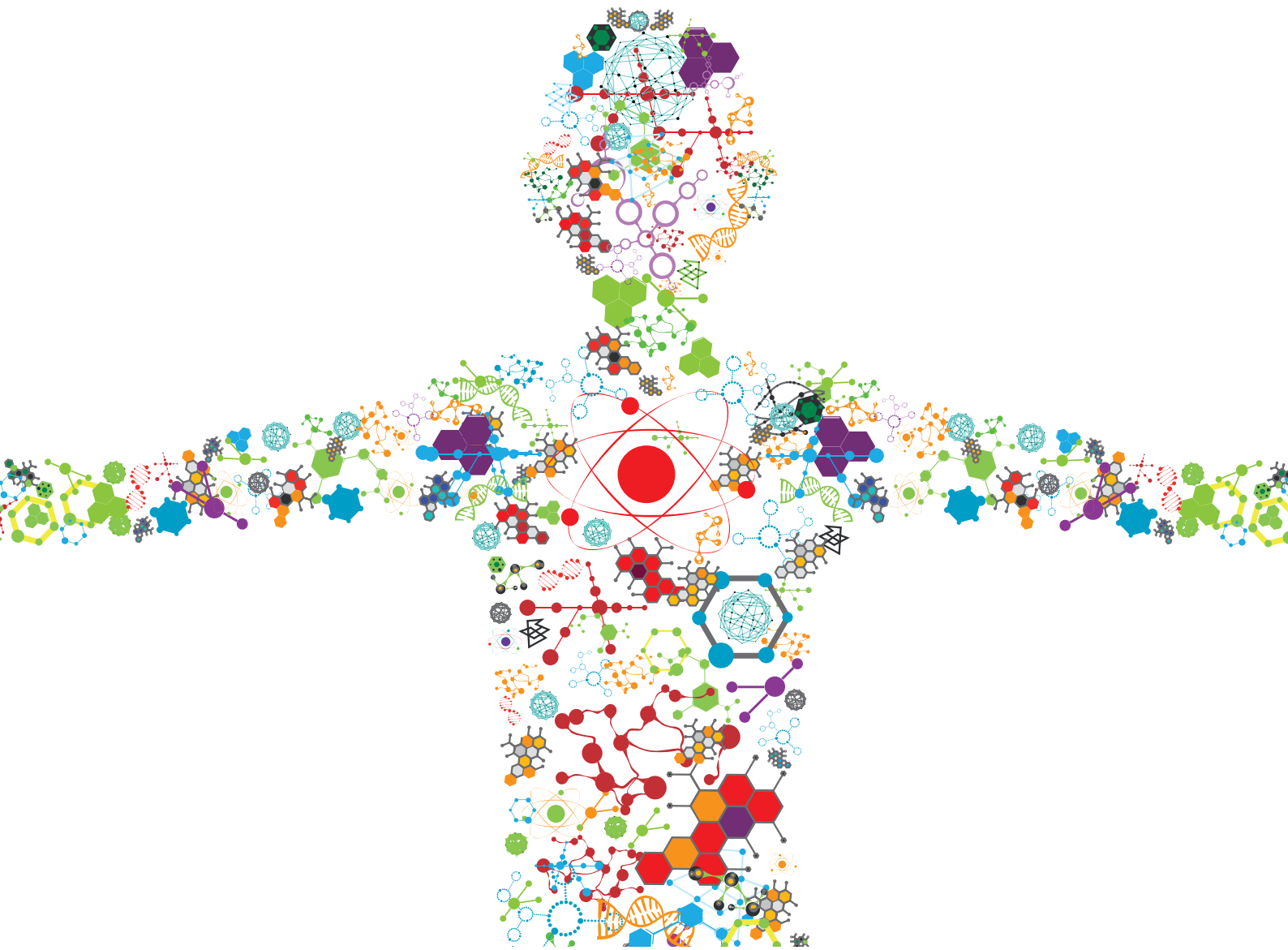
COMPUTATIONAL EPIGENETICS IN HUMAN DISEASES, CELL DIFFERENTIATION, AND CELL REPROGRAMMING, VOLUME I

EDITED BY: Jianzhong Su, Meng Zhou, Yongchun Zuo and Xiaotian Zhang
PUBLISHED IN: Frontiers in Bioengineering and Biotechnology,
Frontiers in Genetics and Frontiers in Cell and Developmental Biology



frontiers Research Topics



frontiers

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 978-2-88966-504-4
DOI 10.3389/978-2-88966-504-4

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

COMPUTATIONAL EPIGENETICS IN HUMAN DISEASES, CELL DIFFERENTIATION, AND CELL REPROGRAMMING, VOLUME I

Topic Editors:
Jianzhong Su, Wenzhou Medical University, China
Meng Zhou, Wenzhou Medical University, China
Yongchun Zuo, Inner Mongolia University, China
Xiaotian Zhang, Van Andel Research Institute (VARI), United States

Citation: Su, J., Zhou, M., Zuo, Y., Zhang, X., eds. (2021). Computational Epigenetics in Human Diseases, Cell Differentiation, and Cell Reprogramming, Volume I. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-504-4

Table of Contents

07 ***Systematic Characterization of Circular RNA-Associated CeRNA Network Identified Novel circRNA Biomarkers in Alzheimer's Disease***
Yan Zhang, Fulong Yu, Siqi Bao and Jie Sun

14 ***Identification of Prognostic and Metastatic Alternative Splicing Signatures in Kidney Renal Clear Cell Carcinoma***
Tong Meng, Runzhi Huang, Zhiwei Zeng, Zongqiang Huang, Huabin Yin, ChenChen Jiao, Penghui Yan, Peng Hu, Xiaolong Zhu, Zhenyu Li, Dianwen Song, Jie Zhang and Liming Cheng

28 ***Enhancer LncRNAs Influence Chromatin Interactions in Different Ways***
Yue Hou, Rongxin Zhang and Xiao Sun

43 ***The Construction and Analysis of ceRNA Network and Patterns of Immune Infiltration in Mesothelioma With Bone Metastasis***
Runzhi Huang, Jiawen Wu, Zixuan Zheng, Guanghua Wang, Dianwen Song, Penghui Yan, Huabin Yin, Peng Hu, Xiaolong Zhu, Haiyun Wang, Qi Lv, Tong Meng, Zongqiang Huang and Jie Zhang

58 ***Prediction of circRNAs Based on the DNA Methylation-Mediated Feature Sponge Function in Breast Cancer***
Yue Gu, Ce Ci, Xingda Zhang, Mu Su, Wenhua Lv, Chuangeng Chen, Hui Liu, Dongwei Zhang, Shumei Zhang and Yan Zhang

70 ***Identification of Common and Subtype-Specific Mutated Sub-Pathways for a Cancer***
Haidan Yan, Xusheng Deng, Haifeng Chen, Jun Cheng, Jun He, Qingzhou Guan, Meifeng Li, Jiajing Xie, Jie Xia, Yunyan Gu and Zheng Guo

79 ***CancerClock: A DNA Methylation Age Predictor to Identify and Characterize Aging Clock in Pan-Cancer***
Tongtong Zhu, Yue Gao, Junwei Wang, Xin Li, Shipeng Shang, Yanxia Wang, Shuang Guo, Hanxiao Zhou, Hongjia Liu, Dailin Sun, Hong Chen, Li Wang and Shangwei Ning

91 ***Whole-Genome Methylation Analysis of Phenotype Discordant Monozygotic Twins Reveals Novel Epigenetic Perturbation Contributing to the Pathogenesis of Adolescent Idiopathic Scoliosis***
Gang Liu, Lianlei Wang, Xinyu Wang, Zihui Yan, Xinzhuang Yang, Mao Lin, Sen Liu, Yuzhi Zuo, Yuchen Niu, Sen Zhao, Yanxue Zhao, Jianguo Zhang, Jianxiong Shen, Yipeng Wang, Guixing Qiu, Zhihong Wu and Nan Wu on behalf of the Deciphering Disorders Involving Scoliosis and Comorbidities (DISCO) Study

99 ***A Novel Approach to Identify Enhancer lincRNAs by Integrating Genome, Epigenome, and Regulatome***
Hui Liu, Tiantongfei Jiang, Shuyuan Wang, Xiang Chen, Xiaoyan Jin, Qi Wang, Xinhui Li, Jiaqi Yin, Tingting Shao, Yongsheng Li, Juan Xu and Qiong Wu

115 ***Integrative Analysis of Methylation and Transcriptome Identified Epigenetically Regulated lncRNAs With Prognostic Relevance for Thyroid Cancer***
Qiuying Li, Peng Wang, Chuanhui Sun, Chao Wang and Yanan Sun

126 ***Co-occurrence and Mutual Exclusivity Analysis of DNA Methylation Reveals Distinct Subtypes in Multiple Cancers***
Wubin Ding, Guoshuang Feng, Yige Hu, Geng Chen and Tieliu Shi

138 ***A Computational Study of Potential miRNA-Disease Association Inference Based on Ensemble Learning and Kernel Ridge Regression***
Li-Hong Peng, Li-Qian Zhou, Xing Chen and Xue Piao

153 ***Dynamic Alternative Splicing During Mouse Preimplantation Embryo Development***
Yongqiang Xing, Wuritu Yang, Guoqing Liu, Xiangjun Cui, Hu Meng, Hongyu Zhao, Xiujuan Zhao, Jun Li, Zhe Liu, Michael Q. Zhang and Lu Cai

170 ***Phylogenetic Tree Inference: A Top-Down Approach to Track Tumor Evolution***
Pin Wu, Linjun Hou, Yingdong Zhang and Liye Zhang

177 ***Nc2Eye: A Curated ncRNAomics Knowledgebase for Bridging Basic and Clinical Research in Eye Diseases***
Yan Zhang, Zhengbo Xue, Fangjie Guo, Fulong Yu, Liangde Xu and Hao Chen

183 ***The Integrated Transcriptome Bioinformatics Analysis Identifies Key Genes and Cellular Components for Spinal Cord Injury-Related Neuropathic Pain***
Runzhi Huang, Tong Meng, Rui Zhu, Lijuan Zhao, Dianwen Song, Huabin Yin, Zongqiang Huang, Liming Cheng and Jie Zhang

201 ***Transcriptomic Analyses for Identification and Prioritization of Genes Associated With Alzheimer's Disease in Humans***
Yuchen Shi, Hui Liu, Changbo Yang, Kang Xu, Yangyang Cai, Zhao Wang, Zheng Zhao, Tingting Shao and Yixue Li

211 ***RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites***
Zhibin Lv, Jun Zhang, Hui Ding and Quan Zou

221 ***A FITM1-Related Methylation Signature Predicts the Prognosis of Patients With Non-Viral Hepatocellular Carcinoma***
Jie Chen, Xicheng Wang, Xining Wang, Wenxin Li, Changzhen Shang, Tao Chen and Yajin Chen

235 ***MicroRNAs Associated With Colon Cancer: New Potential Prognostic Markers and Targets for Therapy***
Junfeng Zhu, Ying Xu, Shanshan Liu, Li Qiao, Jianqiang Sun and Qi Zhao

245 ***Combination of Four Serum Exosomal MiRNAs as Novel Diagnostic Biomarkers for Early-Stage Gastric Cancer***
Shuli Tang, Jianan Cheng, Yuanfei Yao, Changjie Lou, Liang Wang, Xiaoyi Huang and Yanqiao Zhang

255 ***A Mendelian Randomization Analysis to Expose the Causal Effect of IL-18 on Osteoporosis Based on Genome-Wide Association Study Data***
Ni Kou, Wenyang Zhou, Yuzhu He, Xiaoxia Ying, Songling Chai, Tao Fei, Wenqi Fu, Jiaqian Huang and Huiying Liu

262 ***Comprehensive Analysis of the Genetic and Epigenetic Mechanisms of Osteoporosis and Bone Mineral Density***
Hui Dong, Wenyang Zhou, Pingping Wang, Enjun Zuo, Xiaoxia Ying, Songling Chai, Tao Fei, Laidi Jin, Chen Chen, Guowu Ma and Huiying Liu

270 ***Methylation-Driven Genes Identified as Novel Prognostic Indicators for Thyroid Carcinoma***
Liting Lv, Liyu Cao, Guinv Hu, Qinyan Shen and Jinzhong Wu

284 ***PSBP-SVM: A Machine Learning-Based Computational Identifier for Predicting Polystyrene Binding Peptides***
Chaolu Meng, Yang Hu, Ying Zhang and Fei Guo

293 ***A Systematic Way to Infer the Regulation Relations of miRNAs on Target Genes and Critical miRNAs in Cancers***
Peng Xu, Qian Wu, Jian Yu, Yongsheng Rao, Zheng Kou, Gang Fang, Xiaolong Shi, Wenbin Liu and Henry Han

306 ***iRNA5hmC: The First Predictor to Identify RNA 5-Hydroxymethylcytosine Modifications Using Machine Learning***
Yuan Liu, Dasheng Chen, Ran Su, Wei Chen and Leyi Wei

314 ***Characterization of Dysregulated lncRNA-Associated ceRNA Network Reveals Novel lncRNAs With ceRNA Activity as Epigenetic Diagnostic Biomarkers for Osteoporosis Risk***
Meijie Zhang, Luyang Cheng and Yina Zhang

323 ***Identification of Pan-Cancer Prognostic Biomarkers Through Integration of Multi-Omics Data***
Ning Zhao, Maozu Guo, Kuanquan Wang, Chunlong Zhang and Xiaoyan Liu

338 ***Robust Biomarker Screening Using Spares Learning Approach for Liver Cancer Prognosis***
Aman Chandra Kaushik, Aamir Mehmood, Dong-Qing Wei and Xiaofeng Dai

347 ***Effect for Human Genomic Variation During the BMP4-Induced Conversion From Pluripotent Stem Cells to Trophoblast***
Hai-tao Li, Yajun Liu, Hongde Liu and Xiao Sun

359 ***Corrigendum: Effect for Human Genomic Variation During the BMP4-Induced Conversion From Pluripotent Stem Cells to Trophoblast***
Hai-tao Li, Yajun Liu, Hongde Liu and Xiao Sun

360 ***Detection of Allosteric Effects of lncRNA Secondary Structures Altered by SNPs in Human Diseases***
Xiaoyan Lu, Yu Ding, Yu Bai, Jing Li, Guosi Zhang, Siyu Wang, Wenyan Gao, Liangde Xu and Hong Wang

371 ***Differential DNA Methylation Encodes Proliferation and Senescence Programs in Human Adipose-Derived Mesenchymal Stem Cells***
Mark E. Pepin, Teresa Infante, Giuditta Benincasa, Concetta Schiano, Marco Miceli, Simona Ceccarelli, Francesca Megiorni, Eleni Anastasiadou, Giovanni Della Valle, Gerardo Fatone, Mario Faenza, Ludovico Docimo, Giovanni F. Nicoletti, Cinzia Marchese, Adam R. Wende and Claudio Napoli

379 ***Identification and Validation of Novel Long Non-coding RNA Biomarkers for Early Diagnosis of Oral Squamous Cell Carcinoma***
Yue Li, Xiaofang Cao and Hao Li

388 ***Integrative Analysis for Elucidating Transcriptomics Landscapes of Glucocorticoid-Induced Osteoporosis***
Xiaoxia Ying, Xiyun Jin, Pingping Wang, Yuzhu He, Haomiao Zhang, Xiang Ren, Songling Chai, Wenqi Fu, Pengcheng Zhao, Chen Chen, Guowu Ma and Huiying Liu

396 ***Characterization of lncRNA-Associated ceRNA Network to Reveal Potential Prognostic Biomarkers in Lung Adenocarcinoma***
Yang Wang, Ruyi He and Lixin Ma

407 ***Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications***
Rao Zeng and Minghong Liao

415 ***SOD1 Promotes Cell Proliferation and Metastasis in Non-small Cell Lung Cancer via an miR-409-3p/SOD1/SETDB1 Epigenetic Regulatory Feedforward Loop***
Shilong Liu, Bin Li, Jianyu Xu, Songliu Hu, Ning Zhan, Hong Wang, Chunzi Gao, Jian Li and Xiangying Xu

429 ***Bioinformatics Analysis Reveals Biomarkers With Cancer Stem Cell Characteristics in Lung Squamous Cell Carcinoma***
Yi Liao, Hua Xiao, Mengqing Cheng and Xianming Fan

443 ***Identification of a Multi–Long Noncoding RNA Signature for the Diagnosis of Type 1 Diabetes Mellitus***
Guannan Geng, Zicheng Zhang and Liang Cheng

452 ***IGF2BP3 May Contributes to Lung Tumorigenesis by Regulating the Alternative Splicing of PKM***
Huang Xueqing, Zhang Jun, Jiang Yueqiang, Liao Xin, Hu Liya, Fang Yuanyuan, Zhang Yuting, Zeng Hao, Wu Hua, Liu Jian and Yin Tiejun

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 11 September 2019
doi: 10.3389/fbioe.2019.00222

Systematic Characterization of Circular RNA-Associated CeRNA Network Identified Novel circRNA Biomarkers in Alzheimer's Disease

*Yan Zhang†, Fulong Yu†, Siqi Bao and Jie Sun**

School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou, China

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Fang Wang,
University of Texas MD Anderson Cancer Center, United States
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

***Correspondence:**
Jie Sun
suncarajie@wmu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *16 July 2019*
Accepted: *29 August 2019*
Published: *11 September 2019*

Citation:
Zhang Y, Yu F, Bao S and Sun J (2019) Systematic Characterization of Circular RNA-Associated CeRNA Network Identified Novel circRNA Biomarkers in Alzheimer's Disease. Front. Bioeng. Biotechnol. 7:222. doi: 10.3389/fbioe.2019.00222

Alzheimer's disease (AD), a degenerative disease of the central nervous system, is the most common form of dementia in old age. The complexity and behavior of circular RNA (circRNA)-associated competing endogenous RNA (ceRNA) network remained poorly characterized in AD. The aim of this study was to elucidate the regulatory networks of dysregulated circRNAs from ceRNA view and identify potential risk circRNAs involved in AD pathogenesis. Consistent differentially expressed genes (CDEGs) were obtained using meta-analysis for multiple microarrays, and differentially expressed miRNAs (DEmiRs) were identified using empirical Bayes method. The circRNA-associated ceRNA network (cirCeNET) was constructed based on "ceRNA hypothesis" using an integrated system biology method. A total of 1,872 CDEGs and 48 DEmiRs were screened across different datasets. By mapping CDEGs and DEmiRs into the cirCeNET, an AD-related circRNA-associated ceRNA network (ADcirCeNET) was constructed, including 3,907 edges and 1,407 nodes (276 circRNAs, 14 miRNAs and 1,117 mRNAs). By prioritizing AD risk circRNA-associated ceRNAs, we found that the circRNA *KIAA1586* occurred most frequently in the AD risk circRNA-associated ceRNAs and function as a ceRNA that operates by competitively binding three known AD-risk miRNAs. *In silico* functional analysis suggested that circRNA *KIAA1586*-related ceRNA network was significantly enriched in known AD-associated biological processes. Our study provided a global view and systematic dissection of circRNA-associated ceRNA network. The identified circRNA *KIAA1586* may be a key risk factor involved in AD pathogenesis.

Keywords: Alzheimer's disease, circular RNA, competing endogenous RNA, ceRNA network, system biology

INTRODUCTION

Alzheimer's disease (AD), a degenerative disease of the central nervous system, is the most common form of dementia in old age (Jiang et al., 2013; Liu et al., 2019). It is mainly manifested as neuropsychiatric symptoms such as progressive memory disorder, cognitive dysfunction, personality change, and language disorder, which seriously affect social, professional and life functions. With the increasing morbidity year by year, AD has attracted more attention from society. However, the etiology and pathogenesis of AD have not been well-elucidated yet.

