep, S., LaFargue, C., Ivan, C., et al. (2018). ADH1B Promotes Mesothelial Clearance and Ovarian Cancer Infiltration. *Oncotarget* 9, 25115–25126. doi:10.18632/oncotarget.25344

He, F., Fang, L., and Yin, Q. (2019). miR-363 Acts as a Tumor Suppressor in Osteosarcoma Cells by Inhibiting PDZD2. *Oncol. Rep.* 41, 2729–2738. doi:10.3892/or.2019.7078

Javelaud, D., Alexaki, V.-I., Pierrat, M.-J., Hoek, K. S., Dennler, S., Van Kempen, L., et al. (2011). GLI2 and M-MITF Transcription Factors Control Exclusive Gene Expression Programs and Inversely Regulate Invasion in Human Melanoma Cells. *Pigment Cel Melanoma Res.* 24, 932–943. doi:10.1111/j.1755-148X.2011.00893.x

Jia, X., and Zhai, T. (2019). Integrated Analysis of Multiple Microarray Studies to Identify Novel Gene Signatures in Non-alcoholic Fatty Liver Disease. *Front. Endocrinol.* 10, 599. doi:10.3389/fendo.2019.00599

Jiang, X., Sun, L., Qiu, J. J., Sun, X., Li, S., Wang, X., et al. (2013). A Novel Application of Furazolidone: Anti-leukemic Activity in Acute Myeloid Leukemia. *PLoS One* 8, e72335. doi:10.1371/journal.pone.0072335

Kaochar, S., Dong, J., Torres, M., Rajapakshe, K., Nikolos, F., Davis, C. M., et al. (2018). ICG-001 Exerts Potent Anticancer Activity against Uveal Melanoma Cells. *Invest. Ophthalmol. Vis. Sci.* 59, 132–143. doi:10.1167/iovs.17-22454

Karamanakos, P. N. (2013). Possible Role for Furazolidone in the Treatment of Glioblastoma Multiforme. *J. BUON* 18, 1097.

Kitlinska, J., Abe, K., Kuo, L., Pons, J., Yu, M., Li, L., et al. (2005). Differential Effects of Neuropeptide Y on the Growth and Vascularization of Neural Crest-Derived Tumors. *Cancer Res.* 65, 1719–1728. doi:10.1158/0008-5472.CAN-04-2192

Körner, M., Waser, B., and Reubi, J. C. (2008). High Expression of Neuropeptide Y1 Receptors in ewing Sarcoma Tumors. *Clin. Cancer Res.* 14, 5043–5049. doi:10.1158/1078-0432.CCR-07-4551

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust Rank Aggregation for Gene List Integration and Meta-Analysis. *Bioinformatics (Oxford, England)* 28, 573–580. doi:10.1093/bioinformatics/btr709

Lamb, J. (2007). The Connectivity Map: a New Tool for Biomedical Research. *Nat. Rev. Cancer* 7, 54–60. doi:10.1038/nrc2044

Lebbé, C., Meyer, N., Mortier, L., Marquez-Rodas, I., Robert, C., Rutkowski, P., et al. (2019). Evaluation of Two Dosing Regimens for Nivolumab in Combination with Ipilimumab in Patients with Advanced Melanoma: Results from the Phase IIIb/IV CheckMate 511 Trial. *J. Clin. Oncol.* 37, 867–875. doi:10.1200/JCO.18.01998

Lee, H., Lee, W., Chang, S., and Lee, G.-Y. (2015). Hesperidin, A Popular Antioxidant Inhibits Melanogenesis via Erk1/2 Mediated MITF Degradation. *Int. J. Mol. Sci.* 16, 18384–18395. doi:10.3390/ijms160818384

Li, Q.-G., He, Y.-H., Wu, H., Yang, C.-P., Pu, S.-Y., Fan, S.-Q., et al. (2017). A Normalization-free and Nonparametric Method Sharpens Large-Scale Transcriptome Analysis and Reveals Common Gene Alteration Patterns in Cancers. *Theranostics* 7, 2888–2899. doi:10.7150/thno.19425

Liu, L., Xu, Q., Cheng, L., Ma, C., Xiao, L., Xu, D., et al. (2015). NPY1R Is a Novel Peripheral Blood Marker Predictive of Metastasis and Prognosis in Breast Cancer Patients. *Oncol. Lett.* 9, 891–896. doi:10.3892/ol.2014.2721

Liu, X., Li, H., Cong, X., Huo, D., Cong, L., and Wu, G. (2020). α-MSH-PE38KDEL Kills Melanoma Cells via Modulating Erk1/2/MITF/TYR Signaling in an MC1R-dependent Manner. *Onco Targets Ther.* 13, 12457–12469. doi:10.2147/OTT.S268554

Liu, Y., Cui, S., Sun, J., Yan, X., and Han, D. (2021). Identification of Potential Biomarkers for Psoriasis by DNA Methylation and Gene Expression Datasets. *Front. Genet.* 12, 722803. doi:10.3389/fgene.2021.722803

Lv, X., Zhao, F., Huo, X., Tang, W., Hu, B., Gong, X., et al. (2016). Neuropeptide Y1 Receptor Inhibits Cell Growth through Inactivating Mitogen-Activated Protein Kinase Signal Pathway in Human Hepatocellular Carcinoma. *Med. Oncol.* 33, 70. doi:10.1007/s12032-016-0785-1

McGill, G. G., Horstmann, M., Widlund, H. R., Du, J., Motyckova, G., Nishimura, E. K., et al. (2002). Bcl2 Regulation by the Melanocyte Master Regulator Mitf Modulates Lineage Survival and Melanoma Cell Viability. *Cell* 109, 707–718. doi:10.1016/s0092-8674(02)00762-6

Pelster, M. S., and Amaria, R. N. (2019). Combined Targeted Therapy and Immunotherapy in Melanoma: a Review of the Impact on the Tumor Microenvironment and Outcomes of Early Clinical Trials. *Ther. Adv. Med. Oncol.* 11, 175883591983082. doi:10.1177/1758835919830826

Polimanti, R., Kranzler, H. R., and Gelernter, J. (2016). Phenome-Wide Association Study for Alcohol and Nicotine Risk Alleles in 26394 Women. *Neuropsychopharmacol* 41, 2688–2696. doi:10.1038/npp.2016.72

Poźniak, J., Nsengimana, J., Laye, J. P., O'Shea, S. J., Diaz, J. M. S., Droop, A. P., et al. (2019). Genetic and Environmental Determinants of Immune Response to Cutaneous Melanoma. *Cancer Res.* 79, 2684–2696. doi:10.1158/0008-5472.CAN-18-2864

Ruscica, M., Dozio, E., Boghossian, S., Bovo, G., Martos Rianõ, V., Motta, M., et al. (2006). Activation of the Y1 Receptor by Neuropeptide Y Regulates the Growth of Prostate Cancer Cells. *Endocrinology* 147, 1466–1473. doi:10.1210/en.2005-0925

Saginala, K., Barsouk, A., Aluru, J. S., Rawla, P., and Barsouk, A. (2021). Epidemiology of Melanoma. *Med. Sci.* 9, 63. doi:10.3390/medsci9040063

Seitz, H. K., and Stickel, F. (2010). Acetaldehyde as an Underestimated Risk Factor for Cancer Development: Role of Genetics in Ethanol Metabolism. *Genes Nutr.* 5, 121–128. doi:10.1007/s12263-009-0154-1

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Shirasugi, I., Kamada, M., Matsui, T., Sakakibara, Y., Liu, M.-C., and Suiko, M. (2010). Sulforaphane Inhibited Melanin Synthesis by Regulating Tyrosinase Gene Expression in B16 Mouse Melanoma Cells. *Biosci. Biotechnol. Biochem.* 74, 579–582. doi:10.1271/bbb.90778

Steingrímsson, E., Copeland, N. G., and Jenkins, N. A. (2004). Melanocytes and the Microphthalmia Transcription Factor Network. *Annu. Rev. Genet.* 38, 365–411. doi:10.1146/annurev.genet.38.072902.092717

Tam, C. W., Cheng, A. S., Ma, R. Y. M., Yao, K.-M., and Shiu, S. Y. W. (2006). Inhibition of Prostate Cancer Cell Growth by Human Secreted PDZ Domain-Containing Protein 2, a Potential Autocrine Prostate Tumor Suppressor. *Endocrinology* 147, 5023–5033. doi:10.1210/en.2006-0207

Tam, C. W., Liu, V. W. S., Leung, W. Y., Yao, K.-M., and Shiu, S. Y. W. (2008). The Autocrine Human Secreted PDZ Domain-Containing Protein 2 (sPDZD2) Induces Senescence or Quiescence of Prostate, Breast and Liver Cancer Cells via Transcriptional Activation of P53. *Cancer Lett.* 271, 64–80. doi:10.1016/j.canlet.2008.05.047

Tang, X., Liu, D., Shishodia, S., Ozburn, N., Behrens, C., Lee, J. J., et al. (2006). Nuclear Factor-κB (Nf-κB) Is Frequently Expressed in Lung Cancer and Preneoplastic Lesions. *Cancer* 107, 2637–2646. doi:10.1002/cncr.22315

Wawrzyniak, J. A., Bianchi-Smiraglia, A., Bshara, W., Mannava, S., Ackroyd, J., Bagati, A., et al. (2013). A Purine Nucleotide Biosynthesis Enzyme Guanosine Monophosphate Reductase Is a Suppressor of Melanoma Invasion. *Cel Rep.* 5, 493–507. doi:10.1016/j.celrep.2013.09.015

Wellbrock, C., Rana, S., Paterson, H., Pickersgill, H., Brummelkamp, T., and Marais, R. (2008). Oncogenic BRAF Regulates Melanoma Proliferation through the Lineage Specific Factor MITF. *PLoS One* 3, e2734. doi:10.1371/journal.pone.0002734

Yeung, M. L., Tam, T. S. M., Tsang, A. C. C., and Yao, K. M. (2003). Proteolytic Cleavage of PDZD2 Generates a Secreted Peptide Containing Two PDZ Domains. *EMBO Rep.* 4, 412–418. doi:10.1038/sj.embor.embor804

Yu, J. G., Ji, C. H., and Shi, M. H. (2020). The Anti-infection Drug Furazolidone Inhibits NF-κB Signaling and Induces Cell Apoptosis in Small Cell Lung Cancer. *Kaohsiung J. Med. Sci.* 36, 998–1003. doi:10.1002/kjm2.12281

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Liu, Sun, Han, Cui and Yan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 25 August 2022
DOI 10.3389/fcell.2022.886989

Check for updates

OPEN ACCESS

EDITED BY
Liang Cheng,
Harbin Medical University, China

REVIEWED BY
Kunju Zhu,
Jinan University, China
Dongsheng Guo,
Huazhong University of Science and Technology, China

*CORRESPONDENCE
Liming Tan,
tanliming838@163.com
Hao Yuan,
yuanhao666@hunnu.edu.cn
Wen Yin,
silver-88@163.com

SPECIALTY SECTION
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

RECEIVED 01 March 2022
ACCEPTED 03 August 2022
PUBLISHED 25 August 2022

CITATION
Tang G, Tan L, Yuan H and Yin W (2022), Glycosylation modification patterns reveal distinct tumor metabolism and immune microenvironment landscape in lower-grade gliomas.
Front. Cell Dev. Biol. 10:886989.
doi: 10.3389/fcell.2022.886989

COPYRIGHT
© 2022 Tang, Tan, Yuan and Yin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Glycosylation modification patterns reveal distinct tumor metabolism and immune microenvironment landscape in lower-grade gliomas

Guihua Tang[1], Liming Tan[1]*, Hao Yuan[1]* and Wen Yin[2]*

[1]Department of Clinical Laboratory, Hunan Provincial People's Hospital, The First Affiliated Hospital of Hunan Normal University, The College of Clinical Medicine of Human Normal University, Changsha, China, [2]Department of Neurosurgery, National Clinical Research Center for Geriatric Disorders, Xiangya Hospital of Central South University, Changsha, China

Glycosylation alterations, a key driver throughout tumorigenesis and tumor progression, could regulate the microenvironment and immune response as well as lead to harmful metabolism and cell signaling. In this study, we first comprehensively evaluated the glycosylation modification patterns of LGGs based on glycosyltransferase family genes and systematically integrated these modification patterns with tumor metabolism and immune microenvironment characteristics. Glycosylation score was also developed to quantify glycosylation modification patterns of individuals. As a result, two glycosylation modification patterns were identified, with distinct prognosis, metabolism, and immune microenvironment features. The glycosylation subtype A and cluster A were characterized by higher carbohydrates and amino acid metabolism activity, higher levels of infiltrating cells, and poor prognosis, whereas an opposite modification pattern was observed in glycosylation subtype B and cluster B. In addition, a high glycosylation score is closer to a microenvironment characterized by chronic inflammation, immunosuppression, and tumor promotion. Following analysis and validation, the glycosylation score was a reliable and independent prognostic index. More importantly, the glycosylation score influenced the response to immunotherapy, chemotherapy, or targeted therapy, which provided a novel insight into promoting personalized therapy in the future and may contribute to developing novel therapeutic drugs or exploring promising drug combination therapy strategies.

KEYWORDS

glycosylation, lower-grade gliomas, prognosis, tumor metabolism, immune microenvironment landscape

Introduction

Diffuse gliomas, a heterogeneous group of the supporting glial cell neoplasms, are the most prevalent primary malignancies of the central nervous system (CNS). The World Health Organization (WHO) classified gliomas into grades I–IV (Yu H. et al., 2021). Grades II and III gliomas are often defined as lower-grade gliomas (LGGs) because these tumors represent a completely different entity from glioblastoma (grade IV). Up to 80% of LGGs harbor isocitrate dehydrogenase (IDH) mutations, but only approximately 5% of glioblastomas are IDH-mutant (Youssef and Miller, 2020). More importantly, this molecular finding conveys a relatively optimistic and favorable prognosis (Schiff et al., 2019). Complete surgical resection followed by chemoradiotherapy is the currently favored therapy for LGGs (Schiff et al., 2019; Yu H. et al., 2021). However, LGGs cannot be cured completely, and more than 70% of patients will inevitably experience tumor recurrence or malignant progression within 10 years (Gittleman et al., 2020), owing to its high invasiveness and aggressiveness. Thus, reducing tumor recurrence and delaying tumor progression are the most issues and challenges. Even after active efforts of applying multiple treatment modalities, however, the malignant behaviors and clinical outcomes of these tumors vary greatly from person to person (Aoki et al., 2018; Berzero et al., 2021). Due to the high heterogeneity of LGGs, it was insufficient for the existing molecular patterns of LGGs to fully explain clinical outcomes and predict prognosis. Therefore, more comprehensive research to better understand potential molecular mechanisms in LGGs oncogenesis and progression is needed, which will contribute to better patient stratification and further making clinical decisions to improve patient management.

Glycosylation, a recognized hallmark of cancer(Thomas et al., 2021), is defined as an enzymatic process that could accelerate the production of glycosidic linkages between saccharides and other saccharides, lipids, or proteins (Pinho and Reis, 2015). Emerging evidence demonstrated that altered glycans serve as a key driver throughout tumorigenesis and tumor progression; hence, changes in cellular glycosylation, as a key component of malignant progression, have recently attracted more and more attention (Stowell et al., 2015). Glycosylation is involved in almost all important biological processes, such as protein quality control, protein clearance, intracellular trafficking, cell–cell interaction, cell–matrix adhesion, and various signal transduction cascades (Thomas et al., 2021). In addition, abnormal and modified glycosylation alterations caused by cellular and metabolic changes will not only significantly affect the overall charge and conformation of glycoproteins, thereby changing its biological activity and directly affecting tissue cell growth and survival (Stowell et al., 2015). Also, these changes of glycosylation also cause abnormal expression of membrane-localized glycans, which will trigger cellular malignant transformation and mediate cancer cell proliferation, survival, and metastasis (Marsico et al., 2018; Chandler et al., 2019; Thomas et al., 2021). Accumulating findings suggest that abnormal glycan profiles could regulate the microenvironment and immune response (Mereiter et al., 2019) and even hinder an effective immune response (Polmear and Good-Jacobson, 2021). In addition, the metabolic state (Campbell and Wellen, 2018) or metabolic reprogramming (Carvalho-Cruz et al., 2018) of the tumor cell can lead to aberrant glycosylation intra- and extra-cellular. On the contrary, aberrant glycosylation can also lead to harmful metabolism and cell signaling. However, these mechanisms are not fully understood. Therefore, evaluating the association between glycosylation modification patterns and immune landscape, along with metabolism alterations will contribute to strengthening our understanding of the vital role of glycosylation in tumor biology, and is expected to provide valuable strategies for improving personalized management of patients.

In the present study, we integrated the transcriptomic information of LGGs samples to identify the glycosylation patterns associated with distinct prognosis and immune and metabolic characteristics. Moreover, we first proposed a scoring system to quantify the glycosylation patterns of individual LGGs patients, and it can provide clinical decision-makers with a novel perspective to better stratify patients, predict treatment response, and improve individualized treatment strategies.

Materials and methods

Data source and preprocessing

The overview of our study design is shown in Supplementary Figure S1. The LGGs gene expression data and matched clinical annotations were retrieved from TCGA and Chinese Glioma Genome Atlas (CGGA) databases. Only patients with survival time greater than 30 days were involved in this study. In total, four cohorts were enrolled: TCGA-LGG (481 samples), CGGA-mRNAseq_693 (420 samples), CGGA-mRNAseq_325 (170 samples), and CGGA-mRNA-array_301 (159 samples). The TCGA-LGG cohort data by log2(x+1) transformed were downloaded from the UCSC Xena website (https://xena.ucsc.edu/), and the other three cohorts were downloaded from the CGGA official website (http://www.cgga.org.cn/index.jsp). At the same time, the mRNA sequencing (mRNA-seq) data in CGGA was also log2(x+1) normalized.

Glycosylation-based consensus clustering analysis

We collected the genes of the main glycosyltransferase families involved in human species-specific glycosylation

mechanisms (Mohamed Abd-El-Halim et al., 2021), including 221 genes retrieved from the GlycoGene Database (GGDB, http://riodb.ibase.aist.go.jp/rcmg/ggdb) and 213 genes extracted from the Hugo Gene Nomenclature Committee (HGNC, https://www.genenames.org/data/genegroup/#!/group/424). Eventually, a total of 143 overlap genes were selected in our analysis. Consensus clustering (Zhang et al., 2020) was performed to identify distinct glycosylation subtypes according to the expression of these 143 genes in the TCGA cohort. The ConsensusClusterPlus package was used and 1000 times repetitions were conducted to ensure the stability of clustering results. In addition, we applied the T-distributed stochastic neighbor embedding (t-SNE) to verify the reliability of subtype assignments based on the expression of the aforementioned genes (Bai et al., 2021).

Gene set variation analysis

To explore the variations in biological processes between glycosylation modification patterns, we selected the GSVA enrichment analysis to accomplish this purpose by using the "GSVA" package. GSVA is usually applied to estimate the variation of pathway and biological process activity in a nonparametric and unsupervised manner (Hänzelmann et al., 2013). The gene sets of "c2.cp.kegg.v7.4.symbols", GOBP, and metabolism-associated pathway signatures derived from KEGG and REACTOME were both downloaded from the molecular signature database (MsigDB, http://www.gsea-msigdb.org/gsea/msigdb/search.jsp). Adjusted $p < 0.05$ was considered statistically significant.

Estimation of tumor microenvironment characteristics

The single sample gene set enrichment analysis (ssGSEA) was applied to estimate the enrichment scores of 13 immune functions and 16 immune cells. The scores were normalized to specific distribution from 0 to 1 and could quantify the relative abundance of estimated items. The marker gene sets were curated from a current study (Bindea et al., 2013), and the "GSVA" package was introduced to conduct the aforementioned analysis. Moreover, to comprehensively estimate the infiltration level of immune cells and stromal cells in the TME, Immune Cell Abundance Identifier (ImmuCellAI) (Miao et al., 2020), Tumor Immune Estimation Resource (TIMER) (Yu L. et al., 2021), and microenvironment cell population counter (MCP-counter) (Ye et al., 2021) algorithms were also utilized in our study. Among them, the former is calculated in the web tool ImmuCellAI (http://bioinfo.life.hust.edu.cn/ImmuCellAI#!/), while the latter two algorithms are implemented on the Sangerbox platform (http://vip.sangerbox.com/home.html).

Identification of glycosylation-related hub genes

To identify glycosylation-related genes (GRGs), we performed differential expression analysis between glycosylation distinct subtypes. The "limma" package was used to determine differentially expressed genes (DEGs), and the significance filtering criteria were set as $|\log2 \text{ (fold change)}| > 1$ as well as adjusted p value less than 0.0001. In addition, to ensure the reliability of GRGs, weighted gene co-expression network analysis (WGCNA) was also introduced to establish co-expressed gene modules strongly related to glycosylation. Based on the expression of 8619 prognostic genes (univariate Cox analysis $p < 0.01$), a gene co-expression network was established by using the "WGCNA" R package (Langfelder and Horvath, 2008). The Pearson correlation was calculated according to the expression of these filtered genes, and an optimal soft-thresholding power $\beta = 7$ was selected to build an unsigned weighted adjacency matrix followed by a topological overlap matrix (TOM) conversion. Finally, the average linkage hierarchical was clustered with the parameter height = 0.25, and the criterion for gene module identification was set as a cutting height of 0.9 along with the module genes a minimum number of 20. The modules of $|$ correlation coefficient $| > 0.5$ were considered as closely related to glycosylation and selected for a subsequent analysis. The intersections of DEGs and module genes screened by WGCNA were defined as GRGs.

Generation and validation of glycosylation gene signature and glycosylation score

To quantify the glycosylation modification patterns of individual patients, we developed a glycosylation scoring scheme to assess the glycosylation modification pattern of individuals with LGGs, which was glycosylation gene signature and the glycosylation score. Primarily, we divided the GRGs into two groups based on their hazard ratio (HR) of the univariate Cox regression model. Genes with HR > 1 and those with HR < 1 were defined as the glycosylation gene signatures A and B, respectively. Next, we conducted GSVA for dimension reduction of the signatures A and B, the GSVA value of which were separately termed as value A and value B. Finally, both value A and value B were selected to calculate the glycosylation score as follows: glycosylation score = value A- value B. Patients with LGGs were classified into high and low glycosylation score groups based on the median cutoff value, and scores between the glycosylation subtypes or clusters were assessed by the Wilcoxon test.

Moreover, the prognostic value of the glycosylation scores was validated in three cohorts of the CGGA database by using the same signatures and the cutoff value.

Correlation between the glycosylation score and other related biological processes

To explore the association between the glycosylation score and some related biological pathways. The gene sets related to stromal activation were extracted from the supplementary material of a current study (Mariathasan et al., 2018), including epithelial–mesenchymal transition (EMT) signatures (EMT1, EMT2, and EMT3) and pan-fibroblast TGFb response signature (Pan-FTBRS). Moreover, immune checkpoint- and immune activation–related genes were retrieved from the publication of Zhang et al. (2020). In addition, immunosuppressive factors and immunosuppressive cell recruitment factors were obtained from another study (Su et al., 2019). The differences in biological process between distinct glycosylation score groups were assessed *via* the Wilcoxon test.

Small molecule drugs screening and drug sensitivity prediction

Differentially expressed genes (DEGs) between the high and low glycosylation score groups were set as adjusted p-value < 0.001 and | log2 (fold change) | > 1, which were determined by using the "limma" package. The DEGs were visualized into volcano plots and submitted to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses by the "clusterProfiler" package. Furthermore, both of the top 500 up- and downregulated genes were uploaded into the Connectivity Map 02 (CMap, https://portals.broadinstitute.org/cmap/) database, and a CMap mode-of-action (MoA) analysis was applied to discover possible small molecular drugs for LGGs and mechanisms of action.

We evaluate the predictive capacity of glycosylation score in responding to chemotherapeutic agents and immunotherapy. The 50% inhibiting concentration (IC50) value of 138 drugs of each patient was quantified using the pRRophetic algorithm (Tan et al., 2021), and the differences between high and low glycosylation score subgroups compared by using the Wilcoxon test. In addition, the response to immunotherapy was predicted by using immunophenoscore (IPS) (Gui et al., 2021; Wu et al., 2021), which was obtained without bias based on determining components of immunogenicity including effector cells, immunosuppressor cells, major histocompatibility complex (MHC) molecules, and immune modulators. The statistical difference of the IPS value was tested by the Wilcoxon test.

Statistical analysis

R software (version 4.1.0) was applied to conduct statistical analysis. We adopted the Kaplan–Meier (K-M) analysis and log-rank tests to perform the survival analysis with the "survival" R package. The prognostic value of glycosylation score was evaluated by a multivariate Cox regression model and a time-dependent receiver operating characteristic (ROC) analysis using R packages "survival' and "timeROC". Comparisons between glycosylation subgroups, clusters, and score groups were presented *via* the Wilcoxon test.

Results

Glycosylation-related molecular subtypes with distinct survival, metabolism, and immune microenvironment features in LGGs

Based on the expression of 143 genes involved in glycosyltransferase families, 481 patients from the TCGA database were obviously divided into two groups when $k = 2$ (Figure 1A) by the consensus clustering analysis, which was termed as glycosylation subtype A and subtype B, respectively. The t-SNE algorithm confirmed that glycosylation subtype assignments are reliable and can be completely distinguished (Figure 1B). The survival analysis revealed that patients of subtype A showed a significantly shortened survival time compared with patients of subtype B (Figure 1C, log-rank test).

To explore the biological molecular changes between distinct glycosylation subtypes, the KEGG gene set was used to performed GSVA enrichment analysis. As shown in Figure 1D, subtype A presented enrichment pathways prominently related to carcinogenic activation and immune regulation pathways such as nucleotide excision repair, homologous recombination, antigen processing and presentation, and NOD-like receptor signaling pathway, while subtype B is markedly enriched in metabolism-related processes, including steroid biosynthesis, oxidative phosphorylation, citrate cycle TCA cycle, and so on. Next, we further compare the metabolism and immune microenvironment features affected by glycosylation modification patterns between the subtypes. The results of ssGSEA showed a significant up-regulation in immune infiltrating cells (Figure 1E) and immune function (Figure 1F) of subtype A. Furthermore, the GSVA results revealed significant differences in metabolic processes (Figure 2A), with subtype B mainly active in lipid metabolism (Figure 2B) and other metabolism (Figure 2C), while subtype A was mainly active in carbohydrates metabolism (Figure 2D) and amino acid metabolism (Figure 2E).

Survival, metabolism, and immune microenvironment characteristics in glycosylation clusters for LGGs

To further validate the glycosylation modification patterns and the potential biological behavior, GRGs identified by

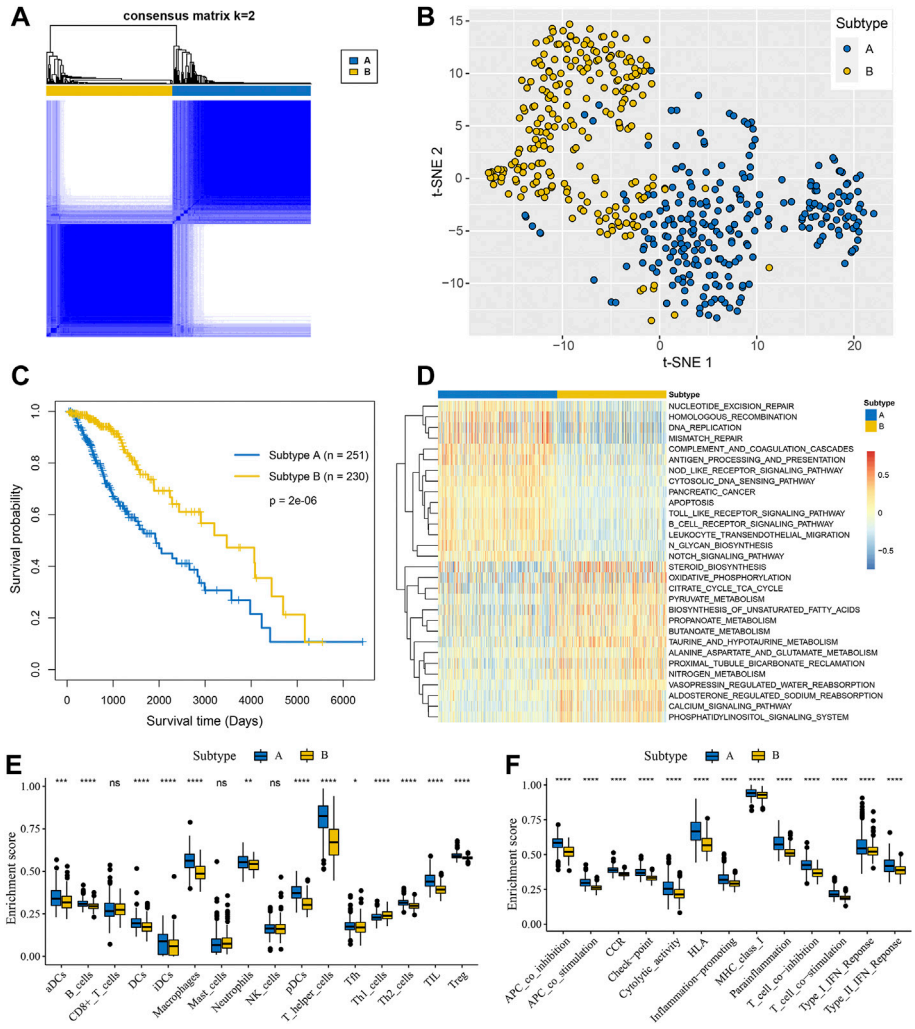


FIGURE 1
Glycosylation-related molecular subtypes with distinct survival and immune microenvironment features in LGGs. **(A)** Consensus clustering matrix of 481 samples from the TCGA dataset for $k = 2$. **(B)** t-SNE of the expression profiles of glycosyltransferase family genes in the TCGA cohort confirmed the subtypes. **(C)** K-M analysis for the subtypes of LGGs patients. **(D)** GSVA enrichment analysis displayed the differences of KEGG pathways between the two subtypes. Red represented higher pathway activity and blue represented lower pathway activity. The boxplots visualized the ssGSEA enrichment scores of immune infiltrating cells **(E)** and immune functions **(F)** between the subtypes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; ns, not significant.

differentially expressed analysis and WGCNA were further inputted into the subsequent investigation. Primarily, 1945 DEGs between glycosylation subtypes including 788 up-regulated genes and 1157 down-regulated genes were visualized in the volcano plot (Figure 3A), and the top 100 ones are displayed in Figure 3B. Then 18 gene modules (Figure 3C) were recognized by WGCNA, and six modules (light yellow, brown, grey60, dark orange, dark grey, and tan) were closely related to glycosylation (Figure 3D). Eventually, 971 interactions (Figure 3E), considered GRGs (Supplementary Table S1), were uploaded into the consensus clustering analysis. Consistent with the subtype grouping of glycosylation modification patterns, the TCGA-LGGs cohort was divided into two distinct clusters, which were defined as glycosylation clusters A and B (Figure 3F). Also, the t-SNE algorithm data confirmed the rationality of the cluster assignments (Figure 3G). As expected, we found that patients in cluster A exhibited significantly shortened the survival time (Figure 3H; log-rank test). This demonstrated that two distinct glycosylation modification patterns did exist and work in LGGs. We further probed into the metabolism and immune microenvironment characteristics of the glycosylation clusters. In short, glycosylation cluster A showed higher scores of immune infiltrating cells (Figure 4A) and immune function (Figure 4B), except for $CD8^+$ T cells, mast cells, and natural killer (NK) cells.

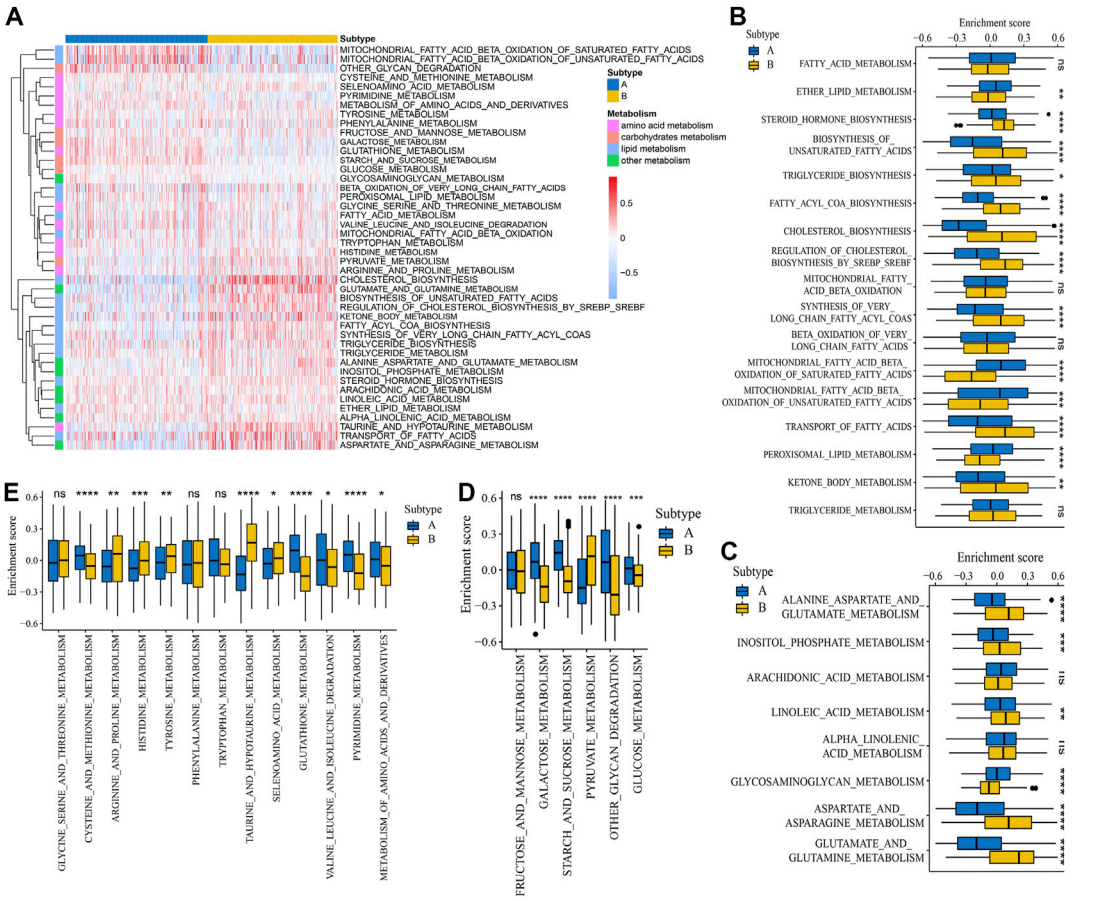


FIGURE 2
Metabolism features of the glycosylation subtypes. **(A)** Heat map depicted the activation states of metabolism-associated process between subtypes based on GSVA enrichment scores. Red represents activated processes and blue represents inhibited processes. Differences of GSVA enrichment scores of the metabolism-associated process including lipid metabolism **(B)**, other metabolism **(C)**, carbohydrates metabolism **(D)**, and amino acid metabolism **(E)** between the subtypes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; ns, not significant.

Moreover, metabolic characteristics also showed significant differences (Figures 4B–E) between glycosylation clusters. Excitedly, these results were much similar to that of between glycosylation subtypes. Collectively, the concordance between prognostic, metabolic, and immune microenvironment characteristics in glycosylation clusters and subtypes confirmed again that glycosylation modification played a vital role in metabolic reprogramming and immune regulation, and this classification was steady and reliable.

Development of the glycosylation scoring system as an independent prognostic factor for LGGs

Although our findings revealed the role of glycosylation modification in prognosis, metabolic reprogramming, and immune modulation, these analyses could not accurately quantify the glycosylation modification patterns of individual tumors. Therefore, to enable the quantification of glycosylation patterns in individual LGGs patients, we developed a glycosylation scoring system termed as the glycosylation score. Based on the median glycosylation score of 0.009856, patients were classified into the high and low glycosylation score subgroups. Figure 5A showed the distribution of clinical traits including age, gender, grade, overall survival, and survival status between high and low glycosylation score subgroups, while Figure 5B further displayed the grade distribution, glycosylation subtypes, and clusters assignment of patients from distinct glycosylation score subgroups, which are summarized in Supplementary Table S2. The Wilcoxon test demonstrated that patients in subtype A had a higher glycosylation score compared to subtype B (Figure 5C); meanwhile, cluster A

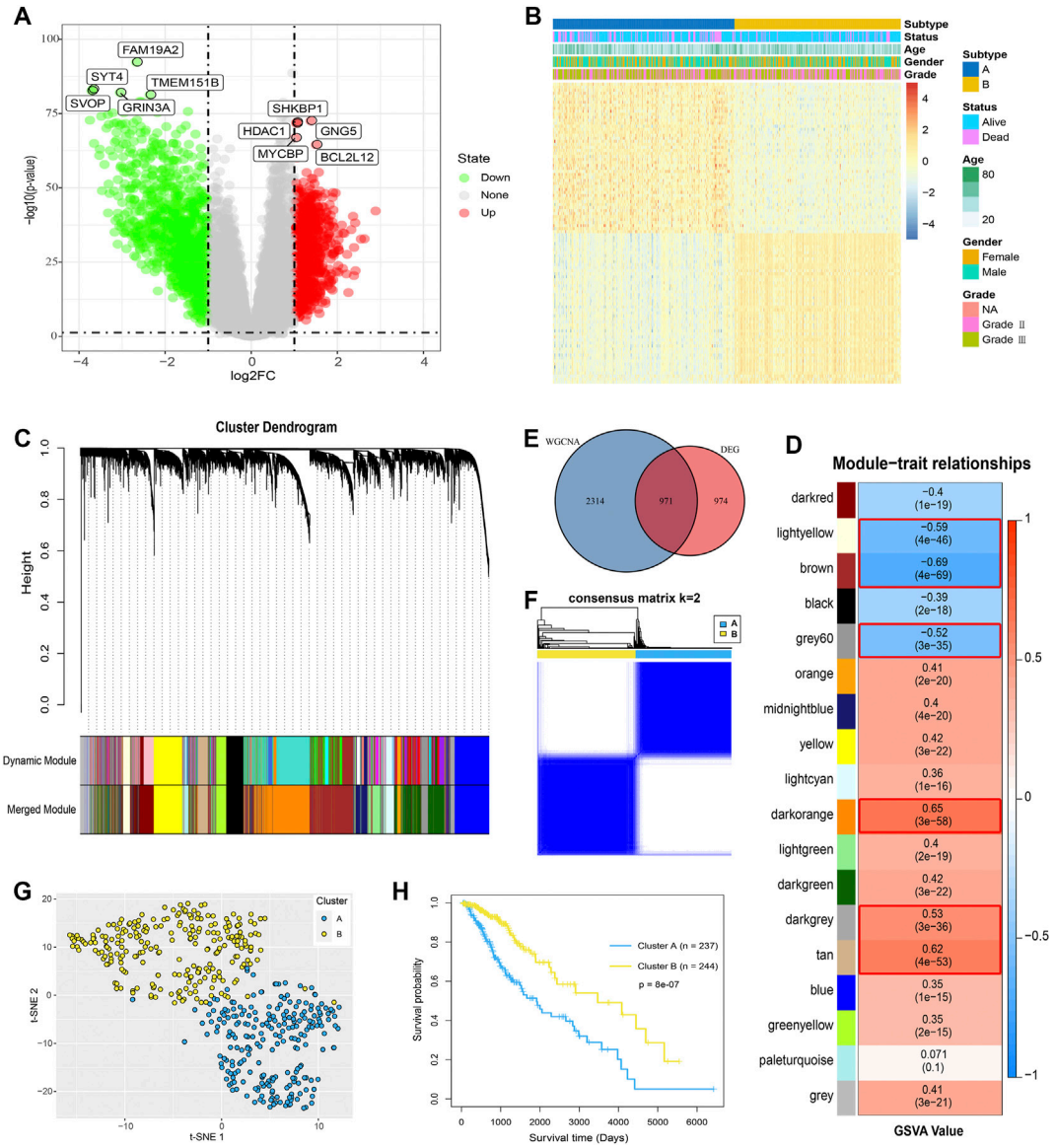


FIGURE 3
Identification of GRGs and glycosylation clusters. **(A)** Volcano map showed the DEGs between the glycosylation subtypes. Red represented up-regulated genes and green represented down-regulated genes. **(B)** Heat map showed clinical features and top 100 DEGs between glycosylation subtypes. **(C)** Gene co-expression network and gene modules were identified by WGCNA. **(D)** In total, 18 gene modules were recognized, and six modules marked with red frames, with | correlation coefficient | > 0.5, were considered as closely related to glycosylation. **(E)** In total, 971 overlapping genes were determined by DEGs and WGCNA were defined as GRGs. **(F)** Consensus clustering matrix for k = 2 and **(G)** t-SNE identified glycosylation clusters, which were both based on the expression profiles of GRGs in the TCGA cohort. **(H)** K-M analysis for the clusters of LGGs patients.

showed a higher glycosylation score than that of cluster B (Figure 5D). This indicated that a high glycosylation score could be closely related to poor prognosis, whereas a low glycosylation score could be closely related to a favorable prognosis. Of course, this hypothesis was confirmed by a subsequent survival analysis (Figure 5E) based on glycosylation scores. We analyzed the prognostic values of glycosylation scores in depth. The multivariate cox regression analysis demonstrated that the glycosylation score was an independent risk factor with its HR = 2.010 for LGGs (Figure 5F). Meanwhile, the time-dependent area under the curve (AUC) suggested that the glycosylation score had a robust and reliable value in predicting the prognosis for LGGs in the TCGA cohort (Figure 5G).