Currently, the rapid advancement of high-throughput technologies offers great opportunities for biomarker identification (Liu et al., 2018a,b; Yu et al., 2018a). Non-coding RNAs as biomarker and therapeutic targets play a significant role in human diseases (Long et al., 2012; Zhou et al., 2015a,b, 2016, 2017, 2018a,b,c; Yu et al., 2018b). Circular RNAs (circRNAs) are a naturally occurring class of non-coding RNA molecules, which has become the latest research focus in the field of RNA (Qu et al., 2015). Unlike traditional linear RNA, circRNA has a closed circular structure and is not affected by RNA exonuclease, so its expression level is more stable and not easily degraded (Li et al., 2015). However, classical RNA sequencing methods can only isolate those molecules with polyA tails. The ends of circRNAs are attached together, lacking these "tails," and are generally ignored. In fact, hundreds of circRNAs are enriched in the mammalian brain tissues (Hansen et al., 2013; Rybak-Wolf et al., 2015) and have important regulatory potency (Memczak et al., 2013). But a key question remains: what do they actually do?

The occurrence of complex diseases is the result of synergism of multiple interacting genes or RNAs. Therefore, we should investigate disease mechanism at the level of system biology and mine useful information from a vast network of interacting genes or RNAs (Zhang et al., 2018). Integration analysis of transcription (protein-coding genes) and post-transcriptional (non-coding genes) regulation has proven to be a valuable strategy for studies of the genetic characteristics of various human complex diseases (Zhang et al., 2015; Arena et al., 2017). In a recent study, Welden et al. found the human MAPT gene which produces the microtubule-associated protein Tau, generates circRNAs to contribute to AD (Welden et al., 2018). Piwecka et al. (2017) demonstrated that loss of circRNA locus affected brain function. They showed circRNA *Cdr1as* functions as miRNA sponge by binding miRNA response elements (MREs) to cause miRNA deregulation. In addition, Lukiw (2013) revealed that circRNA *ciRS-7* also acts as a ceRNA to absorb miRNA. A deficiency in *ciRS-7* "sponging" effects might be expected to increase the expression level of miR-7 in AD-affected brain cells and down-regulate AD-relevant targets. These studies suggest that such circRNA-miRNA-mRNA competing system is an important epigenetic regulatory layer control over gene expression in AD (Salmena et al., 2011; Zhang et al., 2018). However, the complexity and behavior of circRNA-associated competing endogenous RNA (ceRNA) network remain poorly characterized in the pathogenesis of AD.

Therefore, in this study, by comprehensively integrating gene and miRNA expression data of AD, the AD-related circRNA-miRNA-mRNA competitive network (ADcirCeNET) was established. And then, AD risk circRNA-miRNA-mRNA relationships were optimized using the known AD-related data resources. We found circRNA *KIAA1586* could contribute to AD. Our results showed that *KIAA1586* acts as a ceRNA to absorb three miRNAs (*hsa-miR-29b*, *hsa-miR-101*, *hsa-miR-15a*) and lead to the dysregulation of AD-associated biological processes.

MATERIALS AND METHODS

Differential Expression Analysis of Gene and miRNA in AD

We obtained four gene expression profiles of AD from Gene Expression Omnibus (GEO) database, their accession ID were GSE5281 (Liang et al., 2007), GSE1297 (Blalock et al., 2004), GSE12685 (Williams et al., 2009), and GSE16759 (Nunez-Iglesias et al., 2010), respectively. Since the brain is the most complex part of all human organs, it can be divided into multiple regions with different functions. In this study, the gene expression data detected six different brain regions including entorhinal cortex, hippocampus, superior frontal gyrus, posterior cingulated, medial temporal gyrus, and primary visual cortex. To discover the common pathogenic mechanism in different brain regions, the gene expression profiles were divided into six parts according to the brain regions for separate analysis. All of the data were normalized and log2 transformed. The microarray probe IDs were converted to Entrez Gene IDs. To obtain the CDEGs, the differentially expressed gene lists were combined using the R package "metaMA" (Marot et al., 2009). The miRNA expression profile (accession ID GSE16759) was derived from the GEO database, including four AD samples and four normal controls. The DEmiRs were identified using empirical Bayes method (Nunez-Iglesias et al., 2010).

Construction of circRNA-miRNA-mRNA Competitive Network

It has been reported that circRNA dysfunction can lead to the occurrence of AD (Lukiw, 2013), and the main way of its function is competitive regulation (Qu et al., 2015). Here, a competitive network among circRNA, miRNA and mRNA was constructed. First, we downloaded the human miRNA-circRNA and miRNA-mRNA interactions from the RAID database which is a resource of RNA-associated interactions across organisms (Yi et al., 2017). A necessary condition for competitive regulation is the number of shared miRNAs (Zhang et al., 2018). And then, the miRNA-circRNA and miRNA-mRNA interactions shared at least five miRNAs were retained. Finally, all these identified miRNA-mediated circRNA-mRNA ceRNA crosstalk were integrated to build a circRNA-miRNA-mRNA competitive network (cirCeNET).

Identification of circRNA-Associated ceRNA Network of AD

The CDEGs and DEmiRs were mapped into the cirCeNET. The circRNA-miRNA-mRNA competitive relationships were extracted which contained at least one CDEG or DEmiR. These competitive relationships constitute the circRNA-associated ceRNA network of AD (ADcirCeNET).

The coverage rate of known AD-associated gene and miRNA set was used to prioritize AD risk circRNA-associated ceRNAs. If a miRNA-mediated circRNA-mRNA ceRNA crosstalk contained at least one known AD-associated gene or miRNA, it was considered as the AD risk circRNA-associated ceRNA.

Function Enrichment Analysis

The significant enriched biological functions of a gene set were explored using the R package "clusterProfiler" (Yu et al., 2012). The adjusted *P*-values were calculated using the multiple test of Benjamini and Hochberg (BH) method.

RESULTS

AD-Related Gene and miRNA Set

To reveal the common pathogenesis of AD, the gene expression profiles were split into six parts according to different brain region. Through integrated differential expression analysis across

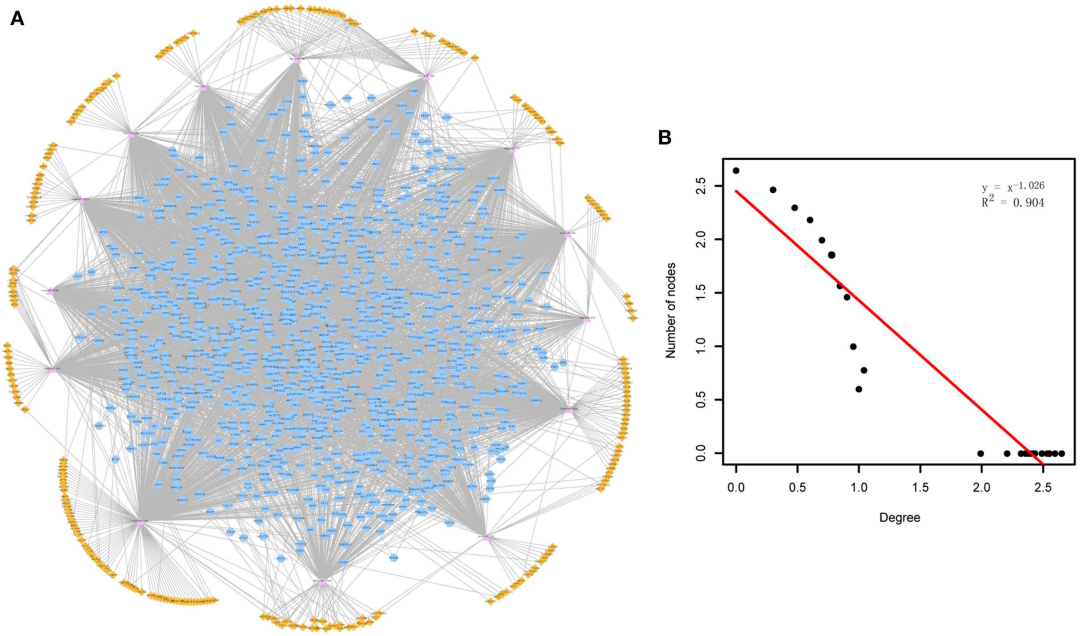


FIGURE 1 | (A) The circRNA-associated ceRNA network in AD. The orange diamonds represented circRNAs, purple triangles represented miRNAs and blue circles represented target mRNAs. Network edges represented competitive interactions. **(B)** Degree distribution of the ADcirCeNET.

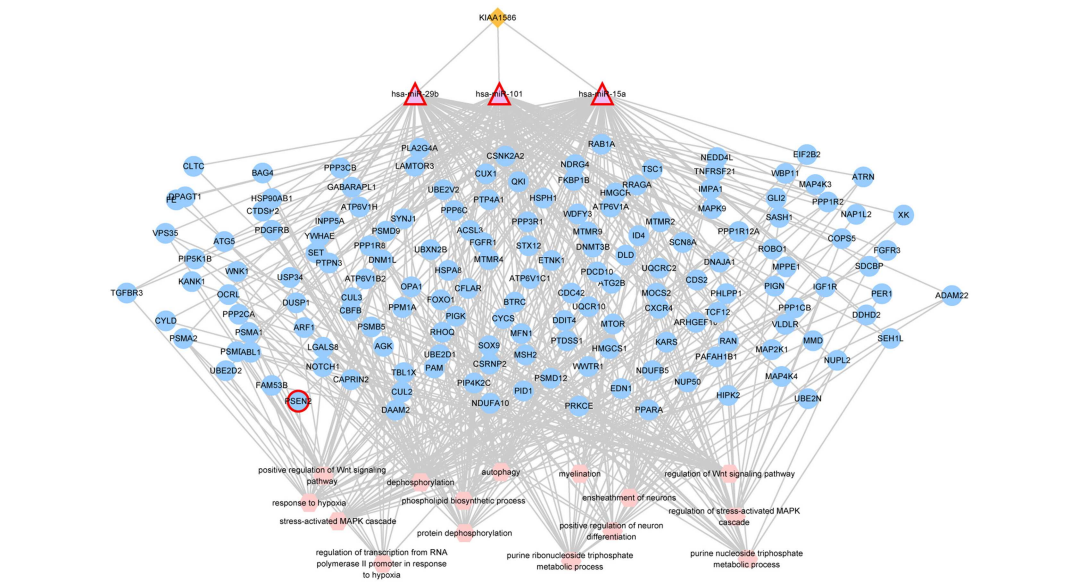


FIGURE 2 | The AD risk circRNA-associated ceRNAs involving in *KIAA1586*. The orange diamonds represented circRNAs, purple triangles represented miRNAs, blue circles represented target mRNAs, and pink hexagons represented GO biological functions. The red border represented known AD-related nodes. The edges among circRNAs, miRNAs and mRNAs represented competitive interactions, and the edges between mRNAs and functions represented that the mRNAs were annotated in the GO terms.

different datasets, 1872 CDEGs were screened under FDR <0.01. The CDEG list can be seen in **Supplementary File 1**. For miRNA expression profile, 48 DEmiRs were obtained under FDR <0.05. The DemiR list can be seen in the **Supplementary File 2**.

The known AD-associated genes and miRNAs were derived from GeneCards (Safran et al., 2010), HMDD (Lu et al., 2008) and miR2Disease (Jiang et al., 2009) databases which are all manually curated data resources. There were 27 genes and 45 miRNAs.

The circRNA-Associated ceRNA Network in AD

We derived 7,896 miRNA-circRNA interactions and 719,442 miRNA-mRNA interactions from RAID database. The miRNA-mediated ceRNA crosstalk between circRNA and mRNA were identified by filtering the number of shared miRNA. Finally, 7,120 miRNA-circRNA and 409,381 miRNA-mRNA interactions were obtained and were integrated to build a tremendous cirCeNET. The constructed cirCeNET contained 11,133 nodes (including 484 circRNAs, 280 miRNAs and 10,369 mRNA) and 416,501 edges.

Through mapping the CDEGs and DEmiRs into the constructed cirCeNET, the miRNA-mediated ceRNA crosstalks were extracted if it contained at least one CDEG or DEmiR. These miRNA-mediated ceRNA crosstalks made up the ADcirCeNET (**Figure 1A**). There were 3,888 edges including 428 miRNA-circRNA and 3,460 miRNA-mRNA interactions, and 1,407 nodes including 276 circRNAs, 14 miRNAs and 1,117 mRNAs in the ADcirCeNET (details see **Supplementary File 3**). Moreover, examination of the degree distribution of the ADcirCeNET revealed a powerlaw distribution, showing that the ADcirCeNET was scale-free, similar to most biological networks (**Figure 1B**).

The coverage rate of known AD-associated gene and miRNA set was used to prioritize AD risk circRNA-associated ceRNAs. If a miRNA-mediated ceRNA crosstalk in the ADcirCeNET contained at least one known AD-associated gene or miRNA, this miRNA-mediated ceRNA crosstalk was considered as the AD risk circRNA-associated ceRNA. Finally, 46,096 AD risk miRNA-mediated ceRNA crosstalk, including 158 circRNAs, 5 miRNAs and 763 mRNAs, were identified.

Dysregulation of circRNA KIAA1586 Contributes to AD

In the AD risk circRNA-associated ceRNAs, the circRNA *KIAA1586* occurred most frequently. The AD risk circRNA-associated ceRNAs involving in circRNA *KIAA1586* were shown in **Figure 2** and details can be seen in **Supplementary File 4**. All of 159 genes and miRNAs were differentially expressed. There

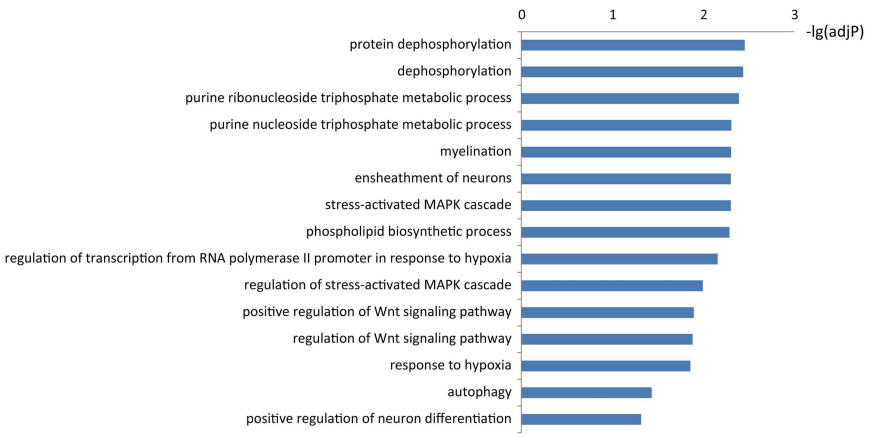


FIGURE 3 | The significantly enriched biological processes of *KIAA1586*-associated ceRNAs.

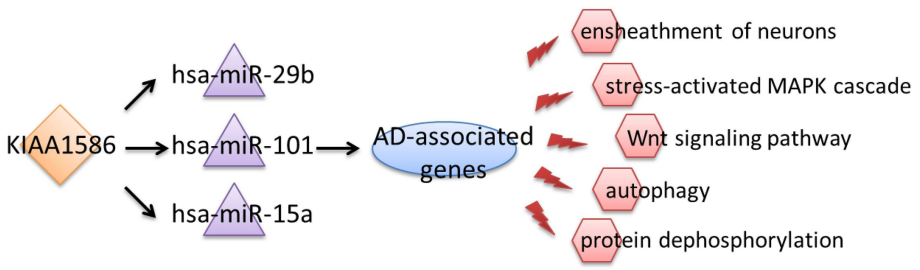


FIGURE 4 | A schematic diagram about circular RNAs, miRNAs, and AD pathologic mechanisms.

are 4 known AD-related genes and miRNAs (PSEN2, hsa-miR-29b, hsa-miR-15a, hsa-miR-101). The crosstalk among *KIAA1586* and AD-risk genes were mediated by hsa-miR-29b, hsa-miR-15a and hsa-miR-101. Thus, we speculate that hsa-miR-29b, hsa-miR-101 and hsa-miR-15a involved in competitive regulation in AD.

The amyloid precursor protein (APP) and its proteolytic product amyloid beta (Aβ) were closely associated with AD (Long and Lahiri, 2011; Long et al., 2019). Aberrant expression and function of miRNAs have been observed in AD and they have important roles in neuropathological conditions. Our results showed that *KIAA1586* could competitive binding three miRNAs (*hsa-miR-29b*, *hsa-miR-101*, and *hsa-miR-15a*). Pereira et al. (2016) demonstrated that *hsa-miR-29b* is decreased in AD patients displaying over-expression of *hBACE1* and subsequent Aβ peptide. Vilardo et al. (2010) revealed that the inhibition of *hsa-miR-101* increased APP levels and affected the accumulation of Aβ. Hebert et al. (2008) also found that *hsa-miR-15a* was significantly altered in AD brain and predicted that *hsa-miR-15a* regulates APP.

Three miRNAs regulated many genes related to AD. For example, Schlatterer et al. (2011) have shown c-Abl activation in AD and its activation in neuronal culture in response to Aβ fibrils and oxidative stress. Oddo (2012) indicated mTOR signaling lead to the progressive cognitive deficits characteristic of AD. And Peterson et al. (2014) uncovered that variants in *PPP3R1* were associated with rapid functional decline in AD. Thus, our results showed that the dysregulation of circRNA *KIAA1586* might disrupt the balance of three miRNA-related ceRNA networks and contribute to AD.

Function of KIAA1586-Associated ceRNAs

To learn about the biological functions of KIAA1586-associated ceRNAs, the GO function enrichment analysis (FDR $<$ 0.05) was performed for genes in the circRNA *KIAA1586*-related ceRNA network. We found that mRNAs in the circRNA *KIAA1586*-related ceRNA network were significantly enriched in known AD-associated biological processes (**Figure 3**, **Supplementary File 5**), such as Wnt signaling pathway (Boonen et al., 2009), protein dephosphorylation (Goedert et al., 1992), ensheathment of neurons (Morawski et al., 2010), stress-activated MAPK cascade (Morawski et al., 2010) and autophagy (Pickford et al., 2008).

Therefore, we inferred that aberrant expression of the circRNA KIAA1586 may be a key risk factor associated with the occurrence and development of AD. Through competitively binding to *hsa-miR-29b*, *hsa-miR-101* and *hsa-miR-15a*, the expressional perturbation of *KIAA1586* and the resultant changes in *KIAA1586*-associated ceRNA crosstalk interactions cause the abnormalities of AD-associated biological processes and contribute to the risk of AD (**Figure 4**).

DISCUSSION

AD is the most common type of dementia and one of the top 10 leading causes of death in the United States (Long et al., 2014; Gupta et al., 2017). The number of people living with AD is projected to nearly triple to 14 million people by 2060. It can seriously affect a person's ability to carry out daily activities. But the cause of AD is poorly understood.

The circRNAs are a class of non-coding RNAs highly expressed in the nervous system. Recent studies showed circRNA to be an important player in the development of neurodegenerative diseases like Alzheimer's disease. Many reports have demonstrated that circRNAs act as a kind of endogenous, competing, anti-complementary miRNA "sponge" to absorb and hence quench normal miRNA functions (Zhang et al., 2018).

In this study, we applied an integrated system biology approach to identify the circRNA-associated ceRNA network in AD. Combining expression information and shared miRNA number, the circRNA-associated ceRNAs were screened through strict threshold setting. We used the target genes of circRNAs to annotate their function through multiple testing (FDR $<$ 0.05). Moreover, through literature review the function annotation was further verified. Our results showed that circRNA *KIAA1586* might contribute to AD and its dysregulation could cause abnormal of AD-related biological functions. Further experimental studies should be conducted to uncover the functional roles of circRNA *KIAA1586* as a potential risk factor in the pathogenesis of Alzheimer. Our method will help to better understand the underlying molecular mechanisms of AD and our results also suggest that circRNA can be taken as a potential biomarker and therapeutic target in AD diagnosis and treatment. With the increasing in available circRNA expression profiles for AD and the accumulation of circRNA regulations or interactions, our method will become more powerful.