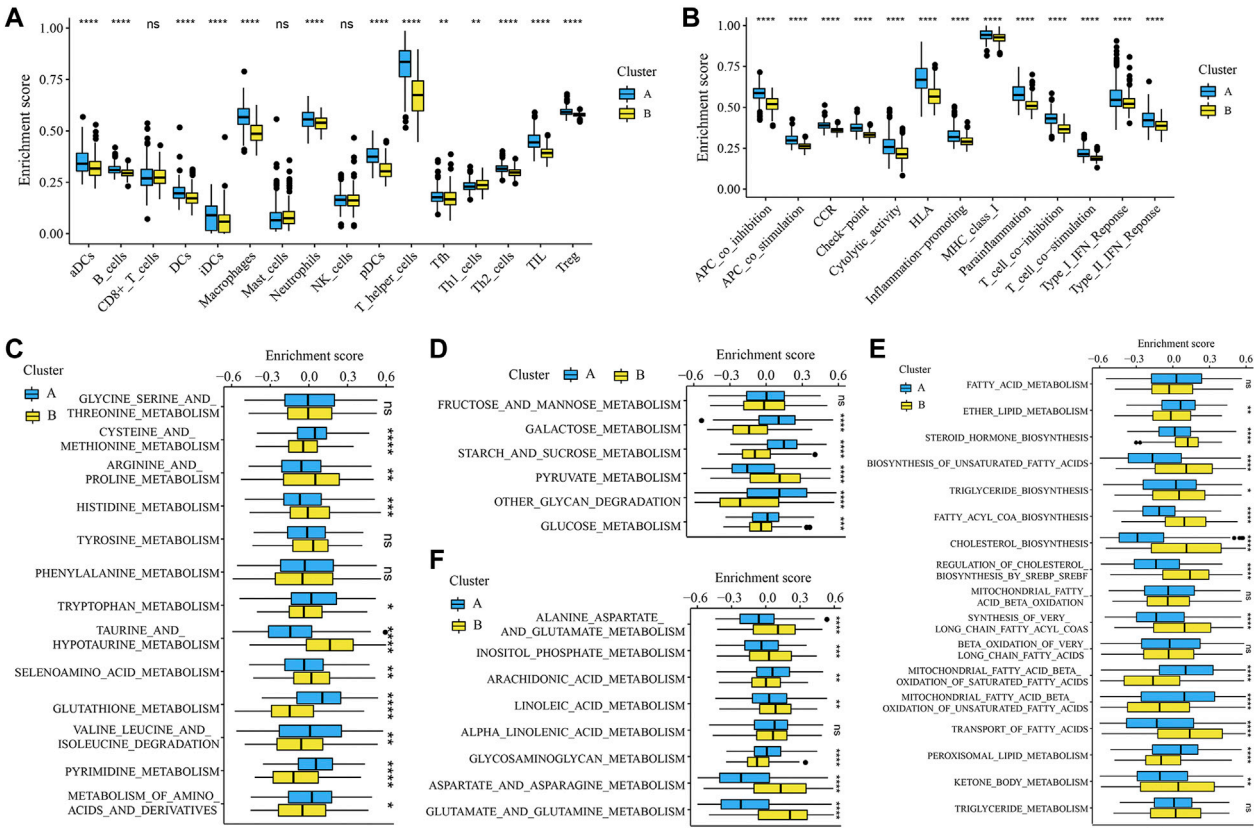


FIGURE 4
Immune microenvironment and metabolism features of the glycosylation clusters. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; ns, not significant. The boxplots showed the enrichment scores of **(A)** immune infiltrations and **(B)** immune functions differed in the clusters. Boxplots depicted the differences in enrichment scores of metabolism-associated processes between glycosylation clusters, which included **(C)** amino acid metabolism, **(D)** carbohydrates metabolism, **(E)** lipid metabolism, and **(F)** other metabolism.

We further verified the predictive effectiveness of glycosylation score in three independent external cohorts according to the same glycosylation gene signatures and cutoff score. Consistent with findings in the TCGA cohort, patients with a high glycosylation score showed a significantly shortened survival than those with a low glycosylation score in validation cohorts including CGGA-mRNAseq_693 (Supplementary Figure S2A), CGGA-mRNAseq_325 (Supplementary Figure S2B), and CGGA-mRNA-array_301 (Supplementary Figure S2C). Similarly, superior time-dependent AUCs were observed in those cohorts (Supplementary Figures S2D–F).

Glycosylation score is associated with metabolism and immune microenvironment features of LGGs

The correlations between the glycosylation score and metabolism as well as immune microenvironment features were further explored in this study. Consistent with findings observed in distinct glycosylation patterns, the glycosylation score was strongly linked to metabolism activity (Supplementary Figure S3A) and immune microenvironment infiltrations (Figure 6). The results showed that the high glycosylation score subgroup was associated with enhanced activity in most amino acid metabolism (Supplementary Figure S3B) and carbohydrates metabolism (Supplementary Figure S3C); on the contrary, the low glycosylation score subgroup was associated with enhanced activity in most lipid metabolism (Supplementary Figure S3D) and other metabolism (Supplementary Figure S3E). Meanwhile, the distribution of infiltrating cells estimated by ImmuCellAI, TIMER, MCP-counter, and ssGSEA algorithms between the high and low glycosylation subgroups were also investigated. Our results (Figure 6) suggested that the high glycosylation score subgroup had a significantly higher infiltrating level than the low glycosylation score subgroup. Most of the infiltrating immune and stromal cells increased in the high glycosylation score subgroup, whereas T helper 1 (Th1) and $CD8^+$ naïve cells decreased in this subgroup. Furthermore, there were higher scores

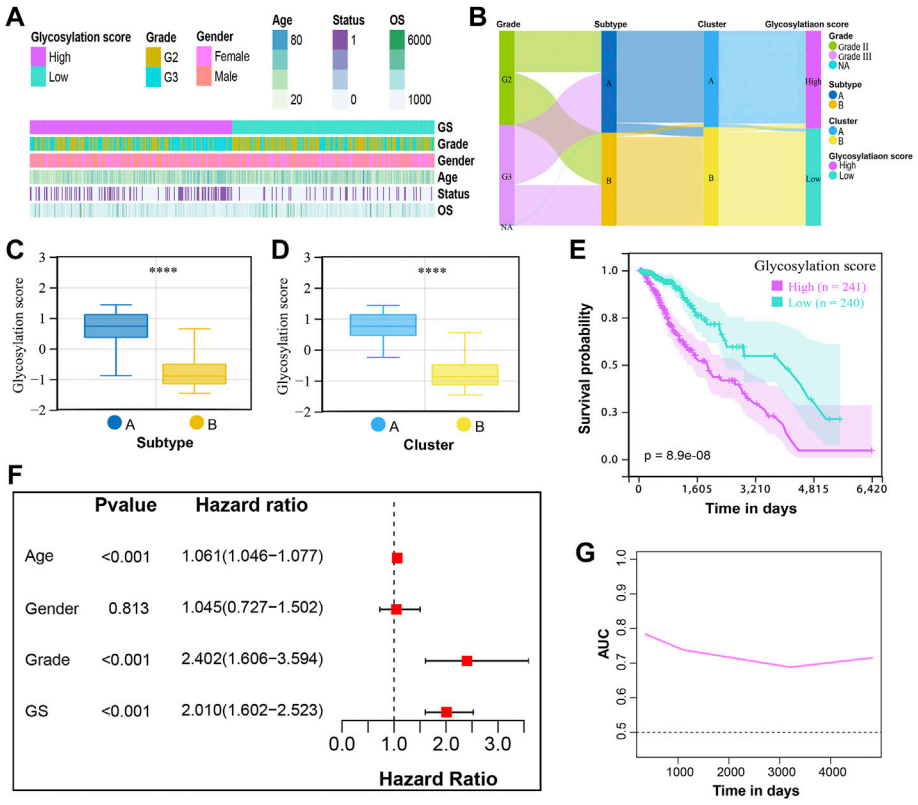


FIGURE 5
Development of the glycosylation scoring system and its prognostic values for LGGs. **(A)** Clinical features in the high and low glycosylation score subgroups. **(B)** Alluvial diagram of glycosylation score in groups with different grades, subtypes, and clusters. The boxplot showed the distribution of glycosylation score in the different **(C)** subtypes and **(D)** clusters. The *p*-values were both less than 0.0001. **(E)** K-M curves for patients with low and high glycosylation scores. **(F)** Multivariate Cox regression analysis for the glycosylation score and clinical variables. GS, glycosylation score. **(G)** Time-dependent AUC value of the glycosylation score demonstrated it as a robust and reliable prognostic index.

of 13 immune functions in the high glycosylation score subgroup (Figure 7A), compared to the low glycosylation score subgroup. Taken together, the glycosylation score presented a close association with metabolism and immune microenvironment features of LGGs.

Correlation of cytokine, chemokine, and other related biological processes in glycosylation score subgroups

To explore potential mechanisms correlated with different glycosylation score subgroups in depth, we compared the expression of chemokine and cytokine between subgroups. These cytokine and chemokine retrieved from published literature, of which, includes the transcripts correlated with immune activation (Figure 7B), immunosuppressive cell recruitment factors (Figure 7C), immunosuppressive factors (Figure 7D), and immune checkpoints (Figure 7E). We observed that most of these genes were expressed higher in the high glycosylation score subgroups, revealing the coexistence of a chronic inflammation state and a suppressed immune microenvironment in this subgroup. To better depict the role of glycosylation scores, we evaluated other related biological processes, including immune activation, stromal activation, and cancer promotion, in patients with LGGs. Consistent with the aforementioned observations, patients in the high glycosylation score subgroup exhibited a much higher enrichment score of immune activation–related biological processes (Figure 7F). On the other hand, higher stromal activation and cancer promotion scores were also enriched in this subgroup. Collectively, the glycosylation score did play a non-negligible role in affecting LGG progression.

Potential small molecule compounds and drug sensitivity prediction

To explore potential small therapeutic drugs for patients with high glycosylation scores, 2090 DEGs, including 923 up-

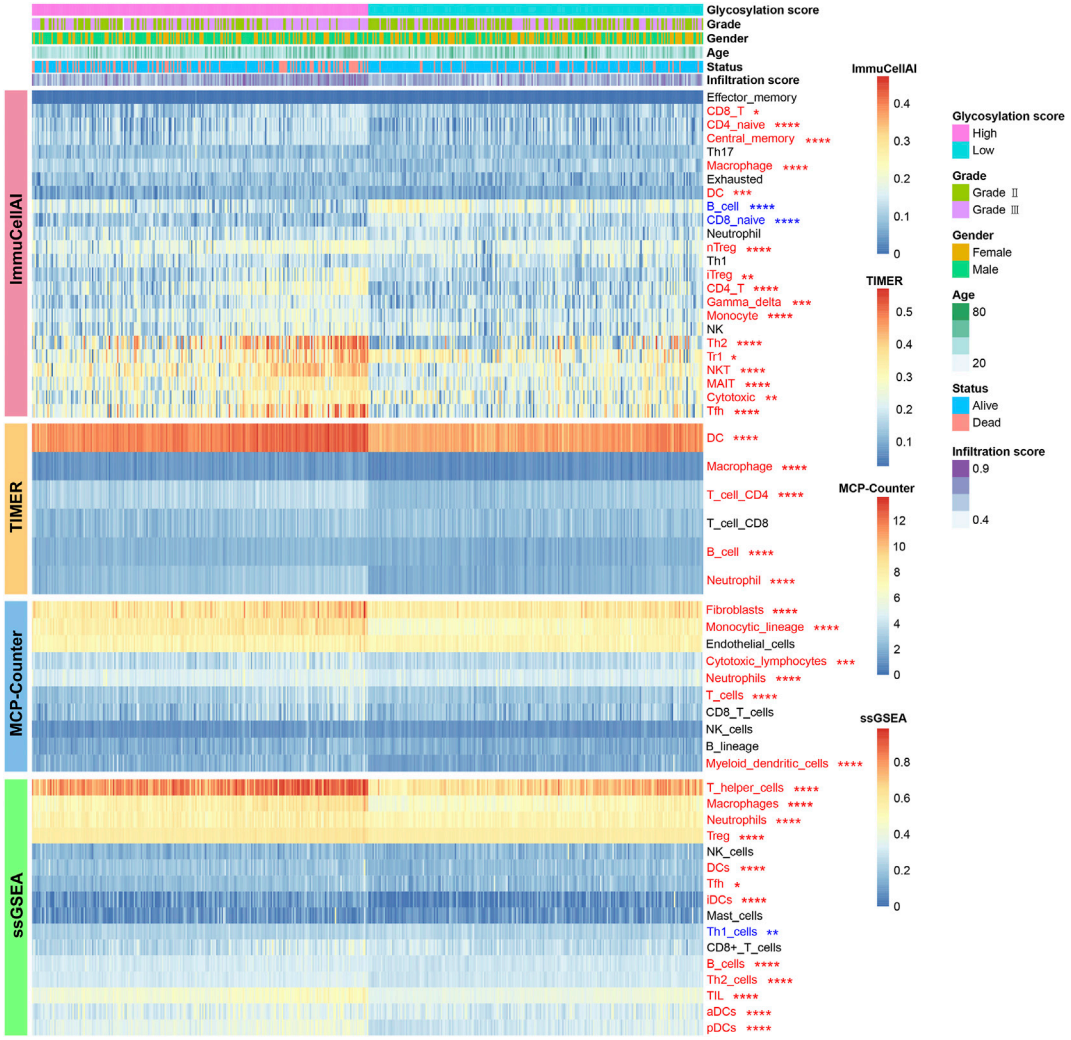


FIGURE 6
Landscape of infiltrating cells in the high and low glycosylation score subgroups. The heat map depicted the immune and stromal cell infiltration scores. Red represents cells infiltrated higher in the high glycosylation score subgroup and blue represents cells infiltrated lower in the high glycosylation score subgroup. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$.

regulated genes and 1167 down-regulated genes, between the high and low glycosylation score subgroups were identified (Figure 8A). The GO analysis results illustrated that the DEGs are primarily involved in modulation of chemical synaptic transmission, regulation of trans-synaptic signaling, synaptic membrane, ion channel complex, gated channel activity, and ion channel activity (Figure 8B). Meantime, these DEGs were significantly enriched in the cholinergic synapse, GABAergic synapse, glutamatergic synapse, neuroactive ligand–receptor interaction, serotonergic synapse, and synaptic vesicle cycle signaling pathways (Figure 8C). Based on the CMap MoA analysis, a total of 14 potential small therapeutic drugs (such as spiradoline, propofol, and dextromethorphan) and 13 drug mechanisms (such as opioid receptor agonist, GABA receptor agonist, and glutamate receptor antagonist) were identified (Figure 8D).

To explore the potential role of glycosylation score in the drug sensitivity prediction, we compared the IC50 value of 138 drugs in TCGA-LGGs patients by using the pRRophetic algorithm; of which, the results showed that there were 29 drugs (such as gefitinib, nilotinib, and axitinib) in which the estimated IC50 value were significantly lower in the low glycosylation score subgroup compared to the high glycosylation score subgroup (Supplementary Figure S4A), indicating that patients in the low glycosylation score subgroup might respond better to these chemotherapy drugs. Similarly, patients in the low glycosylation score subgroup had a higher IPS score, suggesting that patients in this subgroup might be more sensitive to immunotherapy (Supplementary Figure S4B).

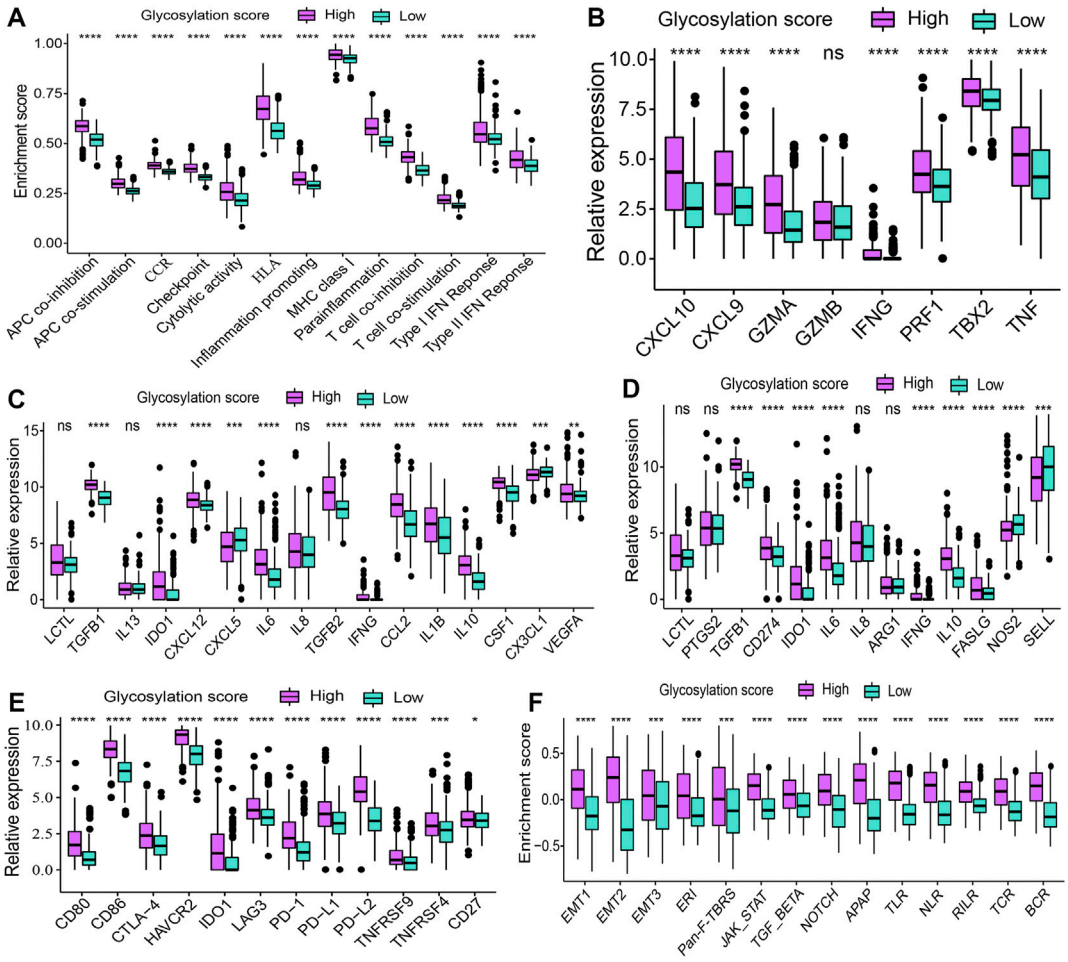


FIGURE 7
Cytokine, chemokine, and other related biological processes in glycosylation score subgroups. (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; ns, not significant) **(A)** Difference in the enrichment score of immune functions between glycosylation score subgroups. **(B)** Difference in the gene expression related to immune activation between glycosylation score subgroups. The difference in the expression of **(C)** immunosuppressive cell recruitment factors and **(D)** immunosuppressive factors between glycosylation score subgroups. **(E)** Difference in the gene expression related to immune checkpoints between glycosylation score subgroups. **(F)** Difference in the gene expression of other related biological processes including immune activation, stromal activation, and cancer promotion between glycosylation score subgroups. ERI, extracellular matrix (ECM) receptor interaction; Pan-F-TBRS, pan-fibroblast TGFb response signature; APAP, antigen processing and presentation; TLR, Toll-like receptor; NLR, NOD-like receptor; RILR, RIG-I-like receptor; TCR, T cell receptor; BCR, B cell receptor.

Discussion

Increasing evidence demonstrated that glycosylation plays an indispensable role in metabolism, immune response, and malignancy (Fujita et al., 2021; Hu et al., 2021; Pijnenborg et al., 2021), which is a complex and multistep process through multiple glycosyltransferase enzymes (Schjoldager et al., 2020). However, most studies concentrated on a single glycosyltransferase enzyme, and the overall metabolism and immune microenvironment landscape characteristics mediated by integrated glycosyltransferase enzymes have not been fully recognized. Therefore, the identification of distinct glycosylation modification patterns in the metabolism and immune microenvironment landscape characteristics will contribute to improve our understanding of the roles of glycosylation modification patterns and provide novel insights for making more effective managements of LGGs.

In the present study, we identified two distinct glycosylation modification patterns, characterized by different prognosis, metabolism, and immune microenvironment features. The glycosylation subtype A and cluster A were characterized by the higher carbohydrates and amino acid metabolism activity, higher level of infiltrating cells, and poor prognosis. However, an opposite modification pattern was observed in glycosylation

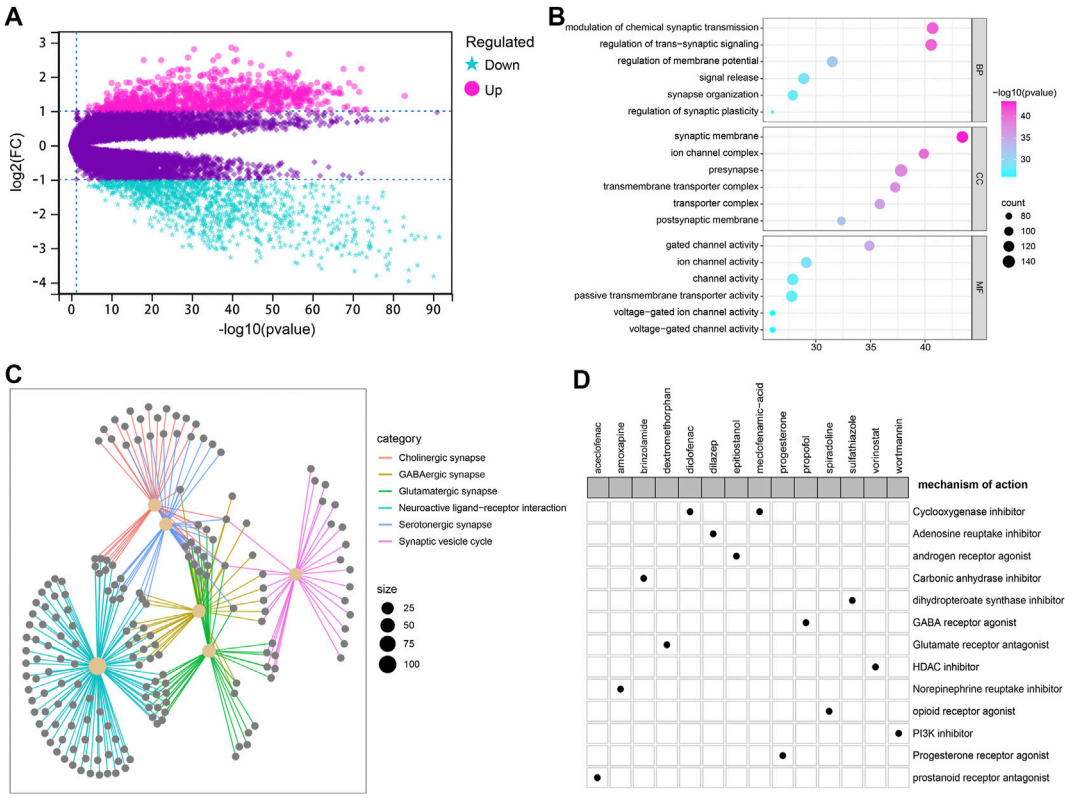


FIGURE 8
Potential small molecule compounds based on the glycosylation score. **(A)** Volcano plot showed 2090 DEGs between the high and low glycosylation score subgroups including 923 up- (red) and 1167 down-regulated (blue) genes. **(B)** Results of GO functional annotation for the DEGs, including biological process (BP), cellular component (CC), and molecular function (MF). **(C)** KEGG pathway enrichment analysis results showed the top six pathways ranked by the *p*-value. **(D)** CMap MoA analysis identified 13 mechanisms of action shared by 14 small therapeutic drugs potentially applicable for high glycosylation score patients.

subtype B and cluster B, that is, characterized by the higher lipid and other metabolism activity, lower level of infiltrating cells, and better prognosis. A previous study demonstrated that the tumor microenvironment plays a crucial role in immune response and tumor progression (Galon and Bruni, 2019). Appropriate stromal status contributes to migration of T cells, whereas loose or dense stromal might hinder the migration of T cells and limit their entry into the tumor parenchyma (Salmon et al., 2012). In addition, the abundance of immune cells would prevent them from penetrating tumor cell nests and being retained in the surrounding stroma (Joyce and Fearon, 2015). Consist with these findings, the glycosylation subtype A and cluster A infiltrated higher immune cells but with poor prognosis, reflecting an inefficient antitumor immune. Metabolic reprogramming is one of the well-established hallmarks of cancers, a flexible metabolic reorganization, which is tailored to meet their energy requirement and maintain the homeostasis of environmental conditions (An and Duan, 2022; Corchado-Cobos et al., 2022). These metabolic variations include the synthesis of protein, cell membranes, and nucleic acids, which all promote cell proliferation (Corchado-Cobos et al., 2022). Confusingly, lipid metabolism is widely believed to be oncogenic and facilitate tumor progression; our analysis showed that hyperactive lipid metabolism was related to a better prognosis, indicating an anticancer role of lipid metabolism in LGGs. However, our finding was confirmed by present studies (Peng et al., 2018; Shen et al., 2020). This is an attractive and novel theory, and the underlying mechanism is still ambiguous and needs to be confirmed.

Furthermore, we developed a scoring system termed ‘glycosylation score’ by using GSVA to quantify the glycosylation modification patterns and provide more precise management strategies for individual LGGs patients. GSVA, a non-parametric and unsupervised method, can map gene expression data from a sample into predefined gene sets and summarize it into a single enrichment score for each gene set. A significant benefit of this gene set–based method is the rationality that genes

functioned collectively and varied by genetic modifications or disease states. Another advantage is the comparability that GSVA could calculate an expression-level statistic with different dynamic ranges to a common scale. In other words, the enrichment scores of the same gene set in different datasets are comparable (Hänzelmann et al., 2013). While, many existing studies adopted least absolute shrinkage and selection operator (LASSO) regression (Tan et al., 2022; Yang and Zhang, 2022) or principal component analysis (PCA) (Zhang et al., 2020; Chong et al., 2021), which cannot satisfy the aforementioned two advantages at the same time. In result, the glycosylation modification pattern subtype A and cluster A showed a higher glycosylation score, while the glycosylation modification pattern subtype B and cluster B exhibited a lower glycosylation score. Consistent with the aforementioned findings, the glycosylation score was closely related to prognosis, metabolism status, and immune microenvironment cell-infiltrating. Moreover, integrated analyses also demonstrated that the glycosylation score was an independent prognostic index in LGGs, and its predictive efficacy was evaluated by a time-dependent ROC analysis and validated in three independent external cohorts. Collectively, the glycosylation score did play a non-negligible role in LGGs progression and became a novel and reliable tool for quantifying glycosylation modification patterns in LGGs.

Immunotherapy, an emerging potential therapy for cancers, has attracted considerable attention in recent years. However, only a subset of patients responds well to this therapy (Ott et al., 2021). Thus, it is particularly important to stratify LGGs patients and screen the ones who may benefit from immunotherapy. Further analyses highlighted that the high glycosylation score subgroup is closer to a microenvironment characterized by chronic inflammation, immunosuppression, and tumor promotion. We observed that this subgroup showed higher levels of cytokine, chemokine, and biological processes activity, involving immune activation, immunosuppression, stromal activation, and cancer promotion. Previous studies demonstrated that activation of EMT- and TGF-β–related pathways could lead to a decrease of lymphocyte cells into tumor parenchyma (Galon and Bruni, 2019) and a weakness of their cytotoxicity effects (Salmon et al., 2012), which were considered T-cell suppressive. In addition, the previous publication confirmed that stromal activation represented a major mechanism of immune evasion (Galon and Bruni, 2019) and mediated resistance to checkpoint immunotherapy (Mariathasan et al., 2018). Accordingly, we speculated that the patients with high glycosylation scores are more likely to experience resistance to immunotherapy. Next, the findings of higher IPS distributed in the low glycosylation score subgroup also suggested that patients with low glycosylation scores may respond better to immunotherapy.

To further gain novel insight into the applicability of glycosylation score and provide clues for optimizing the personalized treatment of patients, the CMap database was applied to explore potential drugs and corresponding targets for high glycosylation score subgroup patients. A total of 14 potential small therapeutic drugs and 13 drug mechanisms were identified, which could contribute to providing new treatment opportunities for LGGs patients; therefore, these findings are needed to be validated by further experiments. Furthermore, the drug sensitivity analysis by the pRRophetic algorithm revealed that 29 antitumor drugs such as gefitinib, nilotinib, and axitinib responded distinctly differently between high and low glycosylation score subgroups. These results suggested that the glycosylation score influenced the therapeutic efficacy of chemotherapy and targeted therapy. Therefore, this points out new directions for the potential clinical application of the glycosylation score. A comprehensive strategy combing immunotherapy, chemotherapy, and targeted therapy based on glycosylation score may be excellent.

In summary, the glycosylation score could be applied to comprehensively evaluate the glycosylation modification patterns, along with their corresponding immune microenvironment and metabolism features. Furthermore, the glycosylation score could be used as an independent prognostic index for predicting LGGs patients' survival. More importantly, our results also provided a novel insight into promoting personalized therapy in the future, which may contribute to developing novel therapeutic drugs or exploring promising drug combination therapy strategies.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

Conceptualization: LT and HY; formal analysis: WY; investigation: GT; methodology: GT; project administration: WY; resources: GT; software: WY; supervision: WY; validation: LT and HY; visualization: WY; writing–original draft: GT; writing–review and editing: LT and HY. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We sincerely acknowledge the public databases CGGA and TCGA for providing the transcriptomic information of LGGs samples.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.886989/full#supplementary-material

References

An, Y., and Duan, H. (2022). The role of m6A RNA methylation in cancer metabolism. *Mol. Cancer* 21 (1), 14. doi:10.1186/s12943-022-01500-4

Aoki, K., Nakamura, H., Suzuki, H., Matsuo, K., Kataoka, K., Shimamura, T., et al. (2018). Prognostic relevance of genetic alterations in diffuse lower-grade gliomas. *Neuro. Oncol.* 20 (1), 66–77. doi:10.1093/neuonc/nox132

Bai, D., Feng, H., Yang, J., Yin, A., Lin, X., Qian, A., et al. (2021). Genomic analysis uncovers prognostic and immunogenic characteristics of ferroptosis for clear cell renal cell carcinoma. *Mol. Ther. Nucleic Acids* 25, 186–197. doi:10.1016/j.omtn.2021.05.009

Berzero, G., Di Stefano, A. L., Ronchi, S., Bielle, F., Villa, C., Guillerm, E., et al. (2021). IDH-Wildtype lower-grade diffuse gliomas: the importance of histological grade and molecular assessment for prognostic stratification. *Neuro. Oncol.* 23 (6), 955–966. doi:10.1093/neuonc/noaa258

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 39 (4), 782–795. doi:10.1016/j.immuni.2013.10.003

Campbell, S. L., and Wellen, K. E. (2018). Metabolic signaling to the nucleus in cancer. *Mol. Cell* 71 (3), 398–408. doi:10.1016/j.molcel.2018.07.015

Carvalho-Cruz, P., Alisson-Silva, F., Todeschini, A. R., and Dias, W. B. (2018). Cellular glycosylation senses metabolic changes and modulates cell plasticity during epithelial to mesenchymal transition. *Dev. Dyn.* 247 (3), 481–491. doi:10.1002/dvdy.24553

Chandler, K. B., Costello, C. E., and Rahimi, N. (2019). Glycosylation in the tumor microenvironment: Implications for tumor angiogenesis and metastasis. *Cells* 8 (6), E544. doi:10.3390/cells8060544

Chong, W., Shang, L., Liu, J., Fang, Z., Du, F., Wu, H., et al. (2021). m(6)A regulator-based methylation modification patterns characterized by distinct tumor microenvironment immune profiles in colon cancer. *Theranostics* 11 (5), 2201–2217. doi:10.7150/thno.52717

Corchado-Cobos, R., García-Sancha, N., Mendiburu-Eliçabe, M., Gómez-Vecino, A., Jiménez-Navas, A., Pérez-Baena, M. J., et al. (2022). Pathophysiological integration of metabolic reprogramming in breast cancer. *Cancers (Basel)* 14 (2), 322. doi:10.3390/cancers14020322

Fujita, K., Hatano, K., Hashimoto, M., Tomiyama, E., Miyoshi, E., Nonomura, N., et al. (2021). Fucosylation in urological cancers. *Int. J. Mol. Sci.* 22 (24), 13333. doi:10.3390/ijms222413333

Galon, J., and Bruni, D. (2019). Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat. Rev. Drug Discov.* 18 (3), 197–218. doi:10.1038/s41573-018-0007-y

Gittleman, H., Sloan, A. E., and Barnholtz-Sloan, J. S. (2020). An independently validated survival nomogram for lower-grade glioma. *Neuro. Oncol.* 22 (5), 665–674. doi:10.1093/neuonc/noz191

Gui, C. P., Wei, J. H., Chen, Y. H., Fu, L. M., Tang, Y. M., Cao, J. Z., et al. (2021). A new thinking: extended application of genomic selection to screen multiomics data for development of novel hypoxia-immune biomarkers and target therapy of clear cell renal cell carcinoma. *Brief. Bioinform.* 22 (6), bbab173. doi:10.1093/bib/bbab173

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Hu, S., Ren, S., Cai, Y., Liu, J., Han, Y., Zhao, Y., et al. (2021). Glycoprotein PTGDS promotes tumorigenesis of diffuse large B-cell lymphoma by MYH9-mediated regulation of Wnt-β-catenin-STAT3 signaling. *Cell Death Differ.* 29, 642–656. doi:10.1038/s41418-021-00880-2

Joyce, J. A., and Fearon, D. T. (2015). T cell exclusion, immune privilege, and the tumor microenvironment. *Science* 348 (6230), 74–80. doi:10.1126/science.aaa6204

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Mariathasan, S., Turley, S. J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., et al. (2018). TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* 554 (7693), 544–548. doi:10.1038/nature25501

Marsico, G., Russo, L., Quondamatteo, F., and Pandit, A. (2018). Glycosylation and integrin regulation in cancer. *Trends Cancer* 4 (8), 537–552. doi:10.1016/j.trecan.2018.05.009

Mereiter, S., Balmaña, M., Campos, D., Gomes, J., and Reis, C. A. (2019). Glycosylation in the era of cancer-targeted therapy: where are we heading? *Cancer Cell* 36 (1), 6–16. doi:10.1016/j.ccell.2019.06.006

Miao, Y. R., Zhang, Q., Lei, Q., Luo, M., Xie, G. Y., Wang, H., et al. (2020). ImmuCellAI: a unique method for comprehensive T-cell subsets abundance prediction and its application in cancer immunotherapy. *Adv. Sci.* 7 (7), 1902880. doi:10.1002/advs.201902880

Mohamed Abd-El-Halim, Y., El Kaoutari, A., Silvy, F., Rubis, M., Bigonnet, M., Roques, J., et al. (2021). A glycosyltransferase gene signature to detect pancreatic ductal adenocarcinoma patients with poor prognosis. *EBioMedicine* 71, 103541. doi:10.1016/j.ebiom.2021.103541

Ott, M., Prins, R. M., and Heimberger, A. B. (2021). The immune landscape of common CNS malignancies: implications for immunotherapy. *Nat. Rev. Clin. Oncol.* 18 (11), 729–744. doi:10.1038/s41571-021-00518-9

Peng, X., Chen, Z., Farshidfar, F., Xu, X., Lorenzi, P. L., Wang, Y., et al. (2018). Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. *Cell Rep.* 23 (1), 255–269. e254. doi:10.1016/j.celrep.2018.03.077

Pijnenborg, J. F. A., Rossing, E., Merx, J., Noga, M. J., Titulaer, W. H. C., Eerden, N., et al. (2021). Fluorinated rhamnosides inhibit cellular fucosylation. *Nat. Commun.* 12 (1), 7024. doi:10.1038/s41467-021-27355-9

Pinho, S. S., and Reis, C. A. (2015). Glycosylation in cancer: mechanisms and clinical implications. *Nat. Rev. Cancer* 15 (9), 540–555. doi:10.1038/nrc3982

Polmear, J., and Good-Jacobson, K. L. (2021). Antibody glycosylation directs innate and adaptive immune collaboration. *Curr. Opin. Immunol.* 74, 125–132. doi:10.1016/j.coi.2021.12.002

Salmon, H., Franciszkiewicz, K., Damotte, D., Dieu-Nosjean, M. C., Validire, P., Trautmann, A., et al. (2012). Matrix architecture defines the preferential localization and migration of T cells into the stroma of human lung tumors. *J. Clin. Invest.* 122 (3), 899–910. doi:10.1172/jci45817

Schiff, D., Van den Bent, M., Vogelbaum, M. A., Wick, W., Miller, C. R., Taphoorn, M., et al. (2019). Recent developments and future directions in adult lower-grade gliomas: society for Neuro-Oncology (SNO) and European Association of Neuro-Oncology (EANO) consensus. *Neuro. Oncol.* 21 (7), 837–853. doi:10.1093/neuonc/noz033

Schjoldager, K. T., Narimatsu, Y., Joshi, H. J., and Clausen, H. (2020). Global view of human protein glycosylation pathways and functions. *Nat. Rev. Mol. Cell Biol.* 21 (12), 729–749. doi:10.1038/s41580-020-00294-x

Shen, X., Hu, B., Xu, J., Qin, W., Fu, Y., Wang, S., et al. (2020). The m6A methylation landscape stratifies hepatocellular carcinoma into 3 subtypes with distinct metabolic characteristics. *Cancer Biol. Med.* 17 (4), 937–952. doi:10.20892/j.issn.2095-3941.2020.0402

Stowell, S. R., Ju, T., and Cummings, R. D. (2015). Protein glycosylation in cancer. *Annu. Rev. Pathol.* 10, 473–510. doi:10.1146/annurev-pathol-012414-040438

Su, J., Ma, Q., Long, W., Tang, H., Wu, C., Luo, M., et al. (2019). LCTL is a prognostic biomarker and correlates with stromal and immune infiltration in gliomas. *Front. Oncol.* 9, 1083. doi:10.3389/fonc.2019.01083