DATA AVAILABILITY

The data analyzed in this study were obtained from the GSE5281 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281), GSE1297 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1297), GSE12685 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12685) and GSE16759 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16759).

AUTHOR CONTRIBUTIONS

JS conceived and designed the experiments. YZ, FY, SB, and JS performed the experiments and analyzed the data. YZ and JS wrote the paper. All authors read and approved the final manuscript.

FUNDING

This study was supported by the Scientific Research Foundation for Talents of Wenzhou Medical University (QTJ18023, QTJ18024, QTJ18029). The funders had no roles 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/fbioe.2019.00222/full#supplementary-material

Supplementary File 1 | The consistent differentially expressed gene list.

Supplementary File 2 | The differentially expressed miRNA list.

Supplementary File 3 | The AD-related circRNA-associated ceRNA network.

Supplementary File 4 | The circRNA KIAA1586-associated ceRNAs.

Supplementary File 5 | The function of KIAA1586-associated ceRNAs.

REFERENCES

Arena, A., Iyer, A. M., Milenkovic, I., Kovacs, G. G., Ferrer, I., Perluigi, M., et al. (2017). Developmental expression and dysregulation of miR-146a and miR-155 in down's syndrome and mouse models of down's syndrome and Alzheimer's Disease. *Curr. Alzheimer Res*. 14, 1305–1317. doi: 10.2174/1567205014666170706112701

Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery, W. R., and Landfield, P. W. (2004). Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. U. S. A*. 101, 2173–2178. doi: 10.1073/pnas.0308512100

Boonen, R. A., van Tijn, P., and Zivkovic, D. (2009). Wnt signaling in Alzheimer's disease: up or down, that is the question. *Ageing Res. Rev*. 8, 71–82. doi: 10.1016/j.arr.2008.11.003

Goedert, M., Cohen, E. S., Jakes, R., and Cohen, P. (1992). p42 MAP kinase phosphorylation sites in microtubule-associated protein tau are dephosphorylated by protein phosphatase 2A1. Implications for Alzheimer's disease [corrected]. *FEBS Lett*. 312, 95–99.

Gupta, P., Bhattacharjee, S., Sharma, A. R., Sharma, G., Lee, S. S., and Chakraborty, C. (2017). miRNAs in Alzheimer disease - a therapeutic perspective. *Curr. Alzheimer Res*. 14, 1198–1206. doi: 10.2174/1567205014666170829101016

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Hebert, S. S., Horre, K., Nicolai, L., Papadopoulou, A. S., Mandemakers, W., Silahtaroglu, A. N., et al. (2008). Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression. *Proc. Natl. Acad. Sci. U. S. A*. 105, 6415–6420. doi: 10.1073/pnas.0710263105

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 37(Database issue), D98-104. doi: 10.1093/nar/gkn714

Jiang, W., Zhang, Y., Meng, F., Lian, B., Chen, X., Yu, X., et al. (2013). Identification of active transcription factor and miRNA regulatory pathways in Alzheimer's disease. *Bioinformatics* 29, 2596–2602. doi: 10.1093/bioinformatics/btt423

Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., et al. (2015). Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res*. 25, 981–984. doi: 10.1038/cr.2015.82

Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Walker, D. G., et al. (2007). Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genom.* 28, 311–322. doi: 10.1152/physiolgenomics.00208.2006

Liu, G., Jin, S., Hu, Y., and Jiang, Q. (2018a). Disease status affects the association between rs4813620 and the expression of Alzheimer's disease susceptibility gene TRIB3. *Proc. Natl. Acad. Sci. U. S. A*. 115, E10519–E10520. doi: 10.1073/pnas.1812975115

Liu, G., Jin, S., and Jiang, Q. (2019). Interleukin-6 receptor and inflammatory bowel disease: a mendelian randomization study. *Gastroenterology* 156, 823–824. doi: 10.1053/j.gastro.2018.09.059

Liu, G., Wang, T., Tian, R., Hu, Y., Han, Z., Wang, P., et al. (2018b). Alzheimer's disease risk variant rs2373115 regulates GAB2 and NARS2 expression in human brain tissues. *J. Mol. Neurosci*. 66, 37–43. doi: 10.1007/s12031-018-1144-9

Long, J. M., and Lahiri, D. K. (2011). MicroRNA-101 downregulates Alzheimer's amyloid-beta precursor protein levels in human cell cultures and is differentially expressed. *Biochem. Biophys. Res. Commun*. 404, 889–895. doi: 10.1016/j.bbrc.2010.12.053

Long, J. M., Maloney, B., Rogers, J. T., and Lahiri, D. K. (2019). Novel upregulation of amyloid-beta precursor protein (APP) by microRNA-346 via targeting of APP mRNA 5'-untranslated region: Implications in Alzheimer's disease. *Mol. Psychiatry* 24, 345–363. doi: 10.1038/s41380-018-0266-3

Long, J. M., Ray, B., and Lahiri, D. K. (2012). MicroRNA-153 physiologically inhibits expression of amyloid-beta precursor protein in cultured human fetal brain cells and is dysregulated in a subset of Alzheimer disease patients. *J. Biol. Chem*. 287, 31298–31310. doi: 10.1074/jbc.M112.366336

Long, J. M., Ray, B., and Lahiri, D. K. (2014). MicroRNA-339-5p down-regulates protein expression of beta-site amyloid precursor protein-cleaving enzyme 1 (BACE1) in human primary brain cultures and is reduced in brain tissue specimens of Alzheimer disease subjects. *J. Biol. Chem*. 289, 5184–5198. doi: 10.1074/jbc.M113.518241

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., et al. (2008). An analysis of human microRNA and disease associations. *PLoS ONE* 3:e3420. doi: 10.1371/journal.pone.0003420

Lukiw, W. J. (2013). Circular RNA (circRNA) in Alzheimer's disease (AD). *Front Genet* 4, 307. doi: 10.3389/fgene.2013.00307

Marot, G., Foulley, J. L., Mayer, C. D., and Jaffrézic, F. (2009). Moderated effect size and P-value combinations for microarray meta-analyses. *Bioinformatics* 25, 2692–2699. doi: 10.1093/bioinformatics/btp444

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, 333–338. doi: 10.1038/nature11928

Morawski, M., Brückner, G., Jäger, C., Seeger, G., and Arendt, T. (2010). Neurons associated with aggrecan-based perineuronal nets are protected against tau pathology in subcortical regions in Alzheimer's disease. *Neuroscience* 169, 1347–1363. doi: 10.1016/j.neuroscience.2010.05.022

Nunez-Iglesias, J., Liu, C. C., Morgan, T. E., Finch, C. E., and Zhou, X. J. (2010). Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation. *PLoS ONE* 5:e8898. doi: 10.1371/journal.pone.0008898

Oddo, S. (2012). The role of mTOR signaling in Alzheimer disease. *Front. Biosci*. 4, 941–952.

Pereira, P. A., Tomás, J. F., Queiroz, J. A., Figueiras, A. R., and Sousa, F. (2016). Recombinant pre-miR-29b for Alzheimer s disease therapeutics. *Sci. Rep*. 6:19946. doi: 10.1038/srep19946

Peterson, D., Munger, C., Crowley, J., Corcoran, C., Cruchaga, C., Goate, A. M., et al. (2014). Variants in PPP3R1 and MAPT are associated with more rapid functional decline in Alzheimer's disease: the cache county dementia progression study. *Alzheimers Dement*. 10, 366–371. doi: 10.1016/j.jalz.2013.02.010

Pickford, F., Masliah, E., Britschgi, M., Lucin, K., Narasimhan, R., Jaeger, P. A., et al. (2008). The autophagy-related protein beclin 1 shows reduced expression in early Alzheimer disease and regulates amyloid beta accumulation in mice. *J. Clin. Invest*. 118, 2190–2199. doi: 10.1172/JCI33585

Piwecka, M., GlaŽar, P., Hernandez-Miranda, L. R., Memczak, S., Wolf, S. A., Rybak-Wolf, A., et al. (2017). Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science* 357:eaam8526. doi: 10.1126/science.aam8526

Qu, S., Yang, X., Li, X., Wang, J., Gao, Y., Shang, R., et al. (2015). Circular RNA: A new star of noncoding RNAs. *Cancer Lett*. 365, 141–148. doi: 10.1016/j.canlet.2015.06.003

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., et al. (2015). Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell*. 58, 870–885. doi: 10.1016/j.molcel.2015.03.027

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., et al. (2010). GeneCards Version 3: the human gene integrator. *Database* 2010:baq020. doi: 10.1093/database/baq020

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, 353–358. doi: 10.1016/j.cell.2011.07.014

Schlatterer, S. D., Acker, C. M., and Davies, P. (2011). c-Abl in neurodegenerative disease. *J. Mol. Neurosci*. 45, 445–452. doi: 10.1007/s12031-011-9588-1

Vilardo, E., Barbato, C., Ciotti, M., Cogoni, C., and Ruberti, F. (2010). MicroRNA-101 regulates amyloid precursor protein expression in hippocampal neurons. *J. Biol. Chem*. 285, 18344–18351. doi: 10.1074/jbc.M110.112664

Welden, J. R., van Doorn, J., Nelson, P. T., and Stamm, S. (2018). The human MAPT locus generates circular RNAs. *Biochim. Biophys. Acta Mol. Basis. Dis*. 1864 (9 Pt B), 2753–2760. doi: 10.1016/j.bbadis.2018.04.023

Williams, C., Mehrian Shai, R., Wu, Y., Hsu, Y. H., Sitzer, T., Spann, B., et al. (2009). Transcriptome analysis of synaptoneurosomes identifies neuroplasticity genes overexpressed in incipient Alzheimer's disease. *PLoS ONE* 4:e4936. doi: 10.1371/journal.pone.0004936

Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., et al. (2017). RAID v2.0: an updated resource of RNA-associated interactions across organisms. *Nucleic Acids Res*. 45, D115–D118. doi: 10.1093/nar/gkw1052

Yu, F., Quan, F., Xu, J., Zhang, Y., Xie, Y., Zhang, J., et al. (2018a). Breast cancer prognosis signature: linking risk stratification to disease subtypes. *Brief Bioinform*. doi: 10.1093/bib/bby073. [Epub ahead of print].

Yu, F., Zhang, G., Shi, A., Hu, J., Li, F., Zhang, X., et al. (2018b). LnChrom: a resource of experimentally validated lncRNA-chromatin interactions in human and mouse. *Database* 2018:bay039. doi: 10.1093/database/bay039

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, Y., Li, X., Zhou, D., Zhi, H., Wang, P., Gao, Y., et al. (2018). Inferences of individual drug responses across diverse cancer types using a novel competing endogenous RNA network. *Mol. Oncol*. 12, 1429–1446. doi: 10.1002/1878-0261.12181

Zhang, Y., Liu, D., Wang, L., Wang, S., Yu, X., Dai, E., et al. (2015). Integrated systems approach identifies risk regulatory pathways and key regulators in coronary artery disease. *J. Mol. Med*. 93, 1381–1390. doi: 10.1007/s00109-015-1315-x

Zhou, M., Guo, M., He, D., Wang, X., Cui, Y., Yang, H., et al. (2015a). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J. Transl. Med*. 13, 231. doi: 10.1186/s12967-015-0556-3

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018a). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Wang, X., Shi, H., Cheng, L., Wang, Z., Zhao, H., et al. (2016). Characterization of long non-coding RNA-associated ceRNA network to reveal potential prognostic lncRNA biomarkers in human ovarian cancer. *Oncotarget* 7, 12598–12611. doi: 10.18632/oncotarget.7181

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018b). An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. *Mol. Neurobiol.* 55, 3684–3697. doi: 10.1007/s12035-017-0572-9

Zhou, M., Zhang, Z., Zhao, H., Bao, S., and Sun, J. (2018c). A novel lncRNA-focus expression signature for survival prediction in endometrial carcinoma. *BMC Cancer* 18, 39. doi: 10.1186/s12885-017-3983-0

Zhou, M., Zhao, H., Wang, Z., Cheng, L., Yang, L., Shi, H., et al. (2015b). Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. *J. Exp. Clin. Cancer Res*. 34:102. doi: 10.1186/s13046-015-0219-5

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

Conflict of Interest Statement: 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 © 2019 Zhang, Yu, Bao and Sun. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 15 October 2019
doi: 10.3389/fbioe.2019.00270

Identification of Prognostic and Metastatic Alternative Splicing Signatures in Kidney Renal Clear Cell Carcinoma

Tong Meng[1,2,3†], Runzhi Huang[1,2†], Zhiwei Zeng[4], Zongqiang Huang[4], Huabin Yin[3], ChenChen Jiao[5], Penghui Yan[4], Peng Hu[4], Xiaolong Zhu[4], Zhenyu Li[6], Dianwen Song[3], Jie Zhang[1,2,6*] and Liming Cheng[1,2*]*

[1] Division of Spine, Department of Orthopedics, Tongji Hospital affiliated to Tongji University School of Medicine, Shanghai, China, [2] Key Laboratory of Spine and Spinal cord Injury Repair and Regeneration, Tongji University, Ministry of Education, Shanghai, China, [3] Department of Orthopedics, School of Medicine, Shanghai General Hospital, Shanghai Jiaotong University, Shanghai, China, [4] Department of Orthopedics, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [5] Department of Pathology, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China, [6] Department of Prevention, Tongji University School of Medicine, Tongji University, Shanghai, China

Background: Kidney renal clear cell carcinoma (KIRC) is the malignancy originated from the renal epithelium, with a high rate of distant metastasis. Aberrant alternative splicing (AS) of pre-mRNA are widely reported to be involved in the tumorigenesis and metastasis of multiple cancers. The aim of this study is to explore the mechanism of alternative splicing events (ASEs) underlying tumorigenesis and metastasis of KIRC.

Methods: RNA-seq of 537 KIRC samples downloaded from the TCGA database and ASEs data from the TCGASpliceSeq database were used to identify ASEs in patients with KIRC. The univariate and Lasso regression analysis were used to screen the most significant overall survival-related ASEs (OS-SEs). Based on those, the OS-SEs model was proposed. The interaction network of OS-SEs and splicing factors (SFs) with absolute value of correlation coefficient value >0.750 was constructed by Pearson correlation analysis. The OS-SEs significantly related to distant metastasis and clinical stage were identified by non-parametric test, and those were also integrated into co-expression analysis with prognosis-related Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways identified by Gene Set Variation Analysis (GSVA). ASEs with significance were selected for multiple online database validation.

Results: A total of prognostic 6,081 overall survival-related ASEs (OS-SEs) were identified by univariate Cox regression analysis and a prediction model was constructed based on 5 OS-SEs screened by Lasso regression with the Area Under Curve of 0.788. Its risk score was also illustrated to be an independent predictor, which the good reliability of the model. Among 390 identified candidate SFs, DExD-Box Helicase 39B (DDX39B) was significantly correlated with OS and metastasis. After external database validation, Retained Intron of Ras Homolog Family Member T2 (RHOT2) and T-Cell Immune Regulator 1 (TCIRG1) were identified. In the co-expression analysis, overlapped co-expression signal pathways for RHOT2 and TCIRG1 were sphingolipid metabolism and N-glycan biosynthesis.

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Fang Zhong,
Icahn School of Medicine at Mount Sinai, United States
Cong Jiang,
Harvard University, United States

***Correspondence:**
Dianwen Song
osongdianwen@126.com
Jie Zhang
jiezhang@tongji.edu.cn
Liming Cheng
limingcheng@tongji.edu.cn

[†]These authors have contributed equally to this work as co-first authors

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 02 September 2019
Accepted: 30 September 2019
Published: 15 October 2019

Citation:
Meng T, Huang R, Zeng Z, Huang Z, Yin H, Jiao C, Yan P, Hu P, Zhu X, Li Z, Song D, Zhang J and Cheng L (2019) Identification of Prognostic and Metastatic Alternative Splicing Signatures in Kidney Renal Clear Cell Carcinoma. Front. Bioeng. Biotechnol. 7:270. doi: 10.3389/fbioe.2019.00270

Conclusions: Based on the results of comprehensive bioinformatic analysis, we proposed that aberrant DDX39B regulated RHOT2-32938-RI and TCIRG1-17288-RI might be associated with the tumorigenesis, metastasis, and poor prognosis of KIRC via sphingolipid metabolism or N-glycan biosynthesis pathway.

Keywords: alternative splicing, kidney renal clear cell carcinoma, prognosis, tumor metastasis, RHOT2, TCIRG1

INTRODUCTION

Kidney renal clear cell carcinoma (KIRC) is a malignant cancer originated from renal epitheliums, accounting for about 75% of kidney tumors (Hsieh et al., 2017). Therapeutically, although radical nephrectomy is performed for localized renal masses, distant metastasis may be observed in a large proportion of patients at diagnosis, especially metastasis in lung, bone and brain (Gupta et al., 2008). With regard to these advanced KIRCs, the treatment option was limited with only sunitinib widely approved (Porta et al., 2019). Even systematic therapy were applied, including immunotherapeutic agents, antiangiogenic agents and mTOR inhibitors, the prognosis was still poor (Jonasch et al., 2014; Jonasch, 2018). In order to prolong the overall survival of patients with KIRC, there is a pressing need to explore its pathogenic mechanism and identify the potential therapeutic targets related to tumorigenesis, metastasis and prognosis.

Nowadays, most studies of KIRC focused on alteration of transcriptome level and the posttranscriptional process was largely underestimated. Alternative splicing (AS), plays an important role in the maturation of mRNAs from its precursors, leading to diverse mRNA isoforms spliced and protein variants translated (Montes et al., 2019). In this process, splicing factors (SFs) work as regulatory catalyst of alternative splicing events (ASEs) and both build up an intricate regulatory network (Frankiw et al., 2019; Wu et al., 2019). Functionally, AS has been reported to take part in cell differentiation, lineage determination and tissue-specificity acquisition (Wang et al., 2008). The aberrant AS of some genes and somatic mutations of SFs, which make network dysregulated, have been shown to modulate malignant transformation of cells and epithelial-mesenchymal transition (EMT) (Sveen et al., 2016; Kouyama et al., 2019; Wu et al., 2019; Xing et al., 2019). Thus, identifying the dysregulated network may shed light upon the molecular biomarkers for prognosis, metastasis, and potential therapeutic targets (Lee and Abdel-Wahab, 2016; Zhou et al., 2016; Wang et al., 2019).

Nowadays, although a systematic analysis of ASEs was unveiled in KIRC, the regulatory network of ASEs and SFs was not explored (Song et al., 2019). Additionally, metastasis-related ASEs, and potential therapeutic targets were also underestimated. In this study, we performed a comprehensive analysis of AS profiling to identify the overall survival-related ASEs (OS-SEs) in patients with KIRC and construct a prognostic model. Additionally, metastasis-related ASEs along with corresponding SFs and pathways were also identified by Pearson correlation analysis to illuminate the underlying mechanism of metastasis in KIRC. The prediction model might assist oncologists in clinical decision-making. Moreover, we also identified a new candidate molecular mechanism and two potential therapeutic targets for KIRC metastasis treatment, especially to the bone metastasis.

Abbreviations: KIRC, Kidney renal clear cell carcinoma; AS, Alternative splicing; ASEs, alternative splicing events; OS-SEs, overall survival-related ASEs; DDX39B, DExD-Box Helicase 39B; RHOT2, Retained Intron of Ras Homolog Family Member T2; TCIRG1, T-Cell Immune Regulator 1; SFs, splicing factors; EMT, epithelial-mesenchymal transition; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; mAb, monoclonal antibody; S1P, sphingosine-1-phosphate; TCGA, the Cancer Genome Atlas; PSI, percent splicing; GSVA, Gene Set Variation Analysis; IHC, Immunohistochemistry; AUC, Area under curve.