Tan, G., Wu, A., Li, Z., Chen, G., Wu, Y., Huang, S., et al. (2021). Bioinformatics analysis based on immune-autophagy-related lncRNAs combined with immune infiltration in bladder cancer. *Transl. Androl. Urol.* 10 (8), 3440–3455. doi:10.21037/tau-21-560

Tan, K., Wu, W., Zhu, K., Lu, L., and Lv, Z. (2022). Identification and characterization of a glucometabolic prognostic gene signature in neuroblastoma based on N6-methyladenosine eraser ALKBH5. *J. Cancer* 13 (7), 2105–2125. doi:10.7150/jca.69408

Thomas, D., Rathinavel, A. K., and Radhakrishnan, P. (2021). Altered glycosylation in cancer: a promising target for biomarkers and therapeutics. *Biochim. Biophys. Acta. Rev. Cancer* 1875 (1), 188464. doi:10.1016/j.bbcan.2020.188464

Wu, P., Sun, W., and Zhang, H. (2021). An immune-related prognostic signature for thyroid carcinoma to predict survival and response to immune checkpoint inhibitors. *Cancer Immunol. Immunother.* 71, 747–759. doi:10.1007/s00262-021-03020-4

Yang, F., and Zhang, Y. (2022). Apoptosis-related genes-based prognostic signature for osteosarcoma. *Aging (Albany NY)* 14, 3813–3825. doi:10.18632/aging.204042

Ye, Y., Ma, J., Zhang, Q., Xiong, K., Zhang, Z., Chen, C., et al. (2021). A CTL/M2 macrophage-related four-gene signature predicting metastasis-free survival in triple-negative breast cancer treated with adjuvant radiotherapy. *Breast Cancer Res. Treat.* 190 (2), 329–341. doi:10.1007/s10549-021-06379-1

Youssef, G., and Miller, J. J. (2020). Lower grade gliomas. *Curr. Neurol. Neurosci. Rep.* 20 (7), 21. doi:10.1007/s11910-020-01040-8

Yu, H., Zhang, D., and Lian, M. (2021a). Identification of an epigenetic prognostic signature for patients with lower-grade gliomas. *CNS Neurosci. Ther.* 27 (4), 470–483. doi:10.1111/cns.13587

Yu, L., Ding, Y., Wan, T., Deng, T., Huang, H., and Liu, J. (2021b). Significance of CD47 and its association with tumor immune microenvironment heterogeneity in ovarian cancer. *Front. Immunol.* 12, 768115. doi:10.3389/fimmu.2021.768115

Zhang, B., Wu, Q., Li, B., Wang, D., Wang, L., and Zhou, Y. L. (2020). m(6)A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. *Mol. Cancer* 19 (1), 53. doi:10.1186/s12943-020-01170-0

frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 01 September 2022
DOI 10.3389/fcell.2022.885785

Check for updates

OPEN ACCESS

EDITED BY
Ramani Ramchandran,
Medical College of Wisconsin,
United States

REVIEWED BY
Zaoqu Liu,
First Affiliated Hospital of Zhengzhou University, China
Joe Harvey Taube,
Baylor University, United States

*CORRESPONDENCE
Sudhanshu Shukla,
sudhanshu@iitdh.ac.in

SPECIALTY SECTION
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

RECEIVED 28 February 2022
ACCEPTED 02 August 2022
PUBLISHED 01 September 2022

CITATION
Kumar P, Khadirnaikar S, Bhandari N, Chatterjee A and Shukla S (2022), An epithelial-mesenchymal plasticity signature identifies two novel LncRNAs with the opposite regulation. *Front. Cell Dev. Biol.* 10:885785. doi: 10.3389/fcell.2022.885785

COPYRIGHT
© 2022 Kumar, Khadirnaikar, Bhandari, Chatterjee and Shukla. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

An epithelial-mesenchymal plasticity signature identifies two novel LncRNAs with the opposite regulation

Pranjal Kumar[1], Seema Khadirnaikar[2], Nikita Bhandari[1], Annesha Chatterjee[1] and Sudhanshu Shukla[1]*

[1]Department of Biosciences and Bioengineering, Indian Institute of Technology Dharwad, Dharwad, Karnataka, India, [2]Department of Electrical Engineering, Indian Institute of Technology Dharwad, Dharwad, Karnataka, India

The epithelial to mesenchymal transition (EMT) is crucial for cancer progression and chemoresistance. EMT is a dynamic process with multiple phases that change cell migration and invasion activity. We used pan-cancer expression data to find 14-LncRNAs that had a high correlation with the EMT markers VIM, CDH1, FN1, SNAI1, and SNAI2. The expression of 14 EMT-associated LncRNA, which also showed high cancer specificity, was used to calculate the pan-cancer EMT score. The EMT score was then applied to the 32 cancer types to classify them as epithelial, epithelial-mesenchymal, mesenchymal-epithelial, or mesenchymal tumors. We discovered that the EMT score is a poor prognostic predictor and that as tumor mesenchymal nature increased, patient survival decreased. We also showed that the cell of origin did not influence the EMT nature of tumors. Pathway analysis employing protein expression data revealed that the PI3K pathway is the most crucial in determining the EMTness of tumors. Further, we divided CCLE-cell lines into EMT classes and discovered that mesenchymal cells, which exhibited higher PI3K pathway activation, were more sensitive to PI3K inhibitors than epithelial cells. We identified Linc01615 as a mesenchymal LncRNA whose expression significantly correlated with survival in several cancer types. We showed that Linc01615 is regulated by the TGFβ-STAT3 pathway in a feedback loop. Knockdown of Linc01615 inhibited cell proliferation and migration by regulating the PI3K pathway and mesenchymal markers. We also identified RP4-568C11.4 as an epithelial cancer marker. We showed that knocking down RP4-568C11.4 decreased cell growth but not migration. In addition, we discovered that ESR1 regulates RP4-5681C11.4 in breast cancer. Taken together, we have developed a pan-cancer EMT signature. Also, we found two new LncRNAs that have different effects on cancer development and EMT.

KEYWORDS

EMT, lncRNAs, PI3K pathway, cellular plasticity, cancer classification

Abbreviations: EMT, Epithelial to mesenchymal transitions; LncRNA, Long non-coding RNA; TCGA, The Cancer Genome atlas.

Introduction

Cancer is the second most common cause of death globally (Rana et al., 2021). According to a recent study, 19.3 million people were diagnosed, and 10 million died in 2020 due to cancer (Sung et al., 2021). Disease progression, metastasis, and chemoresistance are the primary cause of the poor outcome of the majority of cancer patients (Zaorsky et al., 2017). Metastasis, which includes cell migration and invasion, plays an essential role in cancer progression (Novikov et al., 2021). For the movement, cells may need to acquire a mesenchymal phenotype through epithelial-mesenchymal transition (EMT), which is also associated with chemoresistance and cell proliferation (Dudas et al., 2020). Studies on mouse models have shown that EMT inhibition makes cells more resistant to chemotherapeutic drugs (Fischer et al., 2015; Zheng et al., 2015). During EMT, epithelial cells gradually acquire a mesenchymal phenotype, including enhanced migration capabilities, apoptosis resistance, etc. (Kalluri and Weinberg, 2009). However, EMT shows a range of phenotypes, possibly due to variation in the cells' microenvironment and genetic and epigenetic composition (Taki et al., 2021). Also, as EMT is a dynamic and reversible process, many cancers acquire some mesenchymal phenotype and maintain some epithelial nature (Williams et al., 2019). Hence, we expect to see tumors with complete mesenchymal, complete epithelial, or mixed phenotypes; for example, Oesophageal squamous carcinoma has four different EMT phenotype types (Sung et al., 2011). Since EMT is associated with proliferation, chemoresistance, and metastasis, it is crucial to identify each tumor's EMT properties (here on refers to EMTness). This classification may also help in the identification of novel treatment options.

Generally, the presence of mesenchymal markers Vimentin (VIM), fibronectin (FN1), Snail (SNAI1), Slug (SNAI2), and absence of E-Cadherin (CDH1) expression defines the mesenchymal phenotype, and expression of these markers in vice versa fashion defines the epithelial phenotype. However, recent guidelines have suggested that EMT status should not be defined based on a small number of molecular markers (Yang et al., 2020). Also, most proteins are not cancer and tissue-specific and show variable expression (Iyer et al., 2015). In comparison, Long non-coding RNAs (LncRNA) are known to have striking tissue and cancer-specific expression (Iyer et al., 2015). LncRNAs are >200 nt long RNA molecules that lack the capacity to code. Recent research has linked LncRNAs to a variety of cellular processes through a range of mechanisms. HOTAIR, a well-studied LncRNA, for example, influences cancer growth and metastasis in a variety of cancer types (Hajjari and Salavaty, 2015). GAS5, MEG3, NBAT1, PTENP1, ANRIL, MALAT, NKILA, and many other LncRNAs have been extensively studied and are known to promote cancer characteristics such as DNA damage, immunological scape, metabolic dysfunction, stemness, and so on (Pardini et al., 2019). One of the most critical pathways in EMT control is the TGFβ pathway. Many LncRNAs are known to modulate EMT through the TGFβ pathway. PVT1, which is overexpressed in many cancer types, promotes EMT through altering SMAD2/3 phosphorylation and activation (Zhang et al., 2018). SNGH1 is another LncRNA that regulates the TGFβ pathway. SNGH1 acts as a sponge for many miRNAs that modulate the TGFβ pathway (Wang et al., 2018). The LncRNAs stimulated by TGFβ, ATB, and HIT function as a sponge, suppressing the action of the EMT transcription factors ZEB1 and ZEB2 (Yuan et al., 2014; Jia et al., 2016). MEG3 regulates the TGFβ *via* an RNA-DNA triplex structure (Mondal et al., 2015). Similarly, PNUTS has been linked to the control of breast cancer metastasis *via* the regulation of the EMT process (Grelet et al., 2017). In addition, HOTAIR has been demonstrated to influence the EMT process in a range of cancers (Gupta et al., 2010; Padua Alves et al., 2013). Many additional LncRNAs have been linked to the overall EMT process. However, since EMT is a multi-layered process, the involvement of LncRNA in EMT regulation deserves more investigation. As a result, in this study, we used PAN cancer expression data to identify EMT-associated LncRNAs and construct the EMT score. The EMT score was used to categorise TCGA tumor and CCLE cell lines into several EMT classes. We also discovered two new LncRNAs, Linc01615 and RP4-568C11.4, that inhibit cell growth and migration in opposing ways. The workflow used in this study is given in Supplementary Figure S1.

Materials and methods

Data sets and epithelial to mesenchymal transition score analysis

Expression analysis was done using the RNA-sequencing data from TCGA-GDC. RNA-seq data represented by FPKM values for all the samples from the TCGA project were downloaded from the GDC website using the "TCGAbiolinks" library in R for each tumor type separately. Tumor and normal samples from each tumor type were segregated based on the barcode and used for further analysis. Survival data for all the tumor types were also downloaded from the firehose repository. The identification of EMT LncRNA and calculation of EMT score is described in Figure 1A. The expression of VIM, FN2, SNAI1, SNAI2, and CDH1 was correlated with all the annotated LncRNA in TCGA cancers. The LncRNAs having a correlation coefficient of more than 0.35 with VIM and a correlation coefficient of >0.1 with at least two of FN1, SNAI1, SNAI2 and a correlation coefficient of < −0.2 with CDH1 were considered mesenchymal LncRNA. Oppositely, LncRNAs with correlation coefficient of more than 0.35 with CDH1 and negative correlation coefficient with at least two of FN2, SNAI1, SNAI2 and a correlation coefficient of < −0.2 with VIM were considered epithelial LncRNAs. The overview of this is

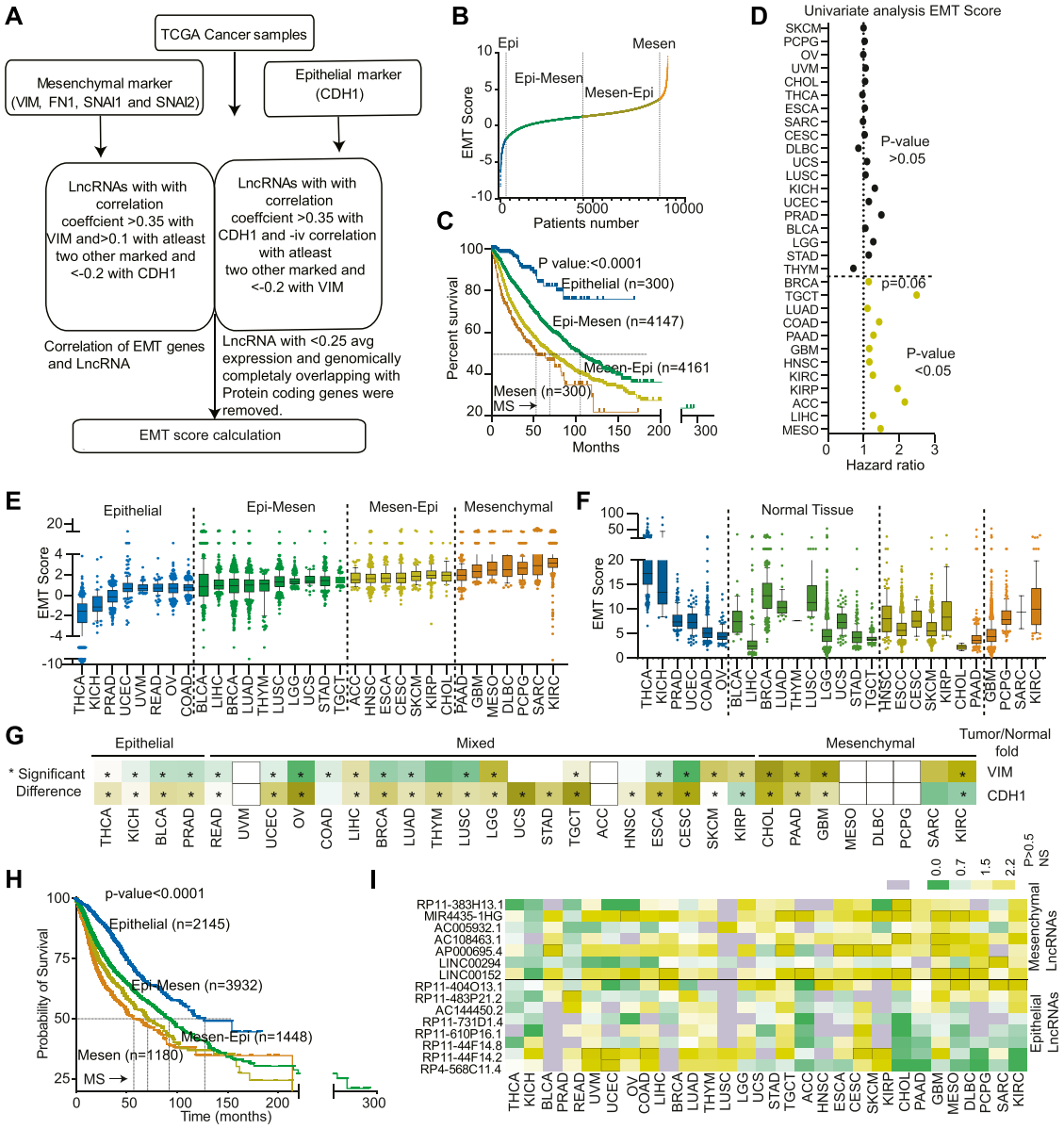


FIGURE 1
Classification of 32 cancer types based on EMTness: **(A)** A flow chart showing the steps used to identify the EMT-related LncRNAs. **(B)** An EMT score was developed for each TCGA cancer patient using the formula described in the method section. The plot shows the division of cancer patients into Epithelial (Epi), Epithelial-Mesenchymal (Epi-Mesen), Mesenchymal-Epithelial (Mesen-Epi), and mesenchymal groups. **(C)** The Kaplan- Meier plot showing the differences in survival of the four EMT groups. **(D)** EMT score for each cancer type was used to perform the univariate Cox-regression analysis. The Hazard ratio and *p*-value are depicted. **(E)** The EMT score for 32 TCGA cancer types were plotted in increasing order of average EMT score for each cancer type. The cancer types were then divided into four EMT groups. **(F)** The EMT score for normal samples was calculated and plotted to show an association between cell origin and EMTness of cancer. **(G)** Expression of VIM and CDH1 was compared between normal and tumor of each cancer type and plotted as a heatmap. Yellow is associated with increased expression, and green is associated with decreased expression. * indicates the significant comparisons. **(H)** The survival of cancers based on EMTness, as identified in Figure 2E, compared using the Kaplan-Meier plot. **(I)** The fold change in expression between normal and cancer of each of the 14 LncRNAs was calculated and plotted. Green indicates significant decreases, and yellow indicates a significant increase in expression in the tumor compared to normal. The grey indicates non-significant differences in expression.

shown in Figure 1A. For EMT score calculation, the correlation coefficient of each selected LncRNA with CDH1 and VIM1 was added. The values were then multiplied by the expression value of each LncRNA and added together to find the EMT score (Supplementary Table S1). The following formula was used for the EMT score calculation-

$$\begin{aligned}\text{EMT Score} = {} & (\text{RP11} - 383\text{H13.1} \times 0.16) + (\text{MIR4435} \\ & -1\text{HG} \times 0.15) + (\text{AC005932.1} \times 0.05) + (\text{AC108463.1} \times 0.15) \\ & + (\text{AP000695.4} \times 0.18) + (\text{LINC00294} \times 0.14) \\ & + (\text{LINC00152} \times 0.13) + (\text{RP11} - 404\text{O13.1} \times 0.09) \\ & + (\text{RP11} - 483\text{P21.2} \times 0.12) + (\text{AC144450.2} \times 0.04) \\ & + (\text{RP11} - 731\text{D1.4} \times 0.06) + (\text{RP11} - 610\text{P16.1} \times 0.06) \\ & + (\text{RP11} - 44\text{F14.8} \times 0.05) + (\text{RP11} - 44\text{F14.2} \times 0.05) \\ & + (\text{RP4} - 568\text{C11.4} \times -0.048)\end{aligned}$$

We also utilized the microarray and RNA-seq data associated with GSE146066 (Lucciola et al., 2020), GSE151251, and GSE153250 (Broome et al., 2021) datasets.

Survival, correlation analysis

To understand the association of various factors with prognosis, Cox regression and Kaplan-Meier plot with Mantle-Cox analysis were performed. For univariate Cox regression analysis, EMT score and for multivariate analysis, EMT score, age, gender, and stage were used as co-variates, and a p-value <0.05 was considered significant. For Kaplan-Meier, plots patients were divided into various groups based on the EMT score, Cancer types, or expression of novel LncRNAs. The p-value<0.05 was considered significant.

For correlation analysis, Graphpad 9.0 software was used. Spearman's non-parametric method was used to calculate the correlation coefficients.

Cell culture

The human cancer cell lines MDA-MB-231, 143B, AGS, MCF7, T47D, HCT116, HT29, H23, H460, and H522 cell lines were obtained from the National Centre for Cell Sciences (NCCS), Pune (India). The cells have been authenticated by the NCCS using STR. H1299 cells were purchased from ATCC, United States. Cell were cultured in recommended cell culture media supplemented with fetal bovine serum (10% v/v) and 100 U/mL Penicillin and 100 µg/ml Streptomycin. Cells were grown in a humidified incubator at 37°C in 5% CO_2/95% atmospheric air. All the cell lines were tested for *mycoplasma* contamination, and experiments were performed in mycoplasma-free cells only.