MATERIALS AND METHODS

Data Collection

Clinical information, RNA sequencing profiles, and SFs of KIRC samples were collected from the Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov). Meanwhile, ASEs data were retrieved from the TCGASpliceSeq database (https://bioinformatics.mdanderson.org/TCGASpliceSeq/) (Ryan et al., 2016) including seven types (alternate acceptor site, AA; exon skip, ES; alternate terminator, AT; mutually exclusive exons, ME; retained intron, RI; alternate donor site, AD; alternate promoter, AP) (Chen et al., 2019). Samples with more than 25 percent of missing percent splicing (PSI) values were excluded. The ASE was presented with gene name, ID number from the TCGASpliceSeq database (AS ID) and splicing pattern.

The Identification of OS-SEs

The K-Nearest Neighbor algorithm was performed to impute ASEs with missing expression data. ASEs, whose means and standard deviations of PSI < 0.05 and 0.01, were excluded, neither were samples without follow-up records. Then, the combined ASEs along with clinical data were put into the univariate Cox regression analysis to evaluate the prognostic value of each filtered ASE. The UpSet plot was developed to illustrate OS-SEs and volcano plot was used to display the prognosis-related and -unrelated ASEs integrally. The bubble plots were generated to present the top 20 OS-SEs for seven types of ASEs, in which the color and size of bubbles symbolize the value of ASEs for overall survival.

The Construction of the Prognostic Model Based on the OS-SEs

The Lasso regression was firstly performed to screen the top 20 significant prognostic OS-SEs and then the significant prognostic OS-SEs were evaluated by the multivariate Cox regression model

with β value, which represented the regression coefficient of each integrated OS-SE in the model. Risk score was thus acquired by the following formula:

$$\sum_{i=1}^{n} \beta i \times \mathrm{PSI}$$

According to the median risk score, samples were divided into two risk groups medially. The area under receiver operating characteristic ROC curve was used to evaluate the accuracy of the model. In addition, Kaplan-Meier survival analysis was also conducted to compare the difference between high- and low-risk group. Samples were reordered according to risk score and then the risk curve, scatterplot and expression heatmap were generated.

The univariate and multivariate Cox regression analysis, modified by baseline information, were applied to evaluate the prognostic role of risk score, along with age, gender, grade, clinical stage, and TNM stage.

The Construction of the Interaction and Correlation Network

In the SpliceAid2 database, 390 splicing factors were retrieved (Piva et al., 2009). Pearson correlation analysis was performed to explore the interaction and correlation between SFs and OS-SEs. The regulation network of SFs and OS-SEs was plotted by Cytoscape (3.7.1) (Shannon et al., 2003), in which the regulation pairs with $P > 0.001$ and the absolute value of correlation coefficient < 0.750 were excluded. In the network, we defined SF and OS-SEs as arrows and ellipses, high and low risk of OS-SEs as red and purple, positive and negative regulations as red and green lines, respectively.

The Identification of Metastasis- and/or Stage-Related OS-SEs

To identify the OS-SEs related to metastasis and/or TNM stage, we performed Kruskal-Wallis test and Mann-Whitney-Wilcoxon test, which were displayed by beeswarm plots. Besides, the regulation network of these metastasis-, and/or stage- related OS-SEs were also explored.

The Co-expression Analysis Between ASEs and Signaling Pathways

The univariate Cox analysis was performed to screen the prognosis-related signaling pathways identified by Gene Set Variation Analysis (GSVA) (Hänzelmann et al., 2013). Then, metastasis and stage-related OS-SEs and prognosis-related

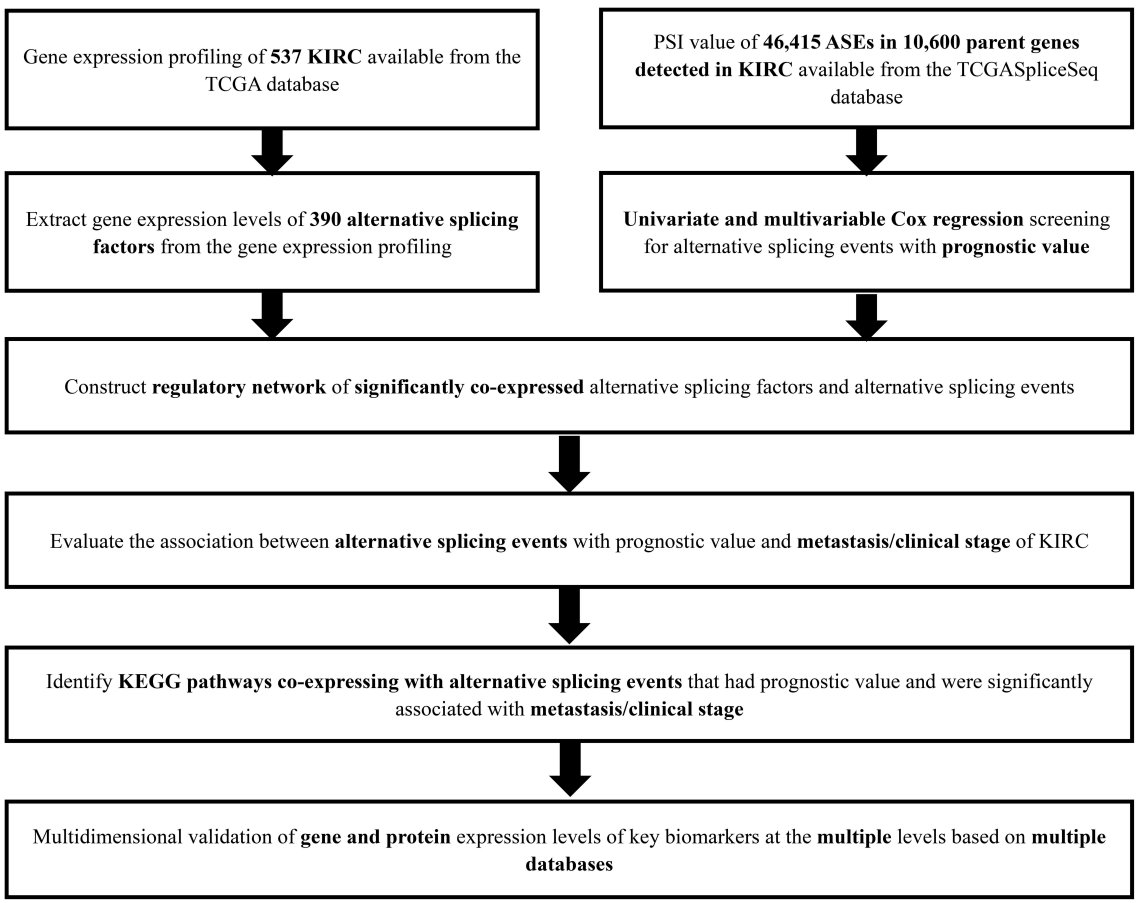


FIGURE 1 | The flowchart of this study.

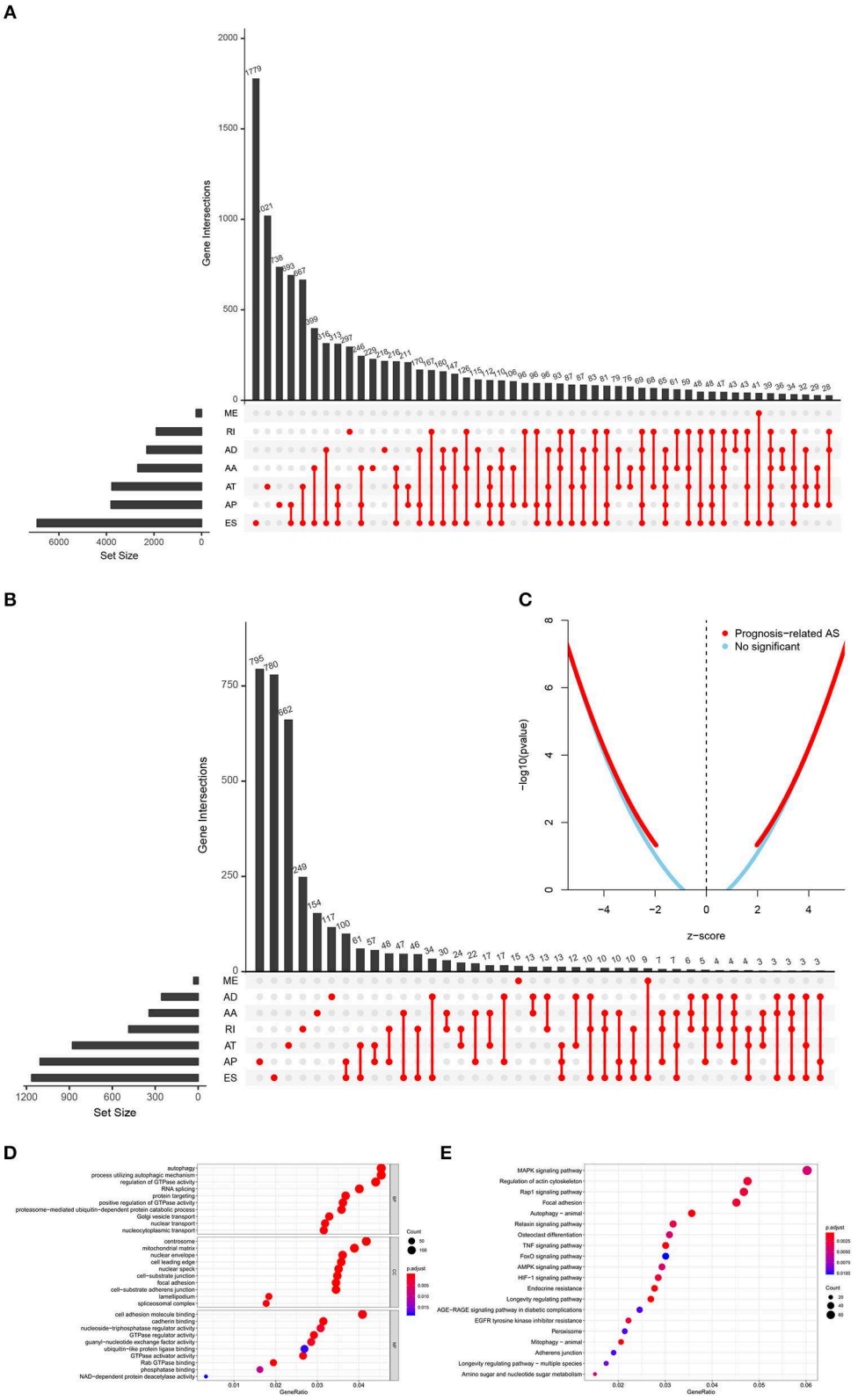


FIGURE 2 | The identification of OS-SEs in KIRC patients. The Upset plots of ASEs and OS-SEs: **(A)** The number of ASEs in different types of splicing patterns; **(B)** The number of OS-SEs in different types of splicing patterns. **(C)** The volcano plot of the prognosis-related and no significant ASEs, respectively; The GO analysis *(Continued)*

FIGURE 2 | (D) and the KEGG pathways enrichment analysis **(E)** of the parent genes of OS-ASEs. ASEs, Alternative splicing events; OS-SEs, overall survival-related ASEs; KIRC, kidney renal clear cell carcinoma; AA, alternate acceptor; AD, alternate donor; AP, alternate promoter; AT, alternate terminator; ES, exon skip; ME, mutually exclusive exons; RI, retained intron; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

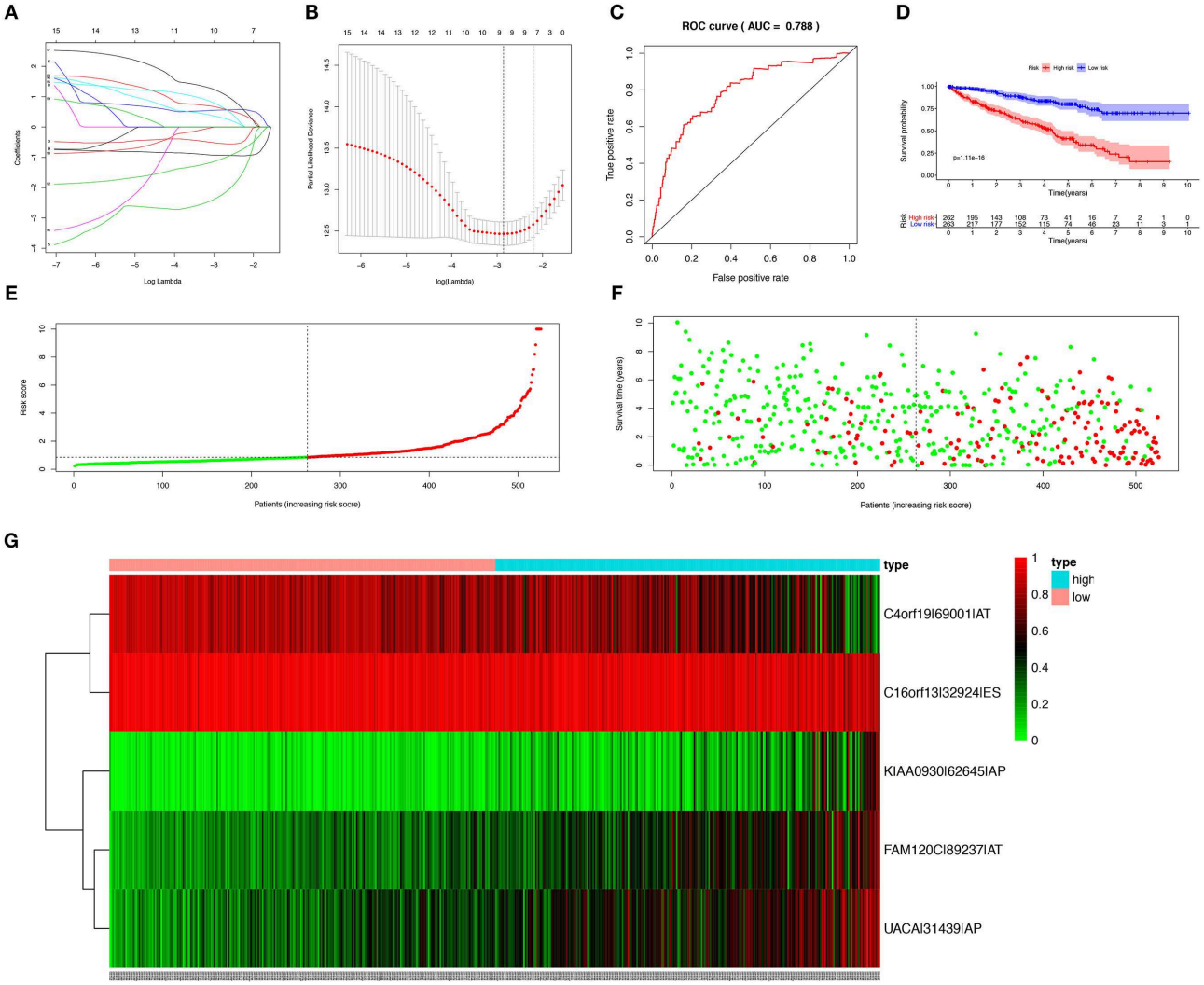


FIGURE 3 | Establishment and assessment of the predict model. **(A)** The coefficients in the Lasso regression for OS-SEs screening; **(B)** Cross-validation for tuning parameter selection in the proportional hazards model; **(C)** The ROC curve for assessing the reliability of the predict model; **(D)** The Kaplan-Meier curve of the predict model; **(E)** The risk curve of each sample reordered by risk score; **(F)** The scatter plot of the samples. The green and red dots representing survival and death, respectively; **(G)** The heatmap of expression level of 5 OS-SEs filtered by Lasso regression. OS-SEs, overall survival-related ASEs; ROC, receiver operating characteristic.

KEGG pathways were put into the co-expression analysis to identify the possible downstream mechanism of OS-SEs.

Online Database Validation

In order to ensure the roles of selected metastasis and stage-related OS-SEs, multiple databases including the UALCAN (Chandrashekar et al., 2017), UCSC Treehouse Childhood Cancer Initiative, Kaplan Meier plotter (Nagy et al., 2018), LinkedOmics (Vasaikar et al., 2018), SurvExpress (Aguirre-Gamboa et al., 2013) and Firebrowse (Deng et al., 2017) were used to detect their gene and protein expression levels in KIRC and normal kidney tissues.

Immunohistochemistry (IHC)

The IHC slides and information were obtained from the Human Protein Atlas. Immunostaining on each slide was assessed by experienced pathologists to examine the percentage of RHOT2 and TCIRG1 positive tumor cells and presented as histochemistry score (H-score). H-score = Σpi(i+1) where i is the intensity score and pi is the percent of the cells with that intensity.

Statistical Analysis

All statistical analysis was applied by R version 3.5.1 (Institute for Statistics and Mathematics, Vienna, Austria; https://www.r-project.org) (Package: impute, UpSetR, ggplot2, rms, glmnet, preprocessCore, forestplot, survminer, survivalROC, beeswarm). For descriptive statistics, mean ± standard deviation was used for the continuous variables in normal distribution while the median (range) was used for continuous variables in abnormal distribution. Categorical variables were described by counts and percentages. Two-tailed $P < 0.05$ was regarded statistically significant.

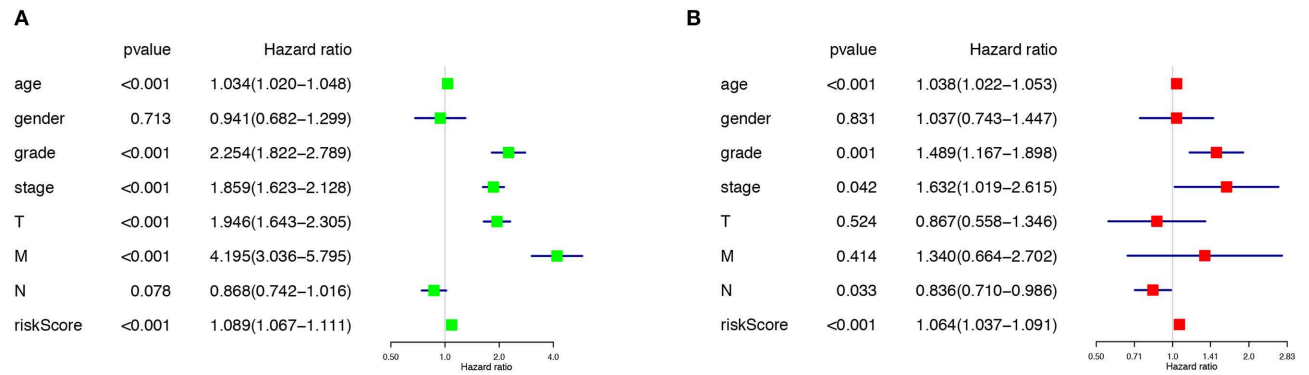


FIGURE 4 | The Cox regression analysis for evaluating the independent prognostic value of the risk score. The univariate **(A)** and multivariate **(B)** Cox regression analysis of risk score, age, gender, grade, and TNM stage.

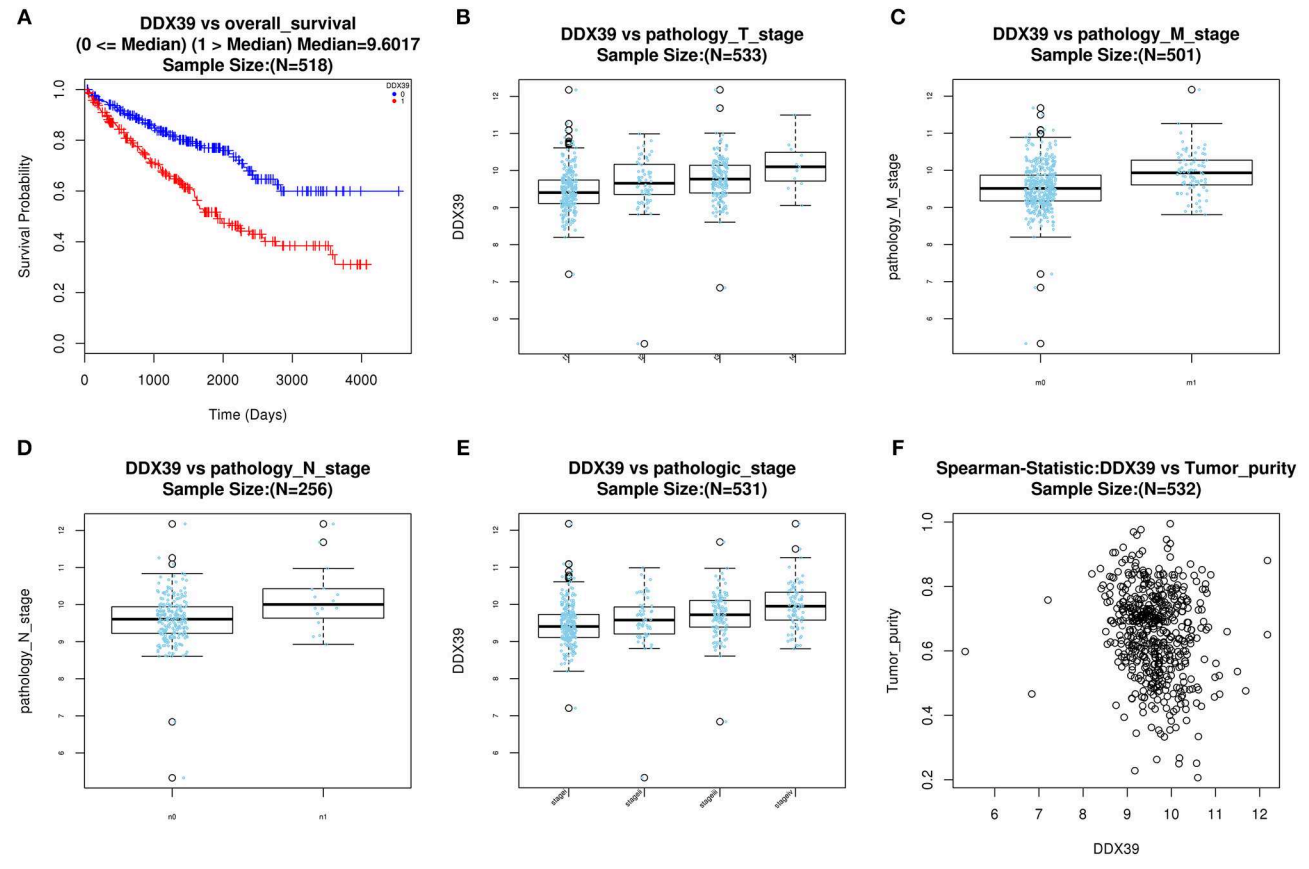


FIGURE 5 | Evaluate prognostic value and clinical correlation of splicing factor DDX39B (Gene symbol: DDX39). **(A)** The Kaplan-Meier curve of DDX39; The expressions of DDX39 according to T **(B)**, N **(C)**, M **(D)** staging system, and clinical stage **(E)**; **(F)** The spearman correlation analysis of DDX39 and tumor purity. DDX39B, DExD-Box Helicase 39B.

RESULTS

Overview of ASEs and OS-SEs in KIRC

The analysis process was presented in the flow chart (**Figure 1**). The sequencing data of 537 cases KIRC were downloaded from the TCGA database, with the median overall survival of 1,091 (range, 0–3,668) days. Throughout the follow-up period, 165 patients died and 496 experienced tumor metastases. A total of 46,415 ASEs in 10,600 parent genes were detected in patients with KIRC, including 3,821 AAs (2,683 genes), 3,270 ADs (2,300 genes), 9,509 APs (3,805 genes), 8,632 ATs (3,770 genes), 18,117 ESs (6,915 genes), 235 MEs (227 genes), and 2,831 RIs (1,902 genes). Thus, one gene could undergo more than 4 splicing patterns (**Figure 2A**). Among the seven types of ASEs, ES was the most prevalent one, followed by AT. A total of 6,081 OS-ASEs from 3,444 parent genes were identified and the UpSet plot revealed that AP was the most common splicing patterns associated with KIRC prognosis (**Figure 2B**). The volcano plot suggested that most of ASEs were OS-SEs in KIRC (**Figure 2C**). The top 20 OS-ASEs in seven types of splicing patterns were illustrated in bubble plots (**Supplementary Figures 1A–G**).

Functional Enrichment Analysis of Prognostic AS Events

In order to illuminate the potential mechanism underlying the OS-ASEs, 2,077 parent genes of the 6,081 OS-ASEs in KIRC were sent for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis (**Figures 2D,E**). The biological process of GO analysis revealed the enrichment of some well-known pathways, in relation to "autophagy", "process utilizing autophagic mechanism" and "regulation of GTPase activity" (**Figure 2D**). Besides, the RNA splicing was also enriched significantly, which meant the active aberrant splicing patterns of KIRC. Additionally, "centrosome" "mitochondrial matrix" and "cell adhesion molecule binding" were also significantly enriched as cellular component or molecular function. The KEGG enrichment analysis suggested some key pathways were associated with the OS of patients with KIRC, such as "MAPK signaling pathway", "Regulation of actin cytoskeleton", "Rap1 signaling pathway" and "Focal adhesion" (**Figure 2E**).

Establishment of the Prediction Model

In order to avoid over-fitting of the predict model, the Lasso regression was performed to screen the top 20 OS-SEs. The result showed that C4orf19-69001-AT, C16orf13-32924-ES, KIAA0930-62645-AP, FAM120C-89237-AT, UACA-31439-AP were included in the multivariate Cox regression analysis (**Figures 3A,B**), with Area Under Curve (AUC) of 0.788 in ROC curve (**Figure 3C**). Accordingly, risk score of each sample was calculated, with a median value of 0.853. Then, Kaplan-Meier curve revealed

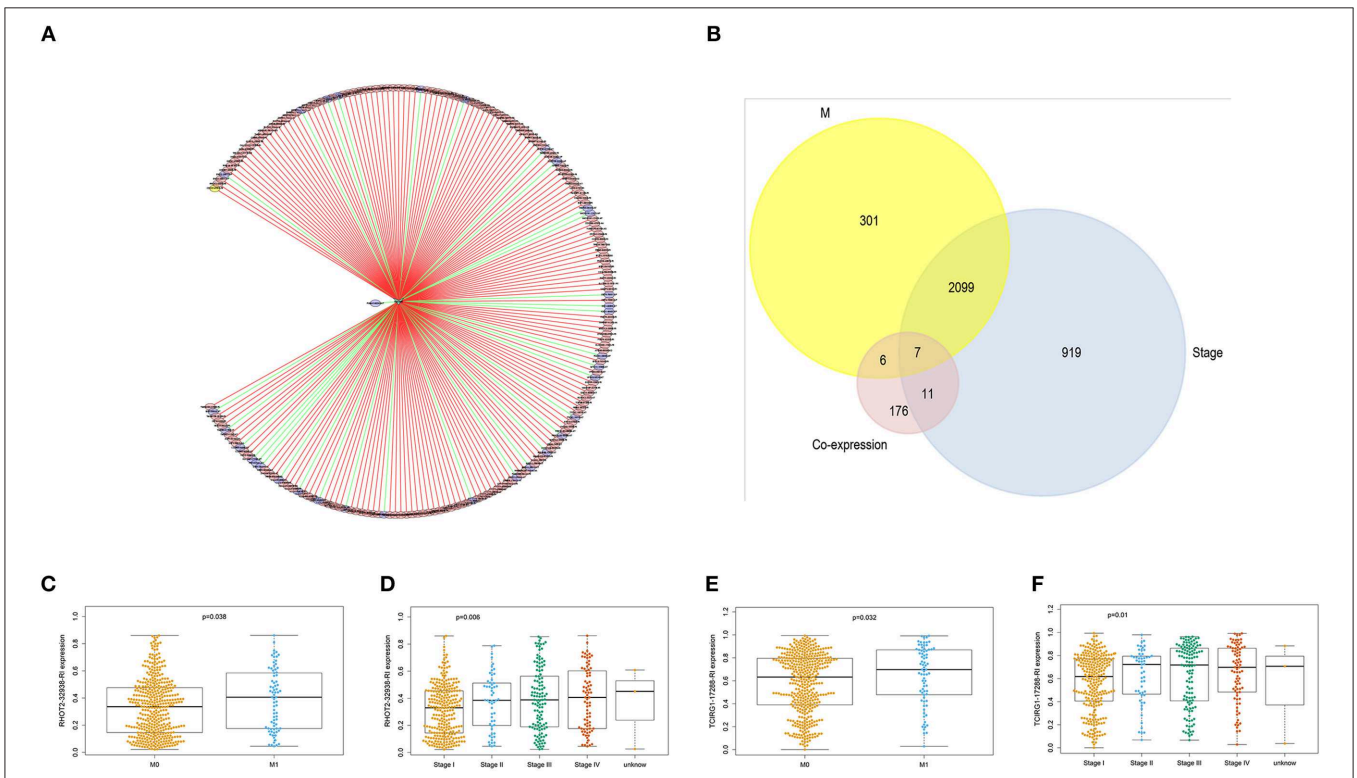


FIGURE 6 | The identification of metastasis- and/or stage-related OS-SEs. **(A)** The network of OS-SEs and prognosis-related SFs; **(B)** The Venn plot to identify the overlapped OS-SEs related to clinical status and metastasis; The beeswarm plots of RHOT2-32938-RI **(C)** and TCIRG1-17288-RI **(D)** according to metastasis or not; The beeswarm plots of RHOT2-32938-RI **(E)**, and TCIRG1-17288-RI **(F)** according to clinical status. OS-SEs, overall survival-related ASEs; SFs, splicing factors; RHOT2, Retained Intron of Ras Homolog Family Member T2; TCIRG1, T-Cell Immune Regulator 1.

TABLE 1 | The external validation of CALCOCO1, CIRBP, P4HTM, RHOT2, TBC1D17, TCIRG1 and THOP1.

Database		CALCOCO1 (anti-oncogene)	CIRBP (anti-oncogene)	P4HTM (anti-oncogene)	RHOT2 (oncogene)	TBC1D17 (oncogene)	TCIRG1 (oncogene)	THOP1 (oncogene)
UALCAN		Expression: $p < 0.001$ Stage: $p < 0.001$ K-M analysis: $P = 0.010$	Expression: $p = 0.023$ Stage: $p < 0.001$ K-M analysis: $p = 0.001$	Expression: $p < 0.001$ Stage: $p < 0.001$ K-M analysis: $p = 0.240$	Expression: $p < 0.001$ Stage: $p < 0.001$ K-M analysis: $p < 0.001$	Expression: $p = 0.010$ Stage: $p < 0.001$ K-M analysis: $p = 0.250$	Expression: $p < 0.001$ Stage: $p < 0.001$ K-M analysis: $p < 0.001$	Expression: $p < 0.001$ Stage: $p < 0.001$ K-M analysis: $p = 0.410$
The human protein atlas		Tumor median Normal high K-M analysis $P < 0.001$	Tumor low Normal high K-M analysis $P < 0.001$	Tumor not detected Normal median K-M analysis $P < 0.001$	Tumor high Normal median K-M analysis $P < 0.001$	Tumor median Normal median K-M analysis $P < 0.001$	Tumor high Normal median K-M analysis $P < 0.001$	Tumor high Normal high K-M analysis $P < 0.001$
Kaplan Meier plotter	Best cutoff	$P < 0.001$	$P < 0.001$	$P = 0.004$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P = 0.008$
	Median value	$P = 0.334$	$P = 0.121$	$P = 0.168$	$P = 0.011$	$P = 0.216$	$P < 0.001$	$P = 0.193$
LinkedOmics		K-M analysis: $P = 0.364$ M: $P = 0.128$ Stage: $P = 0.047$ $R = 0.541$	K-M analysis: $P = 0.044$ M: $P < 0.001$ Stage: $P < 0.001$ $R = 0.545$	K-M analysis: $P = 0.371$ M: $P = 0.630$ Stage: $P = 0.759$ $R = 0.185$	K-M analysis: $P < 0.001$ M: $P = 0.130$ Stage: $P = 0.072$ $R = 0.575$	K-M analysis: $P = 0.021$ M: $P = 0.692$ Stage: $P = 0.930$ $R = 0.469$	K-M analysis: $P < 0.001$ M: $P < 0.001$ Stage: $P < 0.001$ $R = 0.447$	K-M analysis: $P < 0.001$ M: $P = 0.034$ Stage: $P = 0.052$ $R = 0.163$
SurvExpress		K-M analysis: $P = 0.080$	K-M analysis: $P = 0.450$	K-M analysis: $P = 0.083$	K-M analysis: $P = 0.006$	K-M analysis: $P = 0.072$	K-M analysis: $P < 0.001$	K-M analysis: $P = 0.022$
Firebrowse		Fold change: 0.885	Fold change: 0.965	Fold change: 0.768	Fold change: 1.31	Fold change: 0.958	Fold change: 2.71	Fold change: 1.47

calcium binding and coiled-coil domain 1 (CALCOCO1); cold inducible RNA binding protein (CIRBP); prolyl 4-hydroxylase, transmembrane (P4HTM); Retained Intron of Ras Homolog Family Member T2 (RHOT2); TBC1 domain family member 17 (TBC1D17); T-Cell Immune Regulator 1 (TCIRG1); thimet oligopeptidase 1 (THOP1).

that prediction model of risk score had a good effectiveness ($P < 0.001$) (**Figure 3D**).

Risk curve and scatterplot were generated to show the risk score and vital status of each patient with KIRC. Patient in high-risk group had a higher mortality than patient in low-risk group (**Figures 3E,F**). The heatmap showed the expressions of OS-SEs screened by Lasso regression, indicating that C40rf19-69001-AT and C16orf13-32924-ES were lower and KIAA0930-62645-AP, FAM120C-89237-AT, UACA-31439-AP were higher in high-risk group (**Figure 3G**).

The Risk Score Predicted Prognosis

The risk score along with age, gender, grade and TNM stage were evaluated in the univariate and multivariate Cox regression analysis. The risk score was confirmed as an independent predictor in both univariate (HR = 1.089, 95%CI (1.067–1.111), $P < 0.001$), and multivariate Cox regression analysis (HR = 1.064, 95%CI (1.037–1.091), $P < 0.001$), **Figures 4A,B**).

The potential splicing regulatory network of SFs and OS-SEs, and their metastasis or clinical stage correlation.

With access to RNA-seq data and corresponding clinical information of patients with KIRC, we identified 390 candidate SFs whose expression levels were significantly associated with OS of KIRC patients. Among them, DExD-Box Helicase 39B (DDX39B) was the only SF, who was correlated with prognosis (**Figure 5A**), TNM staging system (**Figures 5B–D**), clinical stage (**Figure 5E**), and tumor purity (**Figure 5F**). Then, a network was established to demonstrate the interaction and correlation between SFs and OS-SEs. DDX39B was correlated with 34 favorable OS-SEs (purple ellipses) negatively (green lines) and 166 adverse OS-SEs (red ellipses) positively (red lines) (**Figure 6A**). Among them, 7 OS-SEs (CALCOCO1-22108-RI, CIRBP-46432-RI, P4HTM-64788-ES, RHOT2-32938-RI, TBC1D17-51116-ES, TCIRG1-17288-RI, THOP1-46623-AP) were significantly related to both metastasis and stage in the Venn plot (**Figures 6B–F**, **Supplementary Figures 2A–J**).

External Validation

The parent genes of the 7 OS-SEs were validated in external databases. RHOT2 and TCIRG1 were confirmed in all the external databases (**Table 1**). In the database of UALCAN and LinkedOmics, RHOT2, and TCIRG1 were up-regulated in tumor than normal tissue (**Figure 7A**, **Supplementary Figure 3A**) and associated with tumor stage (**Figure 7B**, **Supplementary Figure 3B**) and OS (**Figure 7C**, **Supplementary Figure 3C**) significantly. In the Kaplan Meier plotter and SurvExpress, RHOT2, and TCIRG1 were also associated with OS significantly (**Figure 7D**, **Supplementary Figure 3D**). The survicalROC was also described in **Figure 7E**. Validation in the Human Protein Atlas revealed that the protein levels of RHOT2 and TCIRG1 in KIRC were significantly higher than those in normal kidney tissue (**Figures 7F,F′**, **Table 2**).

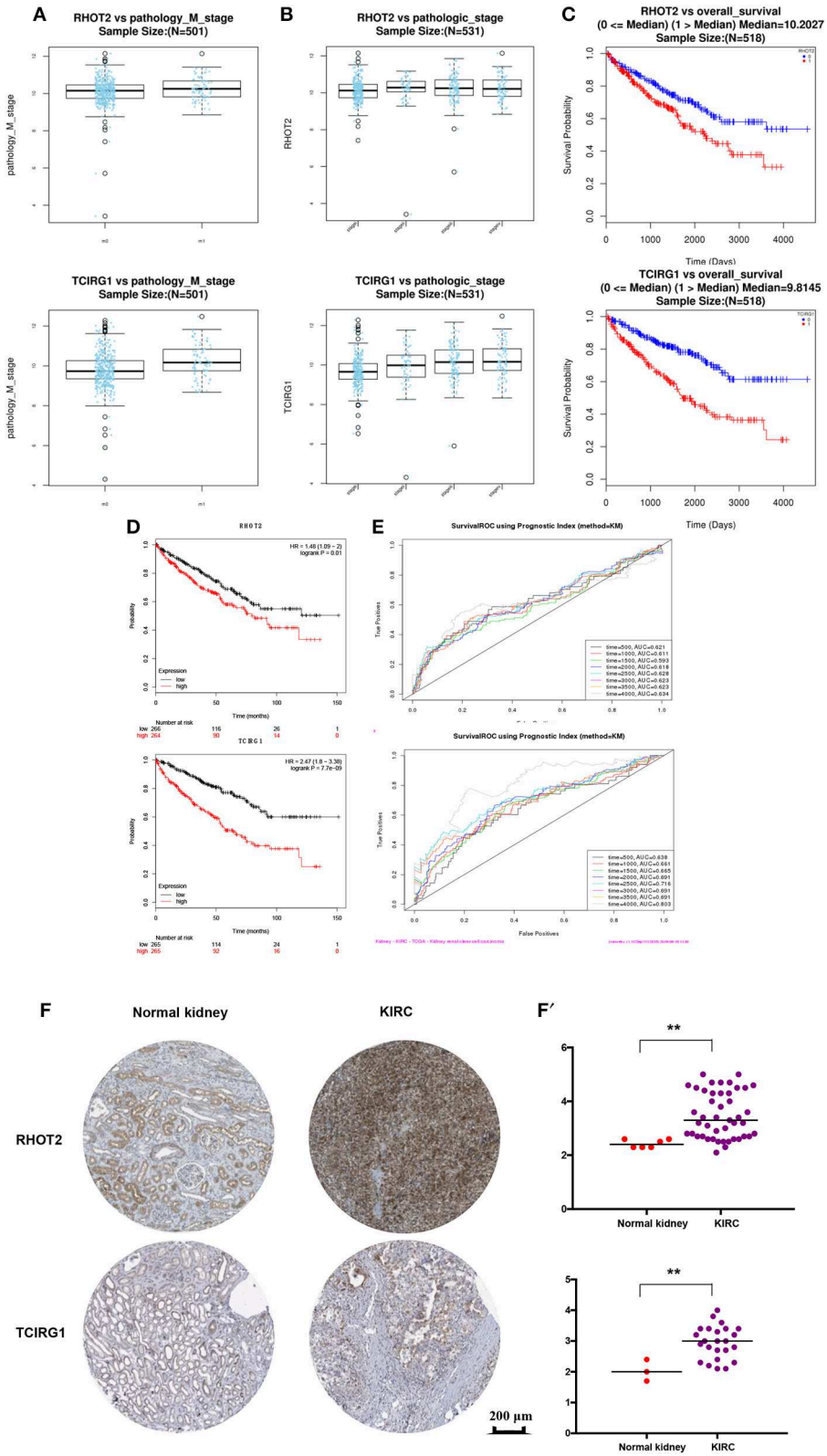


FIGURE 7 | The external validation of RHOT2 and TCIRG1: **(A)** The expressions of RHOT2 and TCIRG1 between normal kidney and KIRC in UALCAN; **(B)** The expressions of RHOT2 and TCIRG1 according to clinical stage in UALCAN; The Kaplan-Meier curve of RHOT2 and TCIRG1 in LinkedOmics **(C)** and Kaplan Meier

(Continued)

FIGURE 7 | plotter **(D)**; **(E)** The survicalROC of RHOT2 and TCIRG1; The IHC **(F)** and H-score **(F')** of RHOT2 and TCIRG1 between normal kidney and KIRC in the Human Protein Atlas. RHOT2, Retained Intron of Ras Homolog Family Member T2; TCIRG1, T-Cell Immune Regulator 1; KIRC, kidney renal clear cell carcinoma; IHC, Immunohistochemistry; H-score, histochemistry score.