Knockdown of Linc01615

The knockdown study was carried out using two different shRNA constructs targeting Linc01615 RNA, along with negative control vector scramble shRNA (Plasmid #1864, and gene). shRNA hairpin sequences were cloned in lentiviral mammalian expression vector pLKO.1 puro (Plasmid #8453, addgene). The sequence pairs for hairpins (target sequences are shown in bold) are: shRNA#1_1 5′-CCGGACGTCTGCTACGAACTCAATTCTCGAGAATTGAGT TCGTAGCAGACGTTTTTTG-3'; shRNA#1_2 5′- AATTCAAAA AACGTCTGCTACGAACTCAATTCTCGAGAATTGAGTTCG TAGCAGACGT-3'; shRNA#2_1 5′- CCGGCACTCCCGGAGCAG CAGGAAACTCGAGTTTCCTGCTGCTCCGGGAGTGTTTTTG-3'; shRNA#2_2 5′- AATTCAAAAACACTCCCGGAGCAGCAG GAAACTCGAGTTTCCTGCTGCTCCGGGAGTG; Scramble shRNA sequence is 5′-CCTAAGGTTAAGTCGCCCTCGCTC GAGCGAGGGCGACTTAACCTTAGG-3'. As per the manufacturer's recommendation, all the transfections were carried out using Xfect polymer (TaKaRa Bio, cat.#631317). The transfected cells were selected using 2 µg/ml puromycin (Sigma, cat.#P8833) for a minimum of 48 h and used for RNA isolation using TRIzol reagent (Ambion, cat.#15596018), and 250 ng of total RNA was used to synthesize cDNA using iScript RT mix (Biorad, cat.#1708840). 20 ng of cDNA was used to perform the qPCR/semi-qPCR using universal SYBR Green mix (Biorad, cat.#1725271) or Emerald GT PCR mix (TaKaRa, cat.#RR310A).

PAGE and western blot

A total of 0.3×10^6 to 0.6×10^6 cells were harvested for protein isolation. Briefly, they were trypsinized and collected in separate micro-centrifuge tubes. They were washed with cold PBS and lysed in a lysis buffer (Pierce IP Lysis Buffer: 87,787) supplemented with a 1X protease inhibitor cocktail (Halt protease inhibitor cocktail: 87,786). The cell lysates were centrifuged at 13,000xg for 15 min, and supernatants were collected. The total protein concentration was estimated before loading onto polyacrylamide gel by the BCA protein assay kit (Pierce: 23,227) and a total of 30 µg of cell lysates from different experimental cells were loaded and resolved in 12% polyacrylamide gel [Biorad TGX FastCast Acrylamide Kit 12%: 161-0175)]. The gel was blotted in the PVDF membrane (Biorad: 1620177), and the membrane was blocked using a 5% blocker solution (Blotting Grade Blocker, Biorad: 1706404). The membrane was cut as per molecular size of proteins and incubated with PARP1 (proteintech: 66520-1-Ig), VIM (Invitrogen: MA5-16409), SNAI1 (proteintech: 13099-1-AP), SNAI2 (proteintech: 12129-1-AP), AKT1 (ABclonal: A10605), phosphor-AKT-S473 (ABclonal: AP1208) and internal control Actin B (ABclonal: AC004) or GAPDH (ABclonal: AC002) antibodies. The membrane was later incubated with HRP-tagged goat anti-mouse (Biorad: 172-1011) or anti-rabbit (Invitrogen: 31,460) secondary antibodies. All the steps were followed by washing with TBS (Biorad: 1706435) consisting of 0.1% Tween 20 (Polysorbate 20, MP Biomedicals: 103,168). The bound antibody complexes were detected with ECL Western Blotting substrate (Pierce: 32,209) or femtoLUCENT PLUS HRP chemiluminescent reagents (G Biosciences: 786-003).

PCR/qPCR

The polymerase chain reactions were carried out using the Emerald GT PCR mix (TaKaRa Bio: RR310A) in a thermal cycler (Biorad). The PCR samples were checked for their amplicon size and utilized for semi-qPCR using agarose gel electrophoresis. For real-time analysis, universal SYBR Green mix (Biorad: 1725271) was used, and PCR was carried out in Biorad CFX96 using two-step amplification and melt curve protocol. The qPCR analysis was performed following the delta-delta Ct value and plotted.

Proliferation assay

The cell proliferation assay was carried out using MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl-tetrazolium bromide] (Sigma, cat.#). Briefly, all the experimental cells were seeded in a 96-well plate at a density of 2,000 cells/well in triplicate for 4 days. MTT was added to the well every 24-h and incubated at 37°C and 5% CO_2 (95% atmospheric air) in a humidified incubator for 4 h. The blue-colored formazan complex was dissolved in DMSO (cat.#) and recorded the absorbance at 570 nm.

Wound healing assay

Cells were seeded in a 24-well plate at a density of 0.3×10^6 cells/well and allowed them to adhere and grow up to ≥95% confluency. Later, the wound was generated using a 200 µl sterile microtip, and the wound size was measured every 12-h up to 36-h. The wound closure was normalized and plotted.

Colony suppression assay

Cells were seeded in a 12-well plate at a density of 300 cells/well in duplicate and allowed then to grow for about 2 weeks. Later, the media was removed and washed with PBS and stained in 0.25% crystal violet (cat.#) solution. The image was captured.

TGF-β and STAT-3 inhibitor treatment

MDA-MB-231 and 143B cells were treated with TGF-β1 inhibitor (Selleckchem: S2805) and STAT-3 inhibitor Niclosamide (Selleckchem: S3030). The TGF- β1 inhibitor was used at 10 µM concentration, whereas STAT3 inhibitor was used at 10 µM for 24 h. Also, recombinant TGF- β1 (R&D Systems: 7754-BH-005/CF) was used at 2.5 ng/ml for 24 h. The recombinant TGF- β1 and STAT3 inhibitors were also used in combination at the same concentration for 24 h. Later, cells were harvested and utilized for qRT-PCR.

Drug sensitivity assay

To study the effect of various drugs/small molecules on cell lines featuring epithelial and/or mesenchymal characteristics (Epithelial: AGS and MCF7; Mesenchymal: 143B), different concentrations of each drug was formulated and used for treatment. Briefly, ~5,000 cells were plated in 96-well plate and grown for ~24 h. Later, the cells were treated with different concentrations of each drug (in duplicate) and after 48 h their viability was assayed using MTT (as described previously). The absorbance readings were normalized and plotted.

Results

Identification of epithelial to mesenchymal transition associated LncRNAs

The epithelial-mesenchymal transition (EMT) is critical in cancer progression and metastasis (Yang et al., 2020; Novikov et al., 2021). We hypothesized that LncRNAs with a high association to EMT marker expression and tumor specificity could aid in the categorization of epithelial and mesenchymal phenotypes. Hence, we analyzed TCGA expression data from 32 cancer types (~9,000 patient samples) to identify LncRNAs associated with EMT markers, VIM, FN1, SNAI1, SNAI2, and CDH1 (Figure 1A, and material and method). We identified mesenchymal and epithelial LncRNA (EMT LncRNA) using the method described in the material and method section (Supplementary Figure S2A, and material and method). To further understand the overall function of these EMT LncRNAs, we performed guilt by association analysis on protein-coding genes (PcGs) with high correlation (identified using the GePIA2 database) with EMT LncRNAs. The analysis showed that mesenchymal LncRNAs were associated with cancer progression pathways (Supplementary Figure S2B). Epithelial LncRNAs, on the other hand, have been connected to epithelial growth, transport, and metabolic signalling (Supplementary Figure S2B). Interestingly, a pan-cancer analysis, tumor vs. normal comparison revealed that mesenchymal LncRNAs are frequently overexpressed in tumors and linked to poor survival (Supplementary Figures S2C,D). On the other hand, epithelial LncRNAs were downregulated or marginally upregulated and associated with better prognosis (Supplementary Figures S2C,D).

Classification of tumors in epithelial to mesenchymal transition classes

Next, we sought to classify tumor based on their EMT potential based on the expression of EMT LncRNA identified in previous section. An EMT score was calculated by adding the

weighted expression data of the LncRNAs as described in the material and method section and illustrated in Figure 1A. EMT score was then used to classify the tumors into Epithelial, Epithelial-Mesenchymal, Mesenchymal-Epithelial, and mesenchymal groups (Figure 1B and Supplementary Table S1). The Kaplan-Meier analysis revealed that patients in the epithelial group survived much better (MS-unidentified) than those in the mesenchymal group, which had a very dismal survival rate (MS-52.8 months) (Figure 1C). The other two groups had intermediate survival, with epithelial-mesenchymal patients outliving mesenchymal-epithelial patients (Median survival 106.66 vs. 70.66 months) (Figure 1C). Next, to understand the association of EMT score with individual cancer types, we performed a univariate Cox regression analysis of EMT scores. The results identified the EMT score as significant predictor of poor survival in the 12 cancer types (Figure 1D). Furthermore, to find the prognostic independence of the EMT score in these 12 cancer types, a multivariate analysis with age, stage, gender, and EMT score was performed. The analyses revealed that except for KIRP (p-value 0.2), GBM (p-value 0.08), LUAD (p-value 0.1) and TGCT (p-value 0.09), in other eight cancers, EMT score was independent prognosticator of survival (p-value<0.05), (Supplementary Table S3).

Classification of cancer types in epithelial to mesenchymal transition classes

In the previous section, we identified EMT classes across the cancer type. Next, we aim to classify different cancer types based on EMT scores. We calculated the average EMT score for each cancer type and ranked them based on it. Interestingly, we discovered variability in EMTness between cancer types and also among tumors of the same type (Figure 1E). Certain cancers, such as THCA, KICH, BLCA, and PRAD, exhibited a significantly high epithelial phenotype (Figure 1E). In contrast, malignancies such as PAAD, GBM, MESO, DLBC, PCPG, SARC, and KIRC were mostly mesenchymal in origin (Figure 1E). Other cancer types showed a mixture of epithelial and mesenchymal characteristics, classified as epithelial-mesenchymal and mesenchymal-epithelial tumors (Figure 1E). We also computed the EMT score for the normal tissues corresponding to each cancer type. We found no association between the EMTness of normal and equivalent malignancy, suggesting that the EMT nature of the tumor is independent of the tissue of origin (Figure 1F). Interestingly, compared to corresponding normal, epithelial cancer types showed higher expression of CDH1 and lower expression of VIM. Similarly, mesenchymal cancers showed higher expression of VIM in comparison to their respective normal (Figure 1G). CDH1 was, however, consistently overexpressed in most of the cancer types (Figure 1G). Additionally, tumors with a high epithelial component had a considerably longer median survival (undefined) compared to tumors with a strong mesenchymal component (median survival of 64 months) (Figure 1H). As expected, epithelial-mesenchymal and mesenchymal-epithelial tumors showed intermediate survival rates, with epithelial-mesenchymal patients living significantly longer (Median survival 92.1 vs. 70.66 months) (Figure 1H). Also, differential expression analysis showed overexpression of mesenchymal LncRNAs in mesenchymal tumors, and epithelial LncRNA showed higher expression in epithelial tumors (Figure 1I).

PI3K pathway plays a crucial role in mesenchymal phenotype determination

Next, to understand the differences in proteins and protein modifications (changes in phosphorylation levels of proteins) associated with EMT phenotype, we performed a correlation analysis between EMT score and 217 proteins assayed in the TCPA database. The proteins/modifications with ±0.2 correlation coefficient and 0.05 FDR were considered significant. The analysis identified 16 proteins with positive and 16 proteins with negative correlation (Figure 2A). The protein with the highest negative correlation with EMT score was CDH1, validating our initial analysis (Supplementary Figure S3A). Some of the know EMT inhibitors like GATA3, CLDN7, INPP4B, JNK2, etc., were also identified as negatively correlating proteins (Hu et al., 2017; Li et al., 2018; Tang et al., 2019; Bai et al., 2021). Of the 16 positively correlating proteins, 50% were proteins with a well-established role in EMT (Figure 2B). More importantly, six proteins/protein modifications with a strong positive correlation with EMT were identified as modulators of the PI3K pathway (Figures 2B,C). We also compared the protein level data and found that PI3K pathway activators like TSC1, 4EBP1 and p4EBP1, PI3K-p85, RICTOR-pT1135 were significantly increased in high mesenchymal tumors, and PI3K inhibitor like LKB1 and INPP4B expression was significantly decreased (Supplementary Figure S3B) (Hoxhaj and Manning, 2020). These observations suggest that the PI3K pathway plays a crucial role in the mesenchymal phenotype (Figure 2L). The results also indicate that patients with a high EMT phenotype may have high PI3K pathway activation and thus may be more sensitive to PI3K inhibitors.

High mesenchymal cell lines are highly sensitive to PI3K pathway inhibitors

To understand the role of the PI3K pathway in EMTness, we utilised the cell line models. First, we calculated the EMT score of cell lines from the CCLE database. As expected, the EMT score showed a significant positive correlation with mesenchymal markers and a negative correlation with

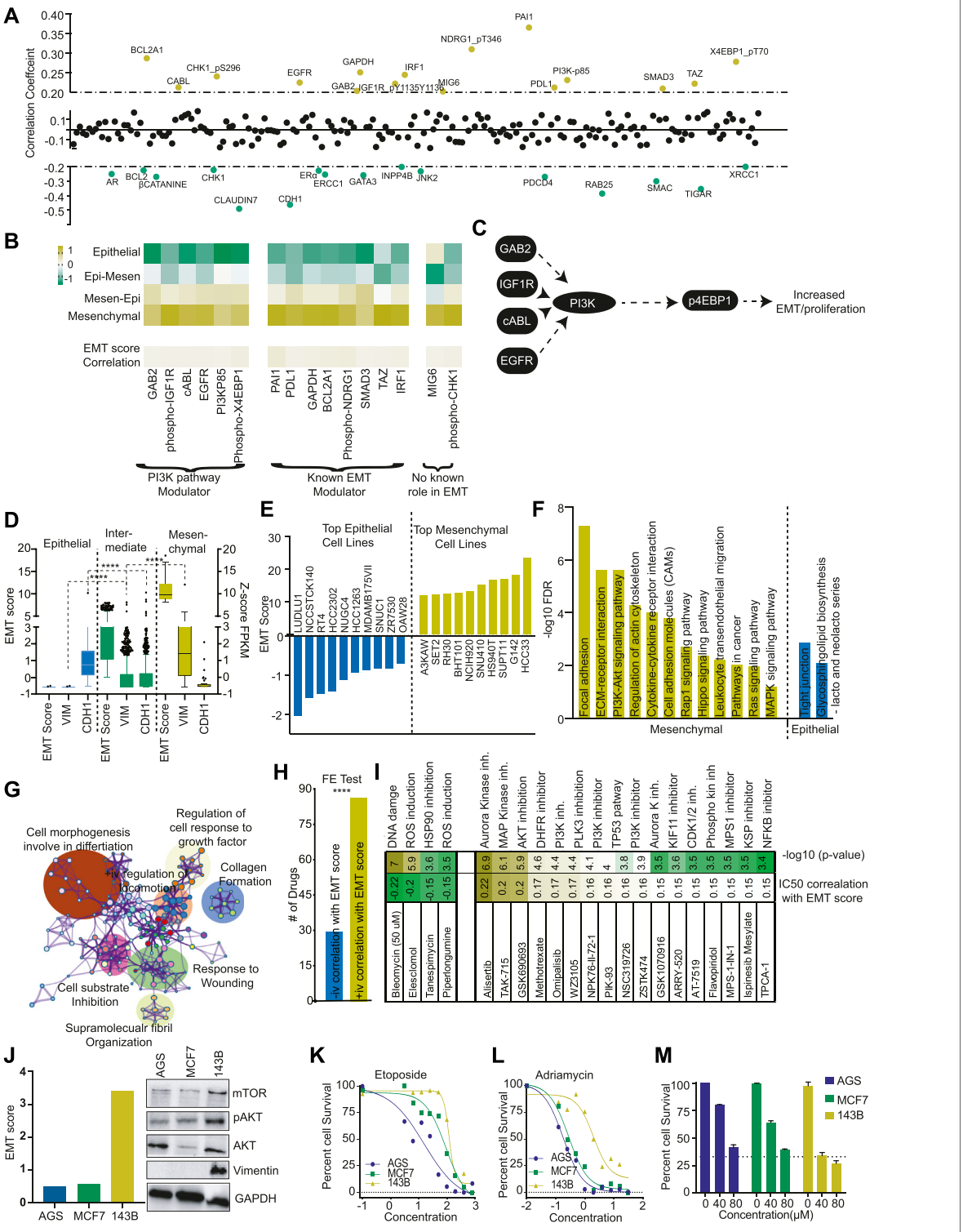


FIGURE 2
Pi3K pathway plays a major role in mesenchymal phenotype. **(A)** The correlation of EMT score with protein expression data of TCGA patients was calculated and plotted. The Yellow dots represent the positive correlation (>+0.2), and the green dots represent the negative correlation (<−0.2). **(B)** The average expression of proteins with a positive correlation with the EMT score was plotted as a heatmap for each EMT class. The proteins are divided into three classes depending upon their role in the PI3K pathway and EMT. **(C)** A diagram showing the PI3K pathway modulator with a positive correlation with the EMT score. **(D)** EMT score for each CCLE cell line was calculated, and cells were divided into epithelial, mixed, and mesenchymal groups. The expression of VIM and CDH1 was compared in three groups and plotted. **(E)** EMT score for each CCLE cell line was

(*Continued*)

FIGURE 2
calculated, and the EMT score of the top 10 epithelial and mesenchymal cells was plotted. **(F)** The GSEA analysis was done to show the activation of various pathways in mesenchymal and epithelial cell lines. **(G)** The differential expression analysis was done between mesenchymal and epithelial cells, and differentially expressed genes were used from the pathway analysis using Metascape. The pathway network was plotted as identified in the analysis. **(H)** The EMT score was correlated with the IC50 value of drugs as obtained from Genomic drug sensitivity in cancer (GDSC) databases. The drugs were divided into two groups (positive and negative correlation with EMT score). Fisher's exact test suggests that significantly more drugs showed increased IC50 in high EMT cells. **(I)** The correlation coefficient of drugs with high correlation (±0.15) with EMT scores were identified and plotted. **(J)** Three cell lines AGS, MCF (epithelial), and one 143B (mesenchymal) were selected, and levels of mTOR, pAKT, AKT, and VIM were determined with western blot. **(K, L)** The effect of DNA damaging compounds **(I)** Etoposide **(J)** Adriamycin was determined in AGS, MCF7 (epithelial), and 143B (mesenchymal cells) and plotted to show the high resistance of 143B for DNA damaging compounds. **(M)** The effect of the PI3K inhibitor was measured at two different concentrations to show the high sensitivity of PI3K inhibitors on mesenchymal cell line 143B.

CDH1 (Supplementary Figure S4A). We classified the cell lines into four classes as TCGA cancer patients samples (Supplementary Table S2). As cell lines are generally selected as models based on epithelial or mesenchymal phenotype, we mixed the two groups in the middle. We further classified the CCLE cells into epithelial, intermediate, and mesenchymal groups (Figure 2D). The cells classified as epithelial cells showed high expression of CDH1 and negligible expression of VIM (Figure 2D). In contrast, mesenchymal cells showed high VIM expression with low expression of CDH1 (Figure 2D). The cells classified as intermediate showed expression of both VIM and CDH1. (Figure 2D). We also listed top cell lines with the highest epithelial and mesenchymal nature for use as EMT models (Figure 2E). Furthermore, GO analysis showed that mesenchymal cells expressed genes associated with Adhesion, ECM-receptor interaction, PI3K signalling, MAPK signalling, etc.(Figure 2F). Similarly, Metascape network analysis showed that the genes differentially expressed between mesenchymal and epithelial cell lines were enriched in cell movement and wound healing processes (Figure 2G). Next, we analysed drug sensitivity data from the DGSC database and found that, generally, epithelial cells were more sensitive to drugs than mesenchymal cells (Figure 2H). However, the closer analysis showed that epithelial cells were more sensitive to general chemotherapeutic drugs affecting DNA damage, ROS activation, etc. In contrast, mesenchymal cells were sensitive to pathway inhibitors, specifically the PI3K pathway (Figure 2I). To further validate the findings, we selected 3 cell lines, MCF7, AGS (Epithelial), and 143B (mesenchymal). MCF7 and AGS showed low EMT scores and low expression of VIM and PI3K pathway markers compared to 143B cells (Figure 2J). As found in DGSC data, MCF7 and AGS cells were more sensitive to DNA damaging agents, Etoposide, and Adriamycin than 143B cells (Figures 2K,L). In contrast, 143B cells were significantly more sensitive to PI3K inhibitor LY294002 at tested concentrations (Figure 2M). These results support the findings from cancer patients' data described in the previous section, which suggest that tumors with high mesenchymal nature (higher EMT score) may be more sensitive to PI3K inhibitors due to higher activation of the PI3K pathway.