TABLE 2 | The mean H-score of RHOT2 and TCIRG1 in Normal kidney and KIRC.

Biomarker	Normal kidney	KIRC	*p*
RHOT2	2.43	3.48	0.005
TCIRG1	2.03	2.96	0.008

Retained Intron of Ras Homolog Family Member T2 (RHOT2); T-Cell Immune Regulator 1 (TCIRG1); Kidney renal clear cell carcinoma (KIRC).

Comprehensive Analysis of ASEs and Signaling Pathways

A total of 90 KEGG pathways were identified as the OS-related KEGG pathways in GSVA and the univariate Cox regression analysis (**Figure 8A**). As shown, RHOT2–32938–RI was associated with sphingolipid metabolism, N-glycan biosynthesis and glycosphingolipid biosynthesis lacto and neolacto series. TCIRG1-17288-RI was associated with sphingolipid metabolism, purine metabolism and N-glycan biosynthesis (**Figure 8B**).

DISCUSSION

KIRC, one of the most prevalent genitourinary malignancies, is often associated with malignant disease progression and poor therapeutic outcomes. Approximately 30 % KIRC patients were found to be metastastic at initial diagnosis (Gupta et al., 2008). Metastatic KIRC evolves from primary KIRC and harbors multiple subpopulations with transcriptomic features, however, the molecular mechanisms of KIRC tumorigenesis and metastasis still remain unclear. In the meantime, effective diagnostic and prognostic biomarkers were still lacking (Song et al., 2018). Recently, aberrant ASs along with corresponding SFs were shown to have great potential value in exploring cancer biology, as both ASEs and SFs have been proved to produce different isoforms of onco-proteins which are associated with cell proliferation, anti-apoptosis and metastasis (Zhang et al., 2019). However, the roles of OS-SEs, SFs and signaling pathways in the tumorigenesis, metastasis, and prognosis of patients with KIRC were not quite clear.

In our study, a total of 6,081 OS-SEs were identified by univariate Cox regression analysis and we constructed a prediction model based on 5 OS-SEs (C4orf19-69001-AT, C16orf13-32924-ES, KIAA0930-62645-AP, FAM120C-89237-AT, UACA-31439-AP) screened by Lasso regression. Compared with previous prediction model of KIRC, the present on had a higher reliability (AUC: 0.788), with less predicters (Song et al., 2019). Additionally, the risk score was proved to be an independent prognostic factor, suggesting the good applicability in clinic for patients with KIRC.

DDX39B, as a DExD RNA helicase, was known to be involved in transportation of mRNA from nuclear to cytoplasm and pre-mRNA splicing (Shen, 2009). Particularly, the ATPase activity of DDX39B played an essential role in unwinding U4/U6 snRNA duplex and connecting U2 snRNP to the pre-mRNA in the process of ASE (Shen et al., 2008). In the present study, DDX39B was the only SF, whose associated OS-SEs were correlated with OS and metastasis by developing the network of OS-SEs and prognosis-related SFs. Similar to our study, mounting evidences regarded DDX39B as an important SF in triggering the progression and metastasis of various cancers. In BioXpress database, the expression level DDX39B was elevated in 75% (12 in 16) types of cancers (Awasthi et al., 2018a,b). Awastthi S et al. found that DDX39B could regulate the transcription and stability of pre-ribosomal RNA, the global translation, cell growth and proliferation. Furthermore, a study regarding to androgen receptor splice variant AR-V7 indicated that DDX39B could serve as the accelerator of AR-V7 mRNA expression and escalated DDX39B could result in resistance to androgen deprivation therapy and poor prognosis in patients with prostate cancer (Nakata et al., 2017).

Among these identified metastasis-associated OS-SEs, the parent genes of RHOT2-32938-RI and TCIRG1-17288-RI were verified by comprehensive databases. RHOT2 gene encodes a member of Rho family of GTPase, which are localized to the outer membrane of mitochondria (Wang et al., 2011). It plays an active role in mitochondrial fusion-fission dynamics, trafficking mitophagy function (Zheng et al., 2015). Mitochondrial dynamics was shown to be reprogrammed in tumor cells via gathering mitochondria at the cortical cytoskeleton (Caino et al., 2016). The mechanism could power the membrane machinery of cell movements, maintained phosphorylation of cell motility kinases, and heightened tumor invasion, chemotaxis, and metastasis (Caino et al., 2016; Agarwal et al., 2019). Besides, remodeling of mitochondrial functions is considered the commonest modified downstream of MYC gene, due to the MYC-dependent transcriptional control of GTPase RHOT1/RHOT2 and posttranslational modifications, such as RHOT phosphorylation by PINK kinase (Wang et al., 2011; Bailey et al., 2018). However, the exact function of RHOT2 has not been explored yet in KIRC. In our study, we found that abnormal expression of ASE of RHOT2 regulated by aberrant DDX39B could result in poor prognosis and tumor metastasis in patients with KIRC. Additionally, we also found out RHOT2-32938-RI was associated with sphingolipid metabolism, N-glycan biosynthesis and glycosphingolipid biosynthesis lacto and neolacto series by co-expression analysis. This might be a novel posttranslational regulation and new function for RHOT2 in KIRC tumorigenesis, progression and metastasis.

TCIRG1 (T-Cell Immune Regulator 1), also known as TIRC7, is essential in T cell activation (Heinemann et al., 1999). Previous studies revealed that TCIRG1 was widely up-regulated in numerous tumors, such as hepatocellular carcinoma,

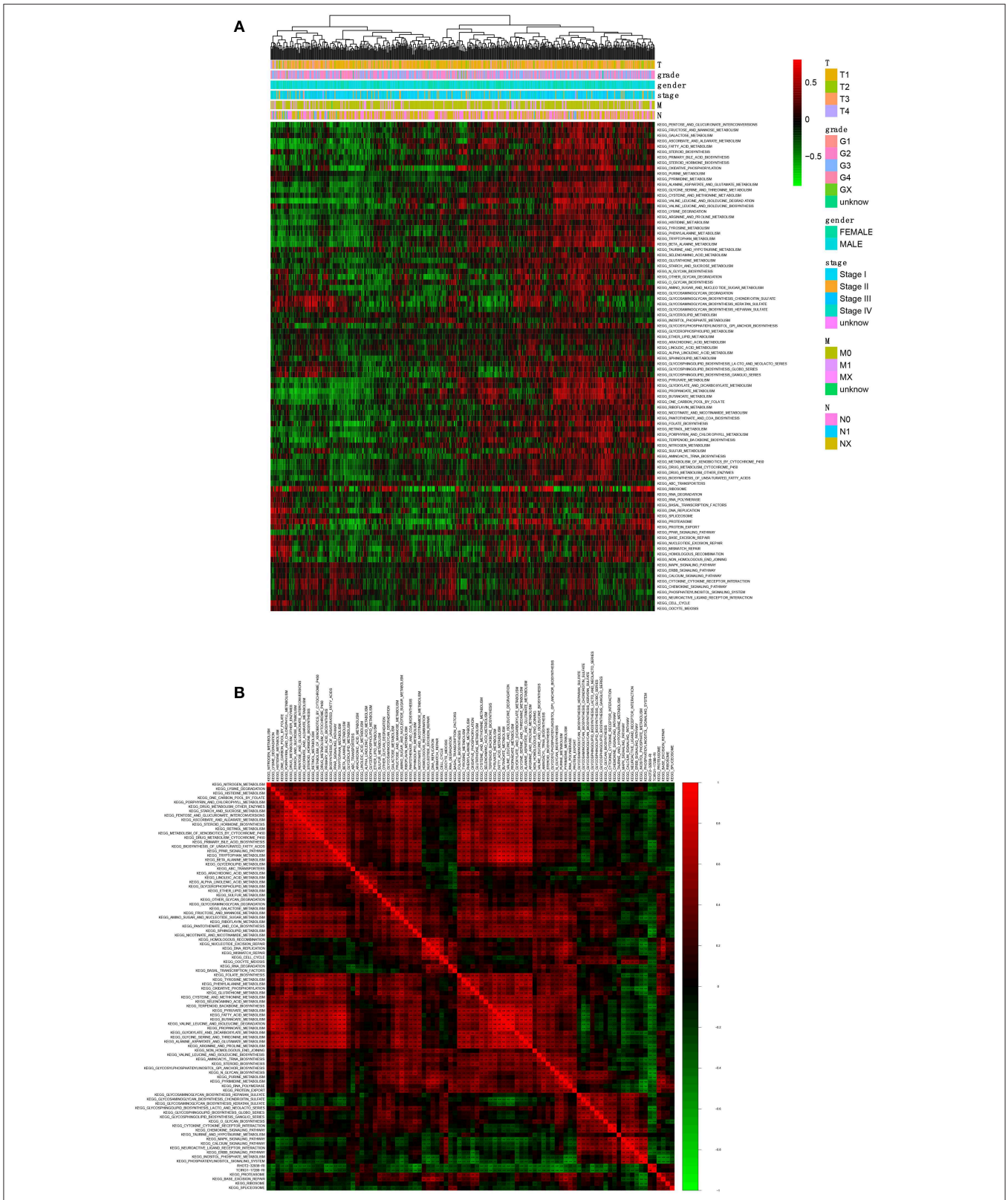


FIGURE 8 | The co-expression analysis between ASEs and signaling pathways. **(A)**: The co-expression heatmap of OS-related KEGG pathways; **(B)** The co-expression heatmap of OS-related KEGG pathways, RHOT2-32938-RI and TCIRG1-17288-RI. KEGG, Kyoto Encyclopedia of Genes and Genomes; RHOT2, Retained Intron of Ras Homolog Family Member T2; TCIRG1, T-Cell Immune Regulator 1.

esophageal adenocarcinoma and breast cancer, which might be associated with autophagic sequestration and degradation (Blair and Athanasou, 2004; Hinton et al., 2009; Botelho et al., 2010). With regard to tumor metastasis, TCIRG1 was reported to modulate the EMT regulatory proteins, such as E-cadherin, N-cadherin, Fibronectin, Vimentin, Snail and Slug, and regulate tumor invasion and metastasis in MDA-MB-231, B16-F10, and SNU475 cells (Hinton et al., 2009; Yotsumoto et al., 2013; Yang et al., 2018; Zhou et al., 2018). In our study, we found that aberrant ASE of TCIRG1 was associated with poor prognosis and tumor metastasis in patients with KIRC. In addition, the parent gene TCIRG1 was verified to be associated with OS and metastasis by external database. With regard to KIRC metastasis, bone is one of the most common sites. In this case, patients often experience local pain and even pathological fracture due to osteolytic destruction. As TCIRG1 encodes the osteoclast-specific 116-kD subunit of the vacuolar proton pump and its defect is responsible for a subset of human autosomal recessive osteopetrosis (Frattini et al., 2000), TCIRG1 could be regarded as the potential targets for KIRC metastasis, especially to skeletal metastasis. Nowadays, many anti-TCIRG1 specific monoclonal antibody (mAb) have been developed for different diseases (Kumamoto et al., 2004; Utku et al., 2006). Thus, further cell and animal experiments should be performed to detect the therapeutic effects of anti-TCIRG1 specific mAb in KIRC skeletal metastasis.

To further investigate the deep mechanism of DDX39B regulating RHOT2-32938-RI and TCIRG1-17288-RI, sphingolipid metabolism and N-glycan biosynthesis were identified as the overlapped co-expression signal pathways. Sphingolipids are the major molecules presenting on the cell membranes, which are composed of sphingosine-1-phosphate (S1P) and ceramide (Ogretmen, 2018). Sphingolipids metabolism is pivotal for normal cellular homeostasis with various events, including endocytosis, nutrient transport and protein synthesis. Bioactive sphingolipid could induce cell motility, migration and phenotypic plasticity, which result in tumor invasion and metastasis (Kumamoto et al., 2004; Bonora et al., 2015; Ogretmen, 2018).

N-glycans biosynthesis play important roles in the immune system, pathogen recognition and tumor metastasis via regulating cell adhesion and ligand recognition (Kadam, 2016). The modification of N-glycans also alter cell-cell or cell-matrix contacts and contribute to EMT, invasion and metastasis (Kadam, 2016). E-cadherin, an adhesion molecule, harbors mainly bisecting N-glycans by MGAT3 enzyme in normal epithelial cells. In the tumorigenesis, MGAT3 is down-regulated by promoter methylation and its counterpart MGAT5 is up-regulated. This change results in the formation of tri- and tetraantennary complex glycans on cadherins and E-cadherin internalization to the cytoplasm (Pinho et al., 2013).

There were still some limitations in our study. First of all, this study was a pure bioinformatics study, and the scientific hypothesis was not proved by biological experiments. Second, although the results were verified by external database, the sequencing data were obtained from one single cohort and the sample size was limited. Third, only the primary samples were found in TCGA database and the lack of samples of metastatic sites, such as lung, bone and brain, made this study less complete. At last, the limitation of single omics analysis was also an inherent defect of this study.

In the future, in order to verify the important roles of RHOT2 and TCIRG1 in KIRC metastasis, the functional experiment, such as wound healing and transwell, will be performed in KIRC in which the RHOT2 or TCIRG1 gene has been knocked out (Calabretta et al., 2016; Qi et al., 2016; Chen et al., 2017; Couture et al., 2017; Zhou et al., 2017). Next, these cells will also be used in the nude mouse tumor metastasis model. In addition, a direct mechanism experiment proving the direct mechanism of the splicing factor DDX39B producing splicing isoforms of RHOT2 and TCIRG1 will also be performed. Furthermore, clinical sampes of lung, bone and brain metastasis from KIRC will also be used to detect the expressions of RHOT2 and TCIRG1.

CONCLUSIONS

In conclusion, we established the prediction model with good with good performance in external validation. Based on the comprehensive bioinformatics analysis, we proposed that aberrant DDX39B regulated RHOT2-32938-RI and TCIRG1-17288-RI might be related to the tumorigenesis, metastasis and poor prognosis of KIRC via sphingolipid metabolism or N-glycan biosynthesis pathway.

DATA AVAILABILITY STATEMENT

All datasets for this study are included in the TCGA-KIRC program.

ETHICS STATEMENT

This study was approved by the Ethics Committee of Tongji Hospital, Tongji University School of Medicine.

AUTHOR CONTRIBUTIONS

TM, RH, ZZ, ZH, HY, CJ, PY, PH, XZ, ZL, DS, JZ, and LC: conception, design, manuscript writing, and final approval of manuscript. DS, JZ, and LC: provision of study material. TM, RH, and ZZ: collection and assembly of data. TM, RH, ZZ, ZH, HY, CJ, PY, PH, XZ, and ZL: Data analysis and interpretation.

FUNDING

This study was supported in part by the National Natural Science Foundation of China (Nos. 81702659, 81772856), Youth Fund of Shanghai Municipal Health Planning Commission (Nos. 2017YQ054, 2017Y0117); Henan medical science and technology research project (Grant No. 201602031).

ACKNOWLEDGMENTS

We thank the TCGA team of the National Cancer Institute for using their data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00270/full#supplementary-material

Supplementary Figure 1 | Bubble plots of the top 20 most significant OS-SEs among seven types of splicing patterns: **(A)** AA; **(B)** AD; **(C)** AP; **(D)** AT; **(E)** ES; **(F)** ME; **(G)** RI. OS-SEs, overall survival-related ASEs; AA, alternate acceptor; AD, alternate donor; AP, alternate promoter; AT, alternate terminator; ES, exon skip; ME, mutually exclusive exons; RI, retained intron.

Supplementary Figure 2 | The beeswarm plots of CALCOCO1-22108-RI **(A)**, CIRBP-46432-RI **(B)**, P4HTM-64788-ES **(C)**, TBC1D17-51116-ES **(D)**, THOP1-46623-AP **(E)** according to metastasis or not; The beeswarm plots of CALCOCO1-22108-RI **(F)**, CIRBP-46432-RI **(G)**, P4HTM-64788-ES **(H)**, TBC1D17-51116-ES **(I)**, THOP1-46623-AP **(J)** according to clinical status. CALCOCO1, calcium binding and coiled-coil domain 1; CIRBP, cold inducible RNA binding protein; P4HTM, prolyl 4-hydroxylase, transmembrane; TBC1D17, TBC1 domain family member 17; THOP1, thimet oligopeptidase 1.

Supplementary Figure 3 | The expressions of RHOT2 and TCIRG1 according to M **(A)** and clinical stage **(B)** in LinkedOmics; The Kaplan-Meier curve of RHOT2 and TCIRG1 in UALCAN **(C)** and SurvExpress **(D)**. RHOT2, Retained Intron of Ras Homolog Family Member T2; TCIRG1, T-Cell Immune Regulator.

REFERENCES

Agarwal, E., Altman, B. J., Ho Seo, J., Bertolini, I., Ghosh, J. C., Kaur, A., et al. (2019). Myc regulation of a mitochondrial trafficking network mediates tumor cell invasion and metastasis. *Mol. Cell. Biol.* 39, e00109–19. doi: 10.1128/MCB.00109-19

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:e74250. doi: 10.1371/journal.pone.0074250

Awasthi, S., Chakrapani, B., Mahesh, A., Chavali, P. L., Chavali, S., and Dhayalan, A. (2018a). DDX39B promotes translation through regulation of pre-ribosomal RNA levels. *RNA Biol.* 15, 1157–1166. doi: 10.1080/15476286.2018.1517011

Awasthi, S., Verma, M., Mahesh, A., K Khan, M. I., Govindaraju, G., Rajavelu, A., et al. (2018b). DDX49 is an RNA helicase that affects translation by regulating mRNA export and the levels of pre-ribosomal RNA. *Nucleic Acids Res.* 46, 6304–6317. doi: 10.1093/nar/gky231

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 174, 1034–1035. doi: 10.1016/j.cell.2018.07.034

Blair, H. C., and Athanasou, N. A. (2004). Recent advances in osteoclast biology and pathological bone resorption. *Histol. Histopathol.* 19, 189–199. doi: 10.14670/HH-19.189.

Bonora, M., Wieckowsk, M. R., Chinopoulos, C., Kepp, O., Kroemer, G., Galluzzi, L., et al. (2015). Molecular mechanisms of cell death: central implication of ATP synthase in mitochondrial permeability transition. *Oncogene* 34:1608. doi: 10.1038/onc.2014.462

Botelho, N. K., Schneiders, F. I., Lord, S. J., Freeman, A. K., Tyagi, S., Nancarrow, D. J., et al. (2010). Gene expression alterations in formalin-fixed, paraffin-embedded Barrett esophagus and esophageal adenocarcinoma tissues. *Cancer Biol. Ther.* 10, 172–179. doi: 10.4161/cbt.10.2.12166

Caino, M. C., Seo, J. H., Aguinaldo, A., Wait, E., Bryant, K. G., Kossenkov, A. V., et al. (2016). A neuronal network of mitochondrial dynamics regulates metastasis. *Nat. Commun.* 7:13730. doi: 10.1038/ncomms13730

Calabretta, S., Bielli, P., Passacantilli, I., Pilozzi, E., Fendrich, V., Capurso, G., et al. (2016). Modulation of PKM alternative splicing by PTBP1 promotes gemcitabine resistance in pancreatic cancer cells. *Oncogene* 35, 2031–2039. doi: 10.1038/onc.2015.270

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Chen, K., Xiao, H., Zeng, J., Yu, G., Zhou, H., Huang, C., et al. (2017). Alternative splicing of EZH2 pre-mRNA by SF3B3 contributes to the tumorigenic potential of renal cancer. *Clin. Cancer Res.* 23, 3428–3441. doi: 10.1158/1078-0432.CCR-16-2020

Chen, Q. F., Li, W., Wu, P., Shen, L., and Huang, Z. L. (2019). Alternative splicing events are prognostic in hepatocellular carcinoma. *Aging (Albany. NY)* 11, 4720–4735. doi: 10.18632/aging.102085

Couture, F., Sabbagh, R., Kwiatkowska, A., Desjardins, R., Guay, S. P., Bouchard, L., et al. (2017). PACE4 undergoes an oncogenic alternative splicing switch in cancer. *Cancer Res.* 77, 6863–6879. doi: 10.1158/0008-5472.CAN-17-1397

Deng, M., Brägelmann, J., Kryukov, I., Saraiva-Agostinho, N., and Perner, S. (2017). FirebrowseR: an R client to the broad institute's firehose pipeline. *Database (Oxford)* 2017:baw160. doi: 10.1093/database/baw160

Frankiw, L., Baltimore, D., and Li, G. (2019). Alternative mRNA splicing in cancer immunotherapy. *Nat. Rev. Immunol.* doi: 10.1038/s41577-019-0195-7. [Epub ahead of print].

Frattini, A., Orchard, P. J., Sobacchi, C., Giliani, S., Abinun, M., Mattsson, J. P., et al. (2000). Defects in TCIRG1 subunit of the vacuolar proton pump are responsible for a subset of human autosomal recessive osteopetrosis. *Nat. Genet.* 25, 343–346. doi: 10.1038/77131

Gupta, K., Miller, J. D., Li, J. Z., Russell, M. W., and Charbonneau, C. (2008). Epidemiologic and socioeconomic burden of metastatic renal cell carcinoma (mRCC): a literature review. *Cancer Treat. Rev.* 34, 193–205. doi: 10.1016/j.ctrv.2007.12.001

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

Heinemann, T., Bulwin, G. C., Randall, J., Schnieders, B., Sandhoff, K., Volk, H. D., et al. (1999). Genomic organization of the gene coding for TIRC7, a novel membrane protein essential for T cell activation. *Genomics* 57, 398–406. doi: 10.1006/geno.1999.5751

Hinton, A., Sennoune, S. R., Bond, S., Fang, M., Reuveni, M., Sahagian, G. G., et al. (2009). Function of a subunit isoforms of the V-ATPase in pH homeostasis and *in vitro* invasion of MDA-MB231 human breast cancer cells. *J. Biol. Chem.* 284, 16400–16408. doi: 10.1074/jbc.M901201200

Hsieh, J. J., Purdue, M. P., Signoretti, S., Swanton, C., Albiges, L., Schmidinger, M., et al. (2017). Renal cell carcinoma. *Nat. Rev. Dis. Primers* 3:17009. doi: 10.1038/nrdp.2017.9

Jonasch, E. (2018). Updates to the management of kidney cancer. *J. Natl. Compr. Canc. Netw.* 16, 639–641. doi: 10.6004/jnccn.2018.0039

Jonasch, E., Gao, J., and Rathmell, W. K. (2014). Renal cell carcinoma. *BMJ* 349:g4797. doi: 10.1136/bmj.g4797

Kadam, P. D. (2016). Rectocutaneous fistula with transmigration of the suture: a rare delayed complication of vault fixation with the sacrospinous ligament. *Int. Urogynecol. J.* 27, 155–157. doi: 10.1007/s00192-015-2823-5

Kouyama, Y., Masuda, T., Fujii, A., Ogawa, Y., Sato, K., Tobo, T., et al. (2019). Oncogenic splicing abnormalities induced by DEAD-Box Helicase 56 (DDX56) amplification in colorectal cancer. *Cancer Sci.* 110, 3132–3144. doi: 10.1111/cas.14163

Kumamoto, Y., Tomschegg, A., Bennai-Sanfourche, F., Boerner, A., Kaser, A., Schmidt-Knosalla, I., et al. (2004). Monoclonal antibody specific for TIRC7 induces donor-specific anergy and prevents rejection of cardiac allografts in mice. *Am. J. Transplant* 4, 505–514. doi: 10.1111/j.1600-6143.2004.00367.x

Lee, S. C., and Abdel-Wahab, O. (2016). Therapeutic targeting of splicing in cancer. *Nat. Med.* 22, 976–986. doi: 10.1038/nm.4165

Montes, M., Sanford, B. L., Comiskey, D. F., and Chandler, D. S. (2019). RNA splicing and disease: animal models to therapies. *Trends Genet.* 35, 68–87. doi: 10.1016/j.tig.2018.10.002

Nagy, A., Lanczky, A., Menyhart, 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/s41598-018-29514-3

Nakata, D., Nakao, S., Nakayama, K., Araki, S., Nakayama, Y., Aparicio, S., et al. (2017). The RNA helicase DDX39B and its paralog DDX39A regulate androgen receptor splice variant AR-V7 generation. *Biochem. Biophys. Res. Commun.* 483, 271–276. doi: 10.1016/j.bbrc.2016.12.153

Ogretmen, B. (2018). Sphingolipid metabolism in cancer signalling and therapy. *Nat. Rev. Cancer* 18, 33–50. doi: 10.1038/nrc.2017.96

Pinho, S. S., Figueiredo, J., Cabral, J., Carvalho, S., Dourado, J., Magalhães, A., et al. (2013). E-cadherin and adherens-junctions stability in gastric carcinoma: functional implications of glycosyltransferases involving N-glycan branching biosynthesis, N-acetylglucosaminyltransferases III and V. *Biochim. Biophys. Acta* 1830, 2690–2700. doi: 10.1016/j.bbagen.2012.10.021

Piva, F., Giulietti, M., Nocchi, L., and Principato, G. (2009). SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans. *Bioinformatics* 25, 1211–1213. doi: 10.1093/bioinformatics/btp124

Porta, C., Cosmai, L., Leibovich, B. C., Powles, T., Gallieni, M., and Bex, A. (2019). The adjuvant treatment of kidney cancer: a multidisciplinary outlook. *Nat. Rev. Nephrol.* 15, 423–433. doi: 10.1038/s41581-019-0131-x

Qi, Y., Yu, J., Han, W., Fan, X., Qian, H., Wei, H., et al. (2016). A splicing isoform of TEAD4 attenuates the Hippo-YAP signalling to inhibit tumour proliferation. *Nat. Commun.* 7:ncomms11840. doi: 10.1038/ncomms11840

Ryan, M., Wong, W. C., Brown, R., Akbani, R., Su, X., Broom, B., et al. (2016). TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. *Nucleic Acids Res.* 44, D1018–1022. doi: 10.1093/nar/gkv1288

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

Shen, H. (2009). UAP56- a key player with surprisingly diverse roles in pre-mRNA splicing and nuclear export. *BMB Rep.* 42, 185–188. doi: 10.5483/BMBRep.2009.42.4.185

Shen, H., Zheng, X., Shen, J., Zhang, L., Zhao, R., and Green, M. R. (2008). Distinct activities of the DExD/H-box splicing factor hUAP56 facilitate stepwise assembly of the spliceosome. *Genes Dev.* 22, 1796–1803. doi: 10.1101/gad.1657308

Song, E., Song, W., Ren, M., Xing, L., Ni, W., Li, Y., et al. (2018). Identification of potential crucial genes associated with carcinogenesis of clear cell renal cell carcinoma. *J. Cell. Biochem.* 119, 5163–5174. doi: 10.1002/jcb.26543

Song, J., Liu, Y. D., Su, J., Yuan, D., Sun, F., and Zhu, J. (2019). Systematic analysis of alternative splicing signature unveils prognostic predictor for kidney renal clear cell carcinoma. *J. Cell. Physiol.* 234, 22753–22764 doi: 10.1002/jcp.28840

Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A., and Skotheim, R. I. (2016). Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* 35, 2413–2427. doi: 10.1038/onc.2015.318

Utku, N., Heinemann, T., Winter, M., Bulwin, C. G., Schlawinsky, M., Fraser, P., et al. (2006). Antibody targeting of TIRC7 results in significant therapeutic effects on collagen-induced arthritis in mice. *Clin. Exp. Immunol.* 144, 142–151. doi: 10.1111/j.1365-2249.2006.03044.x

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963. doi: 10.1093/nar/gkx1090

Wang, C., Zheng, M., Wang, S., Nie, X., Guo, Q., Gao, L., et al. (2019). Whole genome analysis and prognostic model construction based on alternative splicing events in endometrial cancer. *Biomed Res. Int.* 2019:2686875. doi: 10.1155/2019/2686875

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509

Wang, X., Winter, D., Ashrafi, G., Schlehe, J., Wong, Y. L., Selkoe, D., et al. (2011). PINK1 and Parkin target Miro for phosphorylation and degradation to arrest mitochondrial motility. *Cell* 147, 893–906. doi: 10.1016/j.cell.2011.10.018

Wu, H. Y., Peng, Z. G., He, R. Q., Luo, B., Ma, J., Hu, X. H., et al. (2019). Prognostic index of aberrant mRNA splicing profiling acts as a predictive indicator for hepatocellular carcinoma based on TCGA SpliceSeq data. *Int. J. Oncol.* 55, 425–438. doi: 10.3892/ijo.2019.4834

Xing, S., Li, Z., Ma, W., He, X., Shen, S., Wei, H., et al. (2019). DIS3L2 promotes progression of hepatocellular carcinoma via hnRNP U-mediated alternative splicing. *Cancer Res.* 79, 4923–4936. doi: 10.1158/0008-5472.CAN.-19-0376

Yang, H. D., Eun, J. W., Lee, K. B., Shen, Q., Kim, H. S., Kim, S. Y., et al. (2018). T-cell immune regulator 1 enhances metastasis in hepatocellular carcinoma. *Exp. Mol. Med.* 50:e420. doi: 10.1038/emm.2017.166

Yotsumoto, F., Tokunaga, E., Oki, E., Maehara, Y., Yamada, H., Nakajima, K., et al. (2013). Molecular hierarchy of heparin-binding EGF-like growth factor-regulated angiogenesis in triple-negative breast cancer. *Mol. Cancer Res.* 11, 506–517. doi: 10.1158/1541-7786.MCR-12-0428

Zhang, Y., Yan, L., Zeng, J., Zhou, H., Liu, H., Yu, G., et al. (2019). Pan-cancer analysis of clinical relevance of alternative splicing events in 31 human cancers. *Oncogene* 8, 6678–6695. doi: 10.1038/s41388-019-0910-7

Zheng, L., Li, X., Gu, Y., Lv, X., and Xi, T. (2015). The 3'UTR of the pseudogene CYP4Z2P promotes tumor angiogenesis in breast cancer by acting as a ceRNA for CYP4Z1. *Breast Cancer Res. Treat.* 150, 105–118. doi: 10.1007/s10549-015-3298-2

Zhou, M., Diao, Z., Yue, X., Chen, Y., Zhao, H., Cheng, L., et al. (2016). Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer. *Oncotarget* 7:56383. doi: 10.18632/oncotarget.10891

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018). An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. *Mol. Neurobiol.* 55, 3684–3697. doi: 10.1007/s12035-017-0572-9

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

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 © 2019 Meng, Huang, Zeng, Huang, Yin, Jiao, Yan, Hu, Zhu, Li, Song, Zhang and Cheng. 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 Genetics

ORIGINAL RESEARCH
published: 16 October 2019
doi: 10.3389/fgene.2019.00936

Enhancer LncRNAs Influence Chromatin Interactions in Different Ways

*Yue Hou, Rongxin Zhang and Xiao Sun**

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China

More than 98% of the human genome does not encode proteins, and the vast majority of the noncoding regions have not been well studied. Some of these regions contain enhancers and functional non-coding RNAs. Previous research suggested that enhancer transcripts could be potent independent indicators of enhancer activity, and some enhancer lncRNAs (elncRNAs) have been proven to play critical roles in gene regulation. Here, we identified enhancer–promoter interactions from high-throughput chromosome conformation capture (Hi-C) data. We found that elncRNAs were highly enriched surrounding chromatin loop anchors. Additionally, the interaction frequency of elncRNA-associated enhancer–promoter pairs was significantly higher than the interaction frequency of other enhancer–promoter pairs, suggesting that elncRNAs may reinforce the interactions between enhancers and promoters. We also found that elncRNA expression levels were positively correlated with the interaction frequency of enhancer–promoter pairs. The promoters interacting with elncRNA-associated enhancers were rich in RNA polymerase II and YY1 transcription factor binding sites. We clustered enhancer–promoter pairs into different groups to reflect the different ways in which elncRNAs could influence enhancer–promoter pairs. Interestingly, G-quadruplexes were found to potentially mediate some enhancer–promoter interaction pairs, and the interaction frequency of these pairs was significantly higher than that of other enhancer–promoter pairs. We also found that the G-quadruplexes on enhancers were highly related to the expression of elncRNAs. G-quadruplexes located in the promoters of elncRNAs led to high expression of elncRNAs, whereas G-quadruplexes located in the gene bodies of elncRNAs generally resulted in low expression of elncRNAs.

Keywords: chromatin structure, enhancer lncRNA, enhancer–promoter interaction, Hi-C, transcription factor

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Jianfeng Xu,
Baylor College of Medicine, United States
Yan Zhang,
Harbin Institute of Technology, China

****Correspondence:***
Xiao Sun
xsun@seu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *12 July 2019*
Accepted: *05 September 2019*
Published: *16 October 2019*

Citation:
Hou Y, Zhang R and Sun X (2019) Enhancer LncRNAs Influence Chromatin Interactions in Different Ways. Front. Genet. 10:936. doi: 10.3389/fgene.2019.00936

INTRODUCTION

It has been widely accepted that a large proportion of the human genome is transcribed, but that less than 2% of the transcripts are subsequently translated into proteins (Katayama et al., 2005; Djebali et al., 2012; Sallam et al., 2018). Long non-coding RNAs (lncRNAs), transcripts longer than 200 nucleotides, have attracted increasing attention because of their functional relevance in various biological processes (Iyer et al., 2015; Liu et al., 2017; Kopp and Mendell, 2018). Because lncRNAs are expressed at relatively low levels and are weakly conserved during evolution, they

are difficult to annotate and were historically regarded as junk DNA (Uszczynska-Ratajczak et al., 2018). However, convincing evidence has recently emerged that at least some lncRNAs play critical roles in disease (Wapinski and Chang, 2011; Shi et al., 2013; Yan et al., 2015; Wan et al., 2016; Zhang et al., 2017), organism development (Grote et al., 2013; Fatica and Bozzoni, 2014; Sun et al., 2017), and aging (Bianchessi et al., 2015; Yang et al., 2016a; Neppl et al., 2017). Iyer et al. identified 58,648 lncRNA genes in the human genome, of which 1% harbored ultraconserved elements and 7% overlapped with disease-associated SNPs (Iyer et al., 2015). Using a CRISPR interference platform, hundreds of lncRNAs were proven to be required for robust cellular growth for different cell types (Liu et al., 2017). Some lncRNAs can regulate the expression of neighboring (*cis*) or distal (*trans*) genes (Yu et al., 2018). *In cis* means that lncRNAs regulate target genes by the act of transcription (Tehrani et al., 2018). LncRNAs, such as *bxd* lncRNA (Hao et al., 2017), can regulate downstream promoters *in cis* through transcriptional interference (Lin et al., 2018). In addition to acting *in cis*, some lncRNAs translocate from their sites of synthesis and regulate distal target genes *in trans* (Kopp and Mendell, 2018). For instance, *Firre* lncRNA localizes at five distinct *trans*-chromosomal loci through interacting with the nuclear-matrix factor hnRNPU (Hacisuleyman et al., 2014; Yang et al., 2015). Moreover, the *Xist* lncRNA participates in silencing transcription *in trans* by interacting with SHARP (McHugh et al., 2015).

Based on their genomic organization, lncRNAs can be categorized into different subtypes, including intragenic lncRNAs, intergenic lncRNAs, and enhancer lncRNAs (elncRNAs) (Devaux et al., 2015; St Laurent et al., 2015). Enhancers are genomic regions that are bound by transcription factors (TFs) and are capable of interacting with promoters to augment gene expression. Generally, enhancer regions are marked by histone 3 lysine 4 monomethylation (H3K4me1) and histone 3 lysine 27 acetylation (H3K27ac). The binding of the general transcriptional co-activator CBP to enhancers may recruit RNA polymerase II (RNA POLII) and produce enhancer transcripts (Kim et al., 2010). Pioneering research has proven that enhancer RNAs are involved in specific enhancer–promoter looping initiated by ER-α binding (Li et al., 2013). In addition to affecting enhancer–promoter loops, some elncRNAs regulate gene expression by recruiting TFs to the promoters of target genes. LEENE, an elncRNA that enhances eNOS expression, can facilitate the recruitment of RNA POLII to the eNOS promoter to enhance eNOS nascent RNA transcription (Miao et al., 2018). Arc eRNA, an elncRNA that is expressed from the enhancer for Activity-regulated cytoskeletal protein (Arc), can facilitate NELF release from the target promoter (Schaukowitch et al., 2014). Moreover, a muscle-specific elncRNA, DRReRNA, regulates the transcription of myogenin in *trans* by mediating the recruitment of cohesin proteins (Tsai et al., 2018). In principle, nascent RNAs can remain at their sites of synthesis. One of the well-studied mechanisms for retaining nascent RNA is through the formation of an R-loop, which is a double-stranded RNA:DNA hybrid opposite a displaced single strand of DNA (Li and Fu, 2019). R-loops, which are associated with transcription activities under physiological conditions (Skourti-Stathaki et al., 2011; Stork et al., 2016), predominantly form on promoters and enhancers associated with GC-skewed sequences (Ginno et al., 2012; Chen et al., 2017; Li and Fu 2019). These findings suggested that elncRNAs might stay where they are synthesized but exert long-distance regulatory effects on target genes.

Previous studies provided great advances in our understanding of the functions of elncRNAs. However, some studies roughly coupled enhancers to their closest genes, which has been proven to be an imprecise method for identifying the target genes of enhancers. DNA is highly compacted in the nucleus, resulting in a complicated three-dimensional genome conformation. Currently, the developed powerful Hi-C technology has been used to profile the three-dimensional chromatin structure in diverse organisms and cells (Lieberman-Aiden et al., 2009; Rao et al., 2014; Mifsud et al., 2015). As enhancers and their target promoters frequently contact each other despite being separated by thousands or millions of base pairs in genomic distance (Ay et al., 2014), several methods have been proposed to identify enhancers and their target genes using Hi-C (Whalen et al., 2016; Ron et al., 2017). Mifsud et al. proposed that transcriptionally active genes normally interact with regulatory elements and inactive genes frequently interact with genomic regions that are rich in repressive markers (Mifsud et al., 2015). Beagrie et al. found an abundance of three-way contacts among highly transcribed regions (Beagrie et al., 2017). Moreover, specific enhancer transcripts have been proven to be involved in maintaining the formation of loop structures (Lai et al., 2013; Li et al., 2013; Hsieh et al., 2014; Yang et al., 2016b). However, it remains a challenge to decipher the function and mechanism of elncRNAs in the genome-wide range.