Long non-coding RNA Linc01615 is strongly associated with mesenchymal phenotype-

Next, we compared the gene expression of top mesenchymal and top epithelial cells to identify genes associated with epithelial and mesenchymal cell lines (Supplementary Figure S4B). The genes overexpressed in mesenchymal cells showed significant enrichment of GO terms associated with ECM organization, wound healing, etc. (Supplementary Figure S4C). In comparison, genes overexpressed in epithelial cells showed significant enrichment of GO terms associated with epithelial cell development and differentiation (Supplementary Figure S4C). Further, we checked the expression of top differentially expressed LncRNAs between mesenchymal and epithelial cells (Supplementary Figure S4D). The analysis identified 16 LncRNAs with high specificity to mesenchymal (10 LncRNA) or epithelial (6 LncRNAs) cells. We performed survival analysis across TCGA cancer and found that two LncRNAs, Linc01615 (RP11-417E7.1) and AP000695.6, were associated with poor prognosis in more than 15% of the tumor types (Supplementary Figure S4E). Both of the LncRNA were overexpressed and prognostic in multiple cancer types (Figure 3A and Supplementary Figure S5A). However, Linc01615 showed consistently strong expression patterns and mesenchymal association; hence, we studied it in detail. Correlation analysis of Linc01615 with EMT markers showed a significant positive association between Linc01615 and EMT score, SNAI1, SNAI2, VIM, and FN1 (Figure 3B). Also, the expression of Linc01615 was high in mesenchymal tumors (PAAD, GBM, SARC, and KIRC) and lowered in epithelial tumors (THCA, KICH, PRAD, and except BLCA) (Figure 3A). Even in mixed-class cancers, Linc01615 showed significantly higher expression than corresponding normal (Figure 3A). The Kaplan-Meier analysis validated the prognostic association

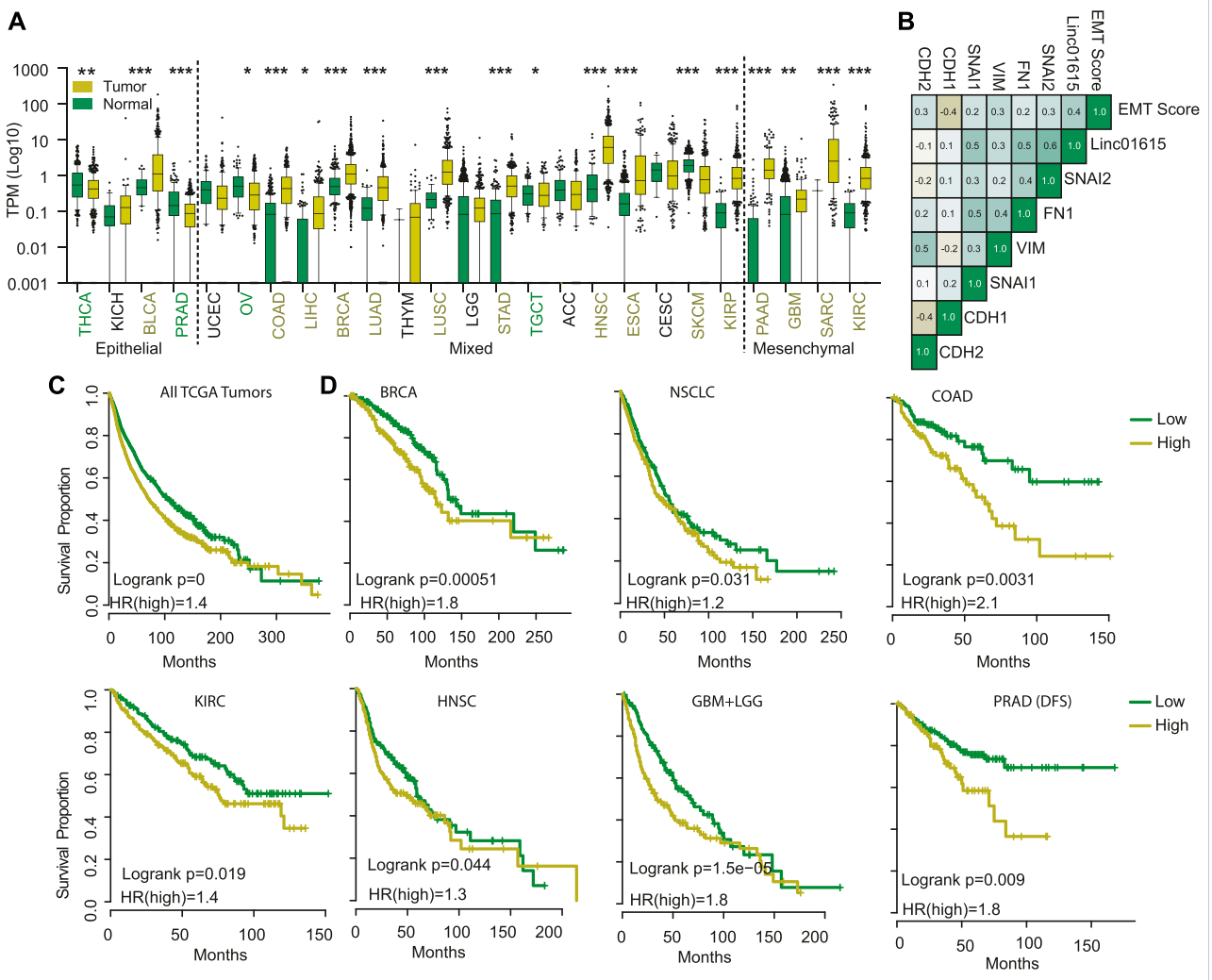


FIGURE 3
Linc01615 is a mesenchymal LncRNA. **(A)** The correlation of Linc01615 expression was done with EMT score and EMT markers and shown. **(B)** Expression of Linc01615 was compared in all the TCGA cancer types and plotted. **(C)** All the TCGA patients were combined and divided into patients with high and low Linc01615 expression, and a Kaplan-Meier analysis was performed. **(D)** Individual cancer types were divided into low and high Linc01615 groups, and survival was compared using Kaplan-Meier analysis.

of Linc01615 expression with poor survival in multiple cancer types (Figures 3C,D).

Linc01615 is regulated *via* TGFβ signalling

To understand the mechanism of the Linc01615 function in cancer, we performed the Panther pathway analysis using the genes significantly correlating with Linc01615 expression (correlation coefficient >0.3). The analysis identified significant enrichment of TGFβ and integrin pathways (Figure 4A). Thus, to check the effect of the TGFβ pathway on Linc01615, we downloaded the three RNA-seq based expression data from GEO and analyzed them. The analysis showed that in all the cell types (MCF10A, Mesenchymal Stem cells, and hepatic stellate cells) expression level of Linc01615 is increased after TGFβ treatment (Figures 4B–D). Next, we used MDAMB321 and 143B cells (Figure 4E) and treated them with TGFβ for 24 h (Figure 4F). The quantification of Linc01615 showed a substantial increase in the expression after TGFβ treatment (Figure 4F). Also, treatment of TGFβ inhibitors showed a small decrease in the expression of Linc01615 in both cells (Figure 4G). The canonical pathway of TGFβ regulates EMT *via* SMAD4. Thus, we checked for the binding of the SMAD4 at the Linc01615 promoter using ENCODE data. However, no binding of SMAD4 was seen in any of the cells. However, we noticed non-canonical TGFβ effector

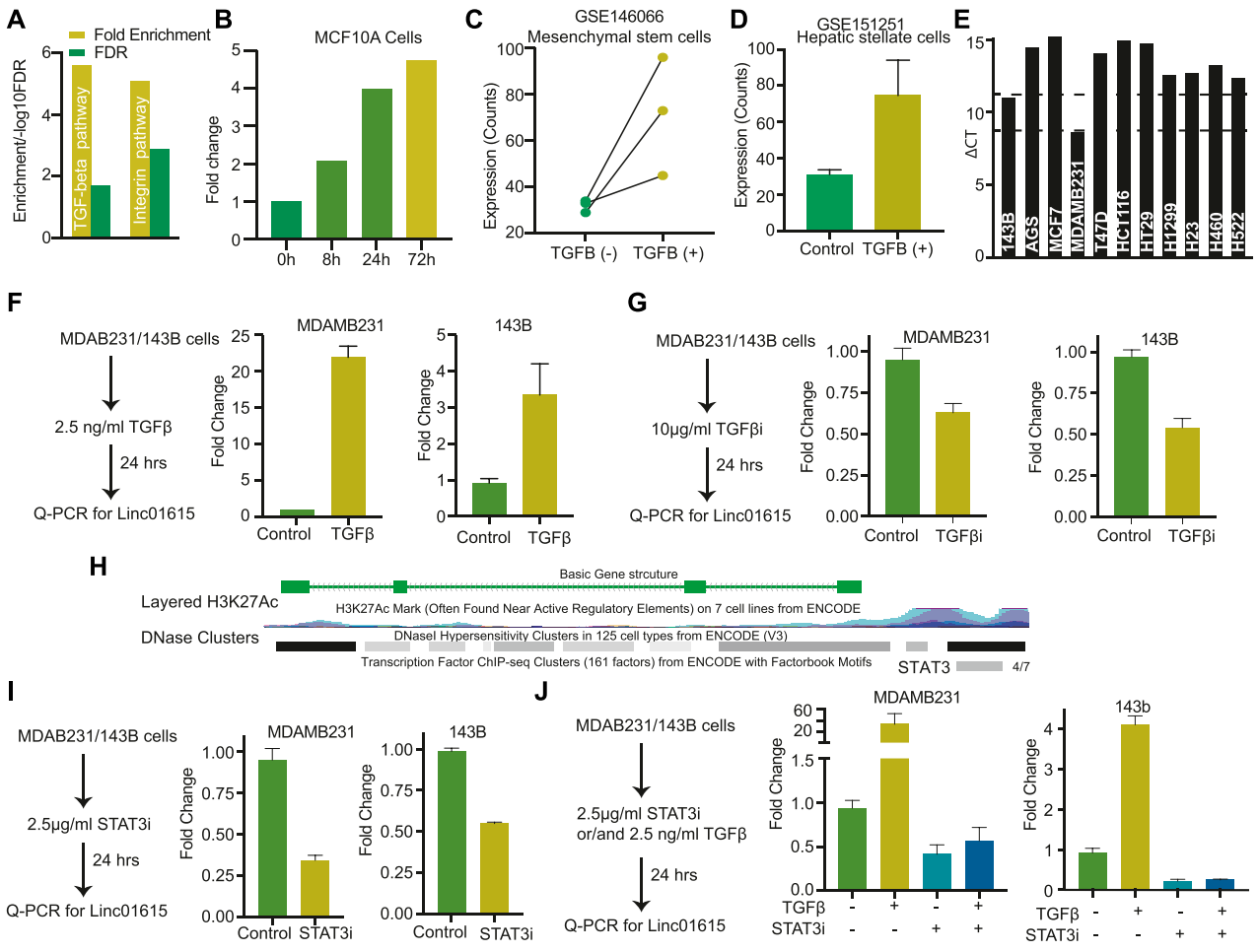


FIGURE 4
Linc01615 is regulated by the TGFβ pathway: **(A)** The genes with a high correlation coefficient with Linc01615 expression were used for KEGG pathway analysis. The Fold enrichment and FDR of enriched pathways are depicted. **(B)** Expression level of Linc001615 in MCF710A cells after treatment with TGFβ at mentioned time points. Expression of Linc01615 after TGFβ treatment in **(C)** mesenchymal stem cells and **(D)** hepatic Stellate cells. **(E)** Expression of Linc01615 in different cells as identified using semi-quantitative PCR. **(F)** Expression of Linc01615 RNA after TGFβ treatment in MDAMB231 and 143B cells. **(G)** Expression of Linc01615 RNA after TGFβ inhibitor in MDAMB231 and 143B cells. **(H)** UCSC genome browser image showing the binding site of STAT3 on the promoter of Linc01615. **(I)** Expression of linc01615 RNA after STAT3 inhibitor treatment in MDAMB231 and 143B cells. **(J)** Expression of linc01615 RNA after TGFβ, STAT3 inhibitor, or combined treatment in MDAMB231 and 143B cells.

STAT3 binding at the promoter of Linc01615 in four out of seven analyzed samples (Figure 4H). Thus, we checked the effect of STAT3 inhibitor on Linc01615 expression in MDAMB231 and 143B cells. Both the cells showed decreased Linc01615 expression after STAT3 inhibitor treatment, confirming the role of STAT3 in Linc01615 expression (Figure 4I). To see the effect of STAT3 inhibitor on TGFβ mediated Linc1615 regulation, we treated the cells with the combination of TGFβ and STAT3 inhibitor (Figure 4J). STAT3 inhibitor decreased the expression of Linc01615 even in the presence of TGFβ, suggesting STAT3 is a downstream regulator of Linc01615 (Figure 4J). These observations confirm that Linc01615 is regulated by the non-canonical TGFβ–STAT3 pathway.

Linc01615 regulated the growth and migration of cells

To understand the role of Linc01615 in cancer cells, we identified the MDAMB231 and 143B cells with high expression of Linc01615 (Figure 4E). We used two shRNA constructs to knock down the Linc01615 expression in MDAMB231 and 143B cells. The knockdown of Linc01615 expression decreased the proliferation and colony formation efficiency in both the cell lines (Figures 5A–F). We also showed that the knockdown of Linc01615 significantly inhibited wound healing in both 143B and MDAMB231 cells (Figures 4G,H). The effect of Linc01615 on cell proliferation was further validated using the inducible shRNA method. Knockdown of Linc01615 using an

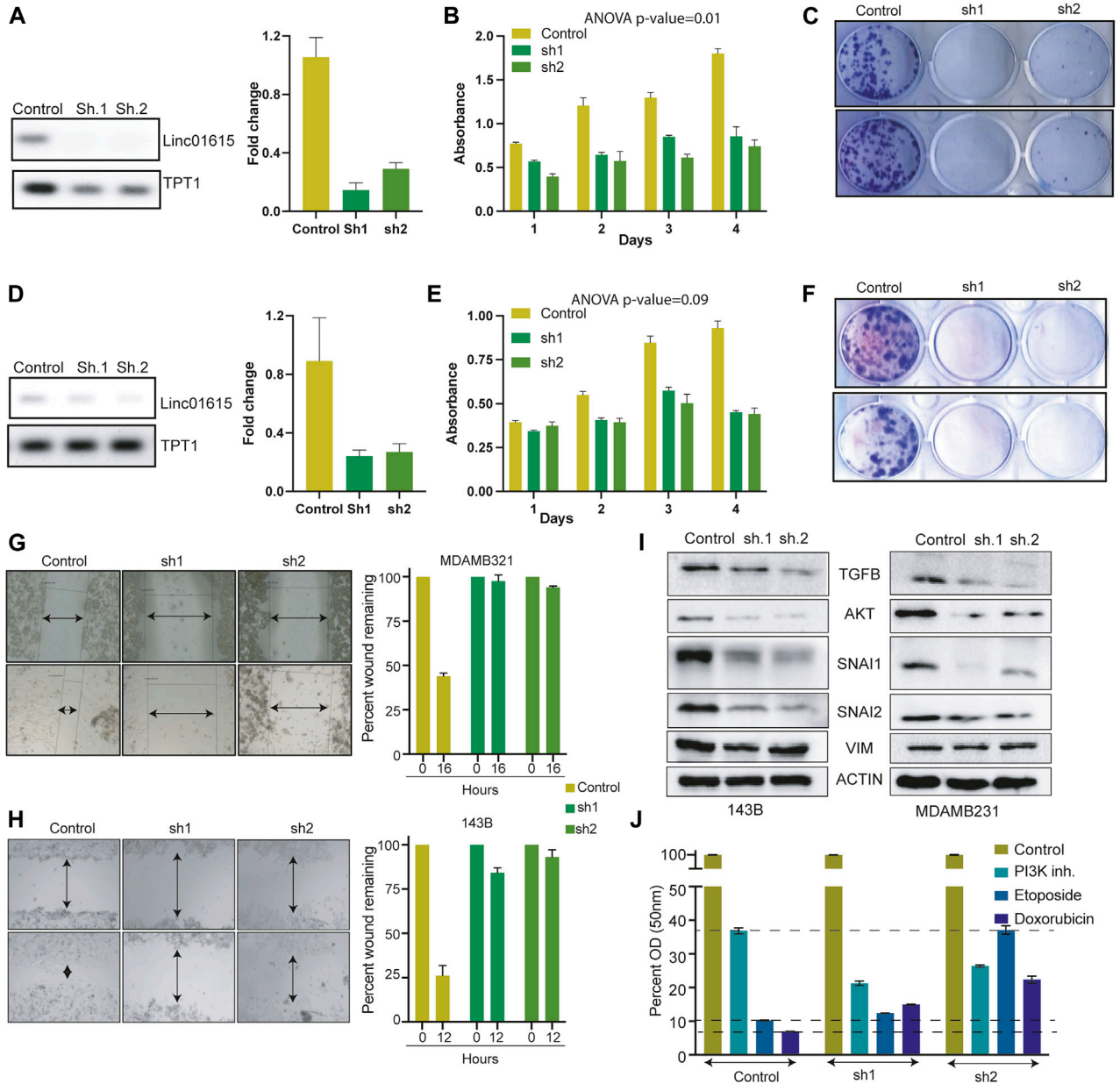


FIGURE 5
Linc01615 is required for cell proliferation and cell migration: **(A)** Semi-quantitative and Q-PCR data showing the decreased RNA level of Linc01615 after knockdown in MDAMB231 cells. Effect of Linc01615 knockdown on MDAMB231 **(B)** Cell proliferation **(C)** Colony formation ability. **(D)** Semi-quantitative and Q-PCR data showing the decreased RNA level of Linc01615 after knockdown in 143B cells. Effect of Linc01615 knockdown on MDAMB231 **(E)** Cell proliferation **(F)** Colony formation ability. The effect of Linc01615 knockdown on **(G)** MDAMB231 and **(H)** 143B cells cell migration ability. **(I)** Linc01615 was knocked down in MDAMB231 and 143B cells, and the level of mentioned protein was checked using western blot. **(J)** 143B control and Linc01615 Sh1 and sh2 cells were treated with PI3K inhibitor, Etoposide and Doxorubicin in triplicate and MTT assay was performed after 48 h.

inducible system also inhibited cell proliferation and migration (Supplementary Figures S6A–D). These observations suggest that Linc01615 expression is required for cancer development and progression. As SNAI1, SNAI2 and VIM play an essential role in cell migration, we check the effect of Linc1615 knockdown on SNAI1, SNAI2, and VIM. Interestingly, the Linc01615 knockdown decreased the expression of SNAI1 and SNAI2 in MDAMB231 and MCF7 (Figure 5I). We also noticed a slight but consistent decrease in VIM expression (Figure 5I). Surprisingly, we also found that knockdown of