In this study, we comprehensively characterized elncRNAs by analyzing the human chromatin structure. Using Hi-C data, chromatin loops and enhancer–promoter interactions were identified in the GM12878 cell line. Our study was intended to resolve the following issues: 1) whether chromatin loops are associated with elncRNAs in the genome-wide range; 2) whether enhancer–promoter interactions are influenced by elncRNAs in the genome-wide range; and 3) the relationship between elncRNAs and transcription factor binding sites (TFBSs). We found that chromatin loops and enhancer–promoter interactions were highly associated with elncRNAs. By analyzing the relationship between elncRNAs and TFBSs, we found that elncRNAs are capable of affecting TFBSs on both local enhancers and target promoters. Our findings suggest that elncRNAs influence enhancer–promoter interactions in different ways.

MATERIALS AND METHODS

Identification of Genomic Elements

The protein-coding and lncRNA genes in the human genome were downloaded from the GENCODE (Harrow et al., 2012) and NONCODE (Fang et al., 2018) databases, respectively. A total of 19,901 protein-coding genes and 96,308 lncRNA genes were identified. In accordance with previous research (He et al., 2014), promoters were defined as regions located 2 kilo-base pairs (kb) upstream and 0.5 kb downstream of transcription start sites (TSSs) annotated in GENCODE (Harrow et al., 2012).

Genomic regions of enhancers in the GM12878 cell line were derived from a previous study (Yip et al., 2012). Enhancers located in promoters and gene bodies of protein-coding genes were excluded. After filtering, a total of 35,939 enhancers in the GM12878 cell line were retained.

Global Nuclear Run-On Sequencing Data and RNA Sequencing Data

The global nuclear run-on sequencing (GRO-seq) data of the GM12878 cell line were generated by Core et al. (GEO accession number: GSE60456) (Core et al., 2014). GRO-seq captures 5′-capped RNAs from active transcriptional regulatory elements with high accuracy (Danko et al., 2015). The obtained GRO-seq reads were mapped to the human reference genome (GRCh37/hg19) using *Bowtie2* (Langmead and Salzberg, 2012). We used dREG, a computational tool for identifying transcriptional regulatory DNA sequences using GRO-seq data, to call peaks (Danko et al., 2015).

The paired-end RNA-seq data of the GM12878 cell line (GEO accession number: GSE90223) were generated by Thomas Gingeras' group of the ENCODE Consortium (Consortium, 2012). RNA-seq reads were mapped to the human reference genome (GRCh37/hg19) by *tophat* (Trapnell et al., 2012). We used *cufflinks* to generate the transcriptome assembly (Trapnell et al., 2010) and *cuffdiff* to test for differential expression (false discovery rate (FDR) <0.05; fold change >1.5) (Trapnell et al., 2013). As reported previously, numerous lncRNAs are expressed at much lower levels than protein-coding genes (Derrien et al., 2012); therefore, we used a threshold of 0.21 fragments per kilobase of transcript per million fragments mapped (FPKM) to define expressed lncRNAs, in accordance with previous studies (Hart et al., 2013; Bonnal et al., 2015).

Identification of elncRNAs

It has been proven that active transcriptional regulatory elements can be identified from GRO-seq data by dREG (Danko et al., 2015; Wang et al., 2019). In addition, GRO-seq reads have been shown to be highly accumulated around active enhancer regions (±1 kb) (Danko et al., 2015; Hu et al., 2017; Wang et al., 2019). Therefore, we designated the enhancers that fall within 1 kb of the GRO-seq peaks that were called by dREG as active enhancers (**Figure 1**). The lncRNAs that overlapped with the active enhancers were defined as elncRNAs (**Figure 1**), consistent with the method described in a previous study (Pefanis et al., 2015). As a result, 5.02% of the lncRNAs were defined as elncRNAs in the GM12878 cell line.

Identification of Enhancer–Promoter Interaction Pairs

Hi-C reads and Hi-C interaction matrixes of the GM12878 cell line generated by Rao et al. were downloaded from the GEO repository under accession number GSE63525 (Rao et al., 2014). Using the chromatin interactions from Hi-C data, the frequency of all enhancer–promoter interactions in the GM12878 cell line was calculated. For example, for an enhancer–promoter interaction pair, the interaction frequency was represented by the count of reads that were located in both gene promoter regions and enhancer regions. To calculate statistical confidence estimates for the interaction pairs, we used the method, *fit-HiC*, as proposed by Duan et al. (2010). Only enhancer–promoter interaction pairs with FDR <0.001 were retained.

We designated pairs that consisted of elncRNA-associated enhancers and their target promoters as elncRNA-associated enhancer–promoter interaction pairs. Other enhancer–promoter interaction pairs were defined as non-elncRNA pairs.

Selection of Structuring Factors

Previous studies have proven that some specific elncRNAs regulate the expression of their target genes by recruiting TFs to the promoter regions of the target genes (Schaukowitch et al., 2014; Miao et al., 2018; Tsai et al., 2018). To find the links between the TFBSs on enhancer–promoter pairs and elncRNAs, we selected eight proteins that potentially influenced enhancer–promoter

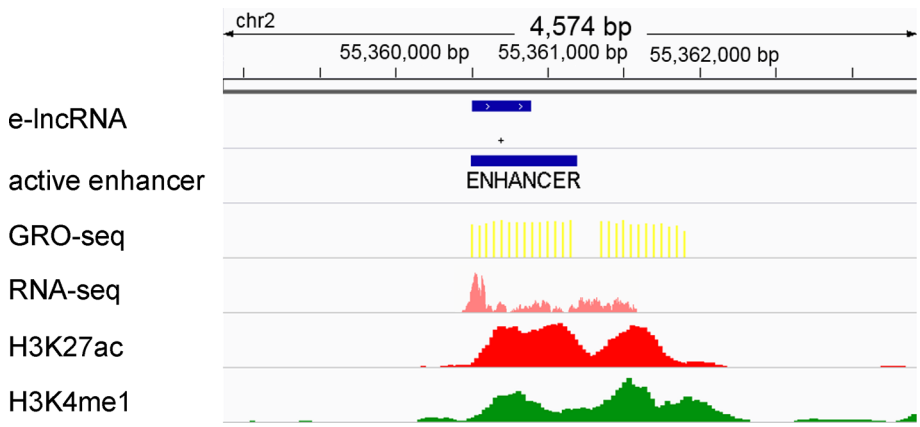


FIGURE 1 | Definition of active enhancers and elncRNAs. Enhancers were predicted by ChromHMM and Segway according to the histone modifications surrounding them. The enhancers that fall within 1 kb of the GRO-seq peaks were defined as active enhancers. LncRNA genes overlapping with active enhancers were defined as elncRNAs.

interactions. CTCF, RAD21, and SMC3 have been well studied in terms of their ability to influence chromatin structure (Rao et al., 2014; Hong and Kim, 2017). RNA POLII can arrange spatial organization and mediate some loop structures which are smaller than CTCF loops (Tang et al., 2015). Weintraub et al. found that YY1 is a structural regulator of enhancer–promoter interactions and facilitates gene expression (Weintraub et al., 2017). In addition to these well-studied structuring factors, we also used ReMap, an integrative ChIP-seq analysis of regulatory regions, to find candidate proteins that can potentially mediate chromatin interaction. ReMap was used to annotate all of the filtered chromatin interactions (FDR < 0.001), and the most enriched TFs in intersection ($p < 1 \times 10^{-500}$) were selected as candidate proteins (Cheneby et al., 2018). After excluding TFs that were not available in a public database or not expressed in the GM12878 cell line, we selected HDGF, GATAD2B and GABPA from the most enriched TFs as structuring factors. Previous study suggested that G-quadruplexes, stable four-stranded non-canonical DNA structures, potentially facilitate enhancer–promoter interactions (Hegyi, 2015; Hou et al., 2019). Therefore, we also selected G-quadruplex sequences, which were derived from the work of Chambers et al. (2015) and can form G-quadruplexes *in vitro*, as a structuring factor.

Although we have selected many structuring factors, a large amount of chromatin interactions are mediated by other TFs. Therefore, we used the ENCODE ChIP-seq data for 137 TFs in the GM12878 cell line, which were merged by ReMap, as an integrated factor. All of the raw data of the structuring factors are shown in **Table 1**.

ChIP-seq Data Analysis

All of the ChIP-seq data were generated by the ENCODE Consortium (Consortium, 2012) and can be retrieved from the GEO database using their accession number (**Table 1**). To identify ChIP-seq peak regions, we performed peak calling using MACS with the default parameters (Zhang et al., 2008).

Normalized ChIP-seq Peak Values on Enhancer–Promoter Pairs

We mapped all selected structuring factors (**Table 1**) onto the identified enhancer–promoter pairs. We defined enhancers/promoters as being associated with a specific structuring factor if they overlapped with a peak region of the selected structuring factor data. For G-quadruplex sequences, the G4-seq values provided by Chambers et al. (2015) were used to characterize the signal values of G-quadruplexes on enhancers/promoters. The peak counts on enhancers/promoters were used to define the signal values of the merged TFs on enhancers/promoters. For other structuring factors, the peak values, which were calculated by MACS (Zhang et al., 2008), were used to define the signal values of the structuring factor of these enhancers/promoters. If multiple peaks of the certain structuring factor overlapped with one enhancer/promoter, the signal value of the structuring factor of the enhancer/promoter equals the maximum peak value.

Because most enhancer–promoter pairs are associated with several structuring factors and the ChIP-seq data of different structuring factors were from different experiments, the signal values on each enhancer–promoter pair were normalized. We used Z-score normalization to standardize different structuring factor signal values of enhancers/promoters.

$$Z_{ij} = \frac{x_{ij} - \mu_i}{\delta_i} (i \in X; j \in Y)$$

Here, Z_{ij} is the normalized signal value of the specified structuring factors (i) on a specified enhancer/promoter (j); the specified structuring factor (i) belongs to the structuring factors (X) in **Table 1**; the specified enhancer/promoter (j) belongs to previously identified enhancer–promoter pairs (Y); x_{ij} represents the raw signal value of the specified structuring factors (i) on specified enhancer/promoter (j); μ_i equals the average signal value of the specified structuring factor (i) of all enhancers/promoters (Y); and δ_i indicates the standard deviation of the specified structuring factor (i) in all enhancers/promoters (Y).

TABLE 1 | The structuring factor data analyzed in this study.

Structuring factors	GEO number	Reference
CTCF	GSM935611	(Consortium 2012, Pope et al. 2014)
RAD21	GSM935332	(Consortium 2012, Pope et al. 2014)
SMC3	GSM935376	(Consortium 2012, Pope et al. 2014)
RNA POLII	GSM803355	(Consortium 2012, Pope et al. 2014, Gertz et al. 2013)
YY1	GSM803406	(Consortium 2012, Pope et al. 2014)
HDGF	GSE91531	(Consortium 2012)
GATAD2B	GSE105881	(Consortium 2012)
GABPA	GSE96120	(Consortium 2012)
G-quadruplex sequence	GSE63874	(Chambers et al. 2015)
ReMap TFs		(Cheneby et al. 2018)

Clustering Enhancer–Promoter Pairs

We performed hierarchical clustering on elncRNA-associated enhancer–promoter pairs and other enhancer–promoter pairs in accordance with their normalized structuring factor signal values. The Clustering software (https://web.stanford.edu/group/sherlocklab/cluster.html) was used to cluster interaction pairs. The Pearson correlation was set as the distance measurement as described previously (Lan et al., 2012). Using all of the normalized signal values, the elncRNA-associated enhancer–promoter pairs and other enhancer–promoter pairs were clustered into 10 and 6 groups, respectively.

RESULTS

ElncRNAs Are Highly Enriched in Chromatin Loop Anchors

A total of 9,449 high-confidence chromatin loops were identified in the GM12878 cell line. Each loop consisted of two interacting anchor points, which were defined as chromatin loop anchors. We calculated the relative density of elncRNAs and other lncRNAs across the entire chromatin loops (**Figure 2A**). We

observed high accumulation of both elncRNAs and other lncRNAs at chromatin loop anchors, with the profiles found to gradually decline towards the central regions of chromatin loops (Student's *t*-test, $p = 1.08 \times 10^{-203}$ and 5.77×10^{-133}, respectively). Furthermore, the relative density of elncRNAs surrounding chromatin loop anchors was significantly higher than that of other lncRNAs (Student's *t*-test, $p = 3.16 \times 10^{-197}$). The relative density of elncRNAs in the central regions of chromatin loops was slightly but significantly lower than that of other lncRNAs (Student's *t*-test, $p = 9.42 \times 10^{-27}$). We next calculated the enrichment of loop anchors with elncRNAs (**Figure 2B**). The high enrichment of loop anchors with elncRNAs indicated that loop anchors tend to localize at sites where elncRNAs are produced, suggesting a potential role of elncRNAs in chromatin loops. Consistent with our observations, it has been reported that AS1eRNA, which is produced by the enhancer downstream of DHRS4-AS1, is involved in the formation of a loop between DHRS4-AS1 and its enhancer (Yang et al., 2016b). In this case, the enhancer and DHRS4-AS1 function as the loop anchors.

Moreover, the chromatin loop anchors containing elncRNAs displayed significantly higher loop anchor interactions (**Figure 2C**, Student's *t*-test, $p = 1.47 \times 10^{-22}$), suggesting that elncRNAs at loop anchors potentially reinforce the interactions of loop anchors, which may help to maintain chromatin loop structures. The distribution of the architectural proteins including CTCF, SMC3, and RAD21 around loop anchors is shown in **Supplementary Figure 1**. Surprisingly, the ChIP-seq peak counts of these architectural proteins showed no significant differences between elncRNA-containing loop anchors and other loop anchors (**Supplementary Figures 1A–C**, Student's *t*-test, $p > 0.001$), indicating that the high interaction strength of elncRNA-containing loop anchors does not arise from these architectural proteins. We found that YY1 ChIP-seq peak counts around elncRNA-containing loop anchors were significantly higher than those around other loop anchors (**Figure 2D**, Student's *t*-test, $p = 1.62 \times 10^{-27}$). Using CLIP-seq, YY1 was found to be capable of interacting with nascent enhancer RNA at the active enhancer regions where it is bound to DNA (Sigova et al., 2015). In addition, YY1 was shown to promote DNA interactions and chromatin looping (Weintraub et al., 2017). These findings suggested that elncRNAs on loop anchors can function to "trap" YY1, thereby increasing the strength of interaction between loop anchors (**Figures 2C, D**). We used ReMap to merge ChIP-seq data of 137 TFs in the GM12878 cell line (Cheneby et al., 2018). The distribution of these TF ChIP-seq peaks around loop anchors is shown in **Figure 2E**. Likewise, we found that the merged TF ChIP-seq peak counts around elncRNA-containing loop anchors were significantly higher than those around other loops (**Figure 2E**, Student's *t*-test, $p = 1.94 \times 10^{-71}$). These results suggested that the highly abundant TFBSs on elncRNA-containing loop anchors promoted the transcription of elncRNAs. As feedback regulatory elements, elncRNAs on loop anchors can facilitate the loop anchor interactions by recruiting TFs such as YY1.

We used Hi-C interaction pairs to select the loop anchors that interact with elncRNA genes (FDR <0.001); these anchors

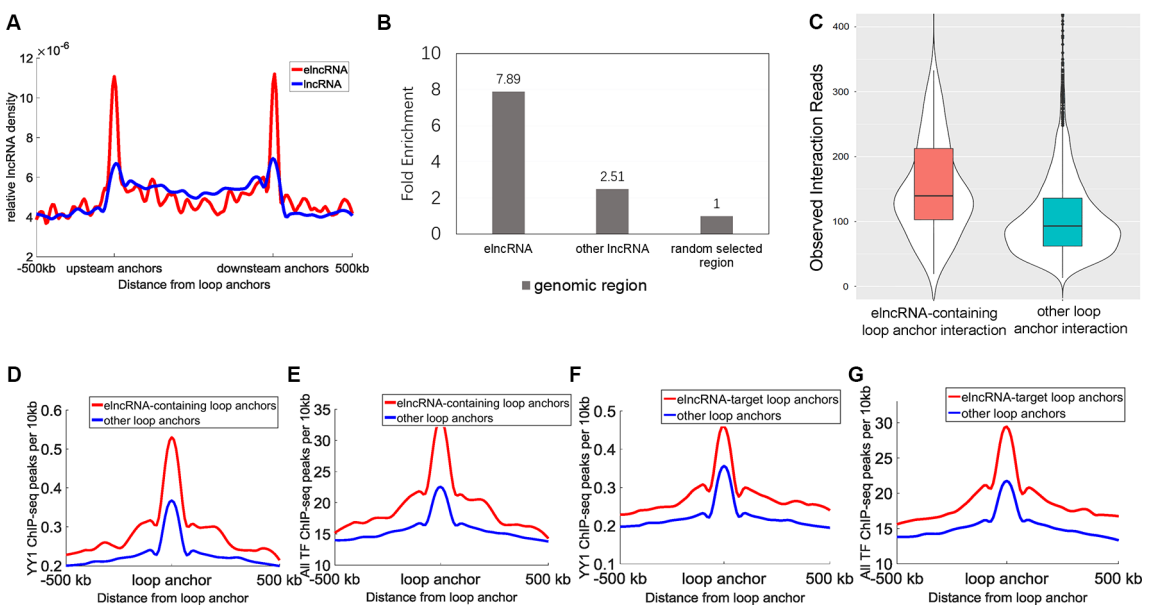


FIGURE 2 | Relationship between elncRNAs and chromatin loops in the GM12878 cell line. **(A)** The distribution of elncRNAs and other lncRNAs across chromatin loops. The y-axis indicates the relative density of lncRNAs. Relative density was calculated from the ratio of the lncRNA counts per 10 kb to the total number of lncRNAs. The red line and blue line indicate elncRNA and other lncRNAs, respectively. **(B)** Enrichment of loop anchors with elncRNAs and other lncRNAs in the GM12878 cell line. The fold-enrichment was calculated by comparing the average counts of loop anchors overlapping per lncRNA to the average counts of loop anchors overlapping per random selected region. **(C)** The boxplot of Hi-C interaction reads between loop anchors. **(D–G)** The distribution of YY1 and the merged TF ChIP-seq peak counts surrounding loop anchors. The red lines and blue lines indicate elncRNA-containing loop anchors and other loop anchors, respectively. The distribution of YY1 **(D)** and the merged TF **(E)** ChIP-seq peak counts surrounding loop anchors. **(F–G)** The distribution of YY1 and the merged TF **(G)** ChIP-seq peak counts surrounding loop anchors.

were defined as elncRNA-target loop anchors. Interestingly, the elncRNA-target loop anchors were also rich in YY1 ChIP-seq peaks and the merged TF ChIP-seq peaks (**Figures 2F**, **G**, Student's *t*-test, $p = 2.57 \times 10^{-24}$ and 2.59×10^{-63} for YY1 and all TF ChIP-seq, respectively). These results suggested that elncRNAs not only influenced loop anchors locally but also potentially affected the target loop anchors through higher-order chromatin structures.

ElncRNAs Are Associated With the Interactions Between Enhancers and Promoters

The average interaction frequency (49.32) of elncRNA-associated