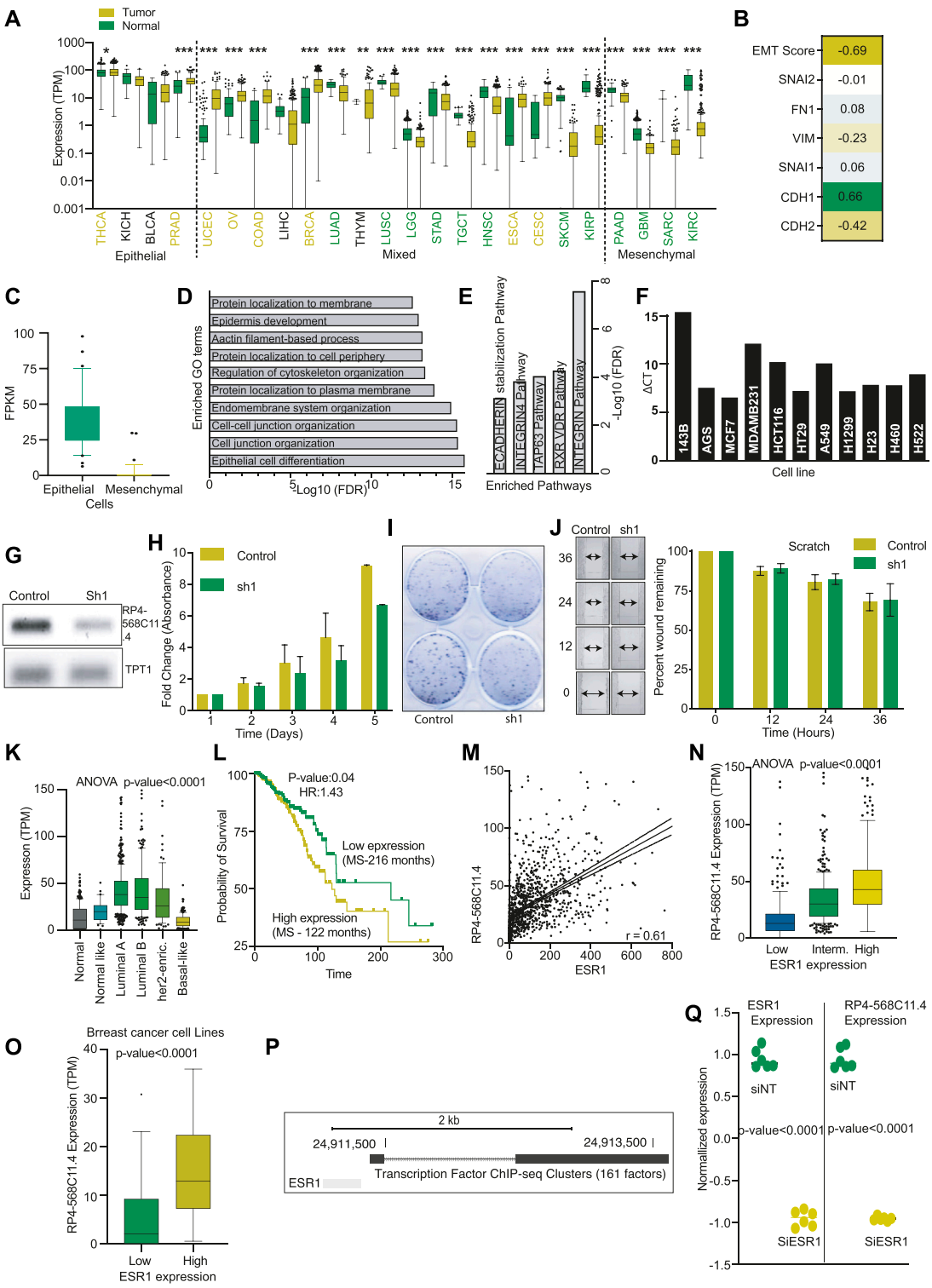


FIGURE 6
RP4-568C11.4 is an epithelial cancer gene and regulates cell proliferation. **(A)** Expression of RP4-568C11.4 was compared in all the TCGA cancer types corresponding normal and plotted. Yellow: Overexpressed in cancer, Green: Unexpressed in cancer, Black: No significant difference. **(B)** Correlation between RP4-568C11.4 and EMT markers was performed, and coefficients are represented as a heatmap. **(C)** Expression of RP4-658C11.4 in epithelial and mesenchymal cells. **(D)** Enriched GO terms and **(E)** Pathways identified using gene correlated with RP4-568C11.4 expression. **(F)** Expression of RP4-568C11.4 as identified by semi-quantitative PCR. **(G)** Expression of RP4-568C11.4 after knockdown in MCF7 cells. **(H)** Effect of RP4-568C11.4 knockdown on **(H)** Proliferation **(I)** Colony formation ability of MCF7 cells. **(J)** Effect of RP4-

(*Continued*)

FIGURE 6
568C11.4 knockdown on MCF7 cell migration. **(K)** RP4-568C11.4 expression level in different subtypes of breast cancer. **(L)** Kaplan-Meier plot showing association of RP4-568C11.4 expression with breast cancer survival. **(M)** Correlation between ESR1 and RP4-568C11.4 in breast cancer. **(N)** Expression of RP4-568C11.4 in breast cancer patients with low, intermediate, or high levels of ESR1. **(O)** Expression of RP4-568C11.4 in breast cancer cells divided into low and high ESR1 levels. **(P)** UCSC genome browser view showing ESR1 binding at RP4-568C11.4 gene. **(Q)** Expression of RP4-568C11.4 and ESR1 was measured in MCF7 cells after ESR1 knockdown to show the transcription regulation of RP4-568C11.4.

Linc01615 consistently decreases total AKT and TGFβ levels (Figure 5J). The decrease in TGFβ level suggests a feedback mechanism of Linc01615 and TGFβ expression.

RP4-568C11.4 is downregulated in mesenchymal tumors and inhibits the cell proliferation

Our analysis also identified another LncRNA, RP4-568C11.4, associated with epithelial tumors and under-expressed in mesenchymal cancers (Figure 6A). Expression of RP4-56811.4 was significantly positively correlated with CDH1 but showed a negative correlation with CDH2 and VIM (Figure 6B). Also, RP4-568C11.4 was down in all the high mesenchymal cells lines compared to epithelial cells (Figure 6C). We performed correlation analysis to identify the genes associated with RP4-568C11.4 expression. The identified genes were then used for the GO analysis. The most significantly enriched GO term was related to epithelial cell differentiation (Figure 6D). Also, the pathway analysis associated the RP4-568C11.4 with Integrin and E-Cadherin stabilization (Figure 6E). These results suggest that RP4-568C11.4 may affect differentiation and negatively regulate cell growth. Next, We performed PCR and identified MCF7 as cell lines with the highest expression of RP4-568C11.4 (Figure 6F). The knockdown of RP4-568C11.4 showed decreased proliferation of MCF7 cells (Figures 6G–I). However, we did not notice any significant difference in the migration of cells (Figure 6J). The expression analysis showed that RP4-568C11.4 was most significantly overexpressed in Breast cancer patients (Figure 6A). Hence, we further investigated the expression of RP4-568C11.4 in different subtypes of breast cancer. The expression of RP4-568C11.4 was highest in Luminal-A tumors and significantly less in Basal type of breast cancers (Figure 6K). Surprisingly, however, RP4-568C11.4 expression was significantly associated with poor survival of breast cancer patients (Figure 6L). As Luminal-A cancers are highly dependent on ESR1 activity, we correlated the expression of ESR1 with RP4-568C11.4, and found a strong positive correlation (Figurse 6M,N). Additionally, cell lines with higher expression of ESR1 showed high expression of RP4-568C11.4 (Figure 6O). We also found the ESR1 binding site at RP4-568C11.4 gene (Figure 6P). We compare the expression of RP4-568C11.4 in MCF7 cells treated with control or ESR1 siRNA using microarray data. The analysis showed that ESR1 knockdown cells have significantly lower expression of RP4-568C11.4 expression (Figure 6Q). These observations suggests that ESR1 regulates RP4-568C11.4 expression at least in breast cancer and RP4-568C11.4 regulates cell proliferation but not cell migration.

Discussion

EMT plays a crucial role in cancer progression. The part of EMT markers like VIM, FN1, SNAI1, SNAI2, CDH1, CDH2, etc., are well documented (Zeisberg and Neilson, 2009). There have been studies where expression of all the mesenchymal genes and epithelial genes were used to calculate the EMT score, and patients were classified as high or low EMT (Gibbons and Creighton, 2018). However, the Expression pattern of these genes was not sufficient to classify the cancer types into EMT classes. Here, we identified 14 LncRNAs strongly correlating EMT markers to calculate the EMT score. The survival correlation of mesenchymal and epithelial LncRNAs suggests that mesenchymal cells are more aggressive than epithelial cells (Figure 1E). Similarly, significant poor survival of patients with high EMT scores showed that patients with high mesenchymal nature could be more aggressive. The aggressiveness of tumors with more EMT scores could be due to high metastatic potential or higher resistance to the therapy. This outcome is supported by earlier reports suggesting that patients with high EMT show increased resistance to chemotherapy (Fischer et al., 2015; Zheng et al., 2015). Further, we classify the cancer types into various EMT groups. We found a majority of cancers have a range of EMT scores; or in other words, cancers have variations in EMT nature. However, some cancers are strongly epithelial (THCA, KICH, BLCA, PRAD) or mesenchymal (CHOL, PAAD, GBM, MESO, DLBC, PCPG, SARC, and KIRC) in nature. Interestingly, there was no significant correlation between the EMT score of the tumors with the tissue of origin. For example, while KIRC was highly mesenchymal compared to BLCA, normal Kidney and normal bladder samples did not show a significant difference in EMT score. Interestingly, while VIM expression was different in cancer types depending on EMT nature, CDH1 expression did not correlate with the EMT nature of the cancers. Suggesting VIM is a better reporter of the EMT than CDH1. Similar to cancer classification, cell lines were also divided into mesenchymal and epithelial groups. We have reported cells with high mesenchymal and epithelial nature, which can

prove significant in selecting cell lines for the EMT assays. We also showed that Mesenchymal cells were more resistant to the general chemotherapeutic agents than epithelial cells, validating the earlier findings. Earlier reports have suggested that mesenchymal cells are dependent on the PI3K-AKT pathway to maintain proliferation in the absence of serum (Xu et al., 2015). Authors have also shown that mesenchymal cells have high PI3K-AKT activity *in vivo*. In our analysis, we also found that mesenchymal cells are highly dependent upon the PI3K-AKT signalling. The analysis with drug sensitivity data also showed that PI3K signalling pathway inhibitors have a higher sensitivity to the PI3K pathway inhibitor than Epithelial cells, which showed increased sensitivity to general chemotherapeutic drugs (drugs inducing DNA damages, ROS, etc.).

The expression comparison of genes between epithelial and mesenchymal cells identified a novel LncRNA, Linc01615, as mesenchymal specific LncRNA. Linc01615 was positively correlated with EMT score and showed a high expression majority of mesenchymal cancers, and was associated with poor prognosis, suggesting the role of Linc01615 on the mesenchymal phenotype. The knockdown of Linc01615 also inhibited the total TGFβ and AKT. Our experiments also showed TGFβ mediated regulation of the Linc01615 expression. These findings indicate the feedback loop between TGFβ and Linc01615.

RP4-568C11.4 was another LncRNA that was identified as epithelial cancer-associated LncRNAs. Expression of RP4-658C11.4 was high in Luminal A breast cancers compared to normal and Basal cancers. Also, knockdown of RP4-568C11.4 inhibited only cell proliferation with no significant effect on migration. These results indicate that RP4-568C11.4 may be required for the development of tumors but not required for the progression of tumors.

Conclusion

In summary, we have developed an EMT signature and classified the tumors based on the EMTness of each cancer. We have shown that tumors with high mesenchymal nature tend to survive poorly. The PI3K pathway plays a crucial role in determining the Mesenchymal nature of the tumors. We also found that mesenchymal cells are more sensitive to PI3K inhibition than epithelial cells. Lastly, we have identified two novel cancer-linked LncRNAs with mesenchymal and epithelial cancer associations.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary Material.

Author contributions

PK: Performed the experiments SK: Performed the experiment and analyzed the data NB: Performed experiments and revised the manuscript. AC: Revised manuscript. SS: Performed the experiments, Analysed the data, and wrote the manuscript.

Funding

This research is supported by funding from the DBT, Govt of India (Grant ID: BT/PR27478/MED/30/1954/2018), the SERB, Govt. of India (Grant ID: ECR/2018/000528), the CSIR, Govt. of India (Grant ID: 27 (0366)/20/EMR-II) and ICMR, Govt. of India (Grant ID: 2020-3806/CMB/ADHOC-BMS).

Acknowledgments

SS acknowledges DBT, Govt of India, and SERB, Govt of India, for financial support. The authors also acknowledge TCGA for the data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.885785/full#supplementary-material

References

Bai, F., Zhang, L. H., Liu, X., Wang, C., Zheng, C., Sun, J., et al. (2021). GATA3 functions downstream of BRCA1 to suppress EMT in breast cancer. *Theranostics* 11 (17), 8218–8233. doi:10.7150/thno.59280

Broome, R., Chernukhin, I., Jamieson, S., Kishore, K., Papachristou, E. K., Mao, S. Q., et al. (2021). TET2 is a component of the estrogen receptor complex and controls 5mC to 5hmC conversion at estrogen receptor cis-regulatory regions. *Cell Rep.* 34 (8), 108776. doi:10.1016/j.celrep.2021.108776

Dudas, J., Ladanyi, A., Ingruber, J., Steinbichler, T. B., and Riechelmann, H. (2020). Epithelial to mesenchymal transition: A mechanism that fuels cancer radio/chemoresistance. *Cells* 9 (2), E428. doi:10.3390/cells9020428

Fischer, K. R., Durrans, A., Lee, S., Sheng, J., Li, F., Wong, S. T., et al. (2015). Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. *Nature* 527 (7579), 472–476. doi:10.1038/nature15748

Gibbons, D. L., and Creighton, C. J. (2018). Pan-cancer survey of epithelial-mesenchymal transition markers across the Cancer Genome Atlas. *Dev. Dyn.* 247 (3), 555–564. doi:10.1002/dvdy.24485

Grelet, S., Link, L. A., Howley, B., Obellianne, C., Palanisamy, V., Gangaraju, V. K., et al. (2017). A regulated PNUTS mRNA to lncRNA splice switch mediates EMT and tumour progression. *Nat. Cell Biol.* 19 (9), 1105–1115. doi:10.1038/ncb3595

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464 (7291), 1071–1076. doi:10.1038/nature08975

Hajjari, M., and Salavaty, A. (2015). Hotair: an oncogenic long non-coding RNA in different cancers. *Cancer Biol. Med.* 12 (1), 1–9. doi:10.7497/j.issn.2095-3941.2015.0006

Hoxhaj, G., and Manning, B. D. (2020). The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. *Nat. Rev. Cancer* 20 (2), 74–88. doi:10.1038/s41568-019-0216-7

Hu, S., Dong, X., Gao, W., Stupack, D., Liu, Y., Xiang, R., et al. (2017). Alternative promotion and suppression of metastasis by JNK2 governed by its phosphorylation. *Oncotarget* 8 (34), 56569–56581. doi:10.18632/oncotarget.17507

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., et al. (2015). The landscape of long non-coding RNAs in the human transcriptome. *Nat. Genet.* 47 (3), 199–208. doi:10.1038/ng.3192

Jia, X., Wang, Z., Qiu, L., Yang, Y., Wang, Y., Chen, Z., et al. (2016). Upregulation of LncRNA-HIT promotes migration and invasion of non-small cell lung cancer cells by association with ZEB1. *Cancer Med.* 5 (12), 3555–3563. doi:10.1002/cam4.948

Kalluri, R., and Weinberg, R. A. (2009). The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* 119 (6), 1420–1428. doi:10.1172/JCI39104

Li, Y., Gong, Y., Ning, X., Peng, D., Liu, L., He, S., et al. (2018). Downregulation of CLDN7 due to promoter hypermethylation is associated with human clear cell renal cell carcinoma progression and poor prognosis. *J. Exp. Clin. Cancer Res.* 37 (1), 276. doi:10.1186/s13046-018-0924-y

Lucciola, R., Vrljicak, P., Gurung, S., Filby, C., Darzi, S., Muter, J., et al. (2020). Impact of sustained transforming growth factor-beta receptor inhibition on chromatin accessibility and gene expression in cultured human endometrial MSC. *Front. Cell Dev. Biol.* 8, 567610. doi:10.3389/fcell.2020.567610

Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., et al. (2015). MEG3 long non-coding RNA regulates the TGF-beta pathway genes through formation of RNA-DNA triplex structures. *Nat. Commun.* 6, 7743. doi:10.1038/ncomms8743

Novikov, N. M., Zolotaryova, S. Y., Gautreau, A. M., and Denisov, E. V. (2021). Mutational drivers of cancer cell migration and invasion. *Br. J. Cancer* 124 (1), 102–114. doi:10.1038/s41416-020-01149-0

Padua Alves, C., Fonseca, A. S., Muys, B. R., de Barros, E. L. B. R., Burger, M. C., de Souza, J. E., et al. (2013). Brief report: The lincRNA Hotair is required for epithelial-to-mesenchymal transition and stemness maintenance of cancer cell lines. *Stem Cells* 31 (12), 2827–2832. doi:10.1002/stem.1547

Pardini, B., Sabo, A. A., Birolo, G., and Calin, G. A. (2019). Noncoding RNAs in extracellular fluids as cancer biomarkers: The new frontier of liquid biopsies. *Cancers (Basel)* 11 (8), E1170. doi:10.3390/cancers11081170

Rana, J. S., Khan, S. S., Lloyd-Jones, D. M., and Sidney, S. (2021). Changes in mortality in top 10 causes of death from 2011 to 2018. *J. Gen. Intern. Med.* 36 (8), 2517–2518. doi:10.1007/s11606-020-06070-z

Sung, C. O., Park, C. K., and Kim, S. H. (2011). Classification of epithelial-mesenchymal transition phenotypes in esophageal squamous cell carcinoma is strongly associated with patient prognosis. *Mod. Pathol.* 24 (8), 1060–1068. doi:10.1038/modpathol.2011.59

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Taki, M., Abiko, K., Ukita, M., Murakami, R., Yamanoi, K., Yamaguchi, K., et al. (2021). Tumor immune microenvironment during epithelial-mesenchymal transition. *Clin. Cancer Res.* 27 (17), 4669–4679. doi:10.1158/1078-0432.CCR-20-4459

Tang, W., Yang, L., Yang, T., Liu, M., Zhou, Y., Lin, J., et al. (2019). INPP4B inhibits cell proliferation, invasion and chemoresistance in human hepatocellular carcinoma. *Onco. Targets. Ther.* 12, 3491–3507. doi:10.2147/OTT.S196832

Wang, H., Wang, G., Gao, Y., Zhao, C., Li, X., Zhang, F., et al. (2018). Lnc-SNHG1 activates the TGFBR2/SMAD3 and RAB11A/wnt/β-catenin pathway by sponging MiR-302/372/373/520 in invasive pituitary tumors. *Cell. Physiol. biochem.* 48 (3), 1291–1303. doi:10.1159/000492089

Williams, E. D., Gao, D., Redfern, A., and Thompson, E. W. (2019). Controversies around epithelial-mesenchymal plasticity in cancer metastasis. *Nat. Rev. Cancer* 19 (12), 716–732. doi:10.1038/s41568-019-0213-x

Xu, W., Yang, Z., and Lu, N. (2015). A new role for the PI3K/Akt signaling pathway in the epithelial-mesenchymal transition. *Cell Adh. Migr.* 9 (4), 317–324. doi:10.1080/19336918.2015.1016686

Yang, J., Antin, P., Berx, G., Blanpain, C., Brabletz, T., Bronner, M., et al. (2020). Guidelines and definitions for research on epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* 21 (6), 341–352. doi:10.1038/s41580-020-0237-9

Yuan, J. H., Yang, F., Wang, F., Ma, J. Z., Guo, Y. J., Tao, Q. F., et al. (2014). A long non-coding RNA activated by TGF-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell* 25 (5), 666–681. doi:10.1016/j.ccr.2014.03.010

Zaorsky, N. G., Churilla, T. M., Egleston, B. L., Fisher, S. G., Ridge, J. A., Horwitz, E. M., et al. (2017). Causes of death among cancer patients. *Ann. Oncol.* 28 (2), 400–407. doi:10.1093/annonc/mdw604

Zeisberg, M., and Neilson, E. G. (2009). Biomarkers for epithelial-mesenchymal transitions. *J. Clin. Invest.* 119 (6), 1429–1437. doi:10.1172/JCI36183

Zhang, X., Feng, W., Zhang, J., Ge, L., Zhang, Y., Jiang, X., et al. (2018). Long non-coding RNA PVT1 promotes epithelial-mesenchymal transition via the TGF-β/Smad pathway in pancreatic cancer cells. *Oncol. Rep.* 40 (2), 1093–1102. doi:10.3892/or.2018.6462

Zheng, X., Carstens, J. L., Kim, J., Scheible, M., Kaye, J., Sugimoto, H., et al. (2015). Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature* 527 (7579), 525–530. doi:10.1038/nature16064

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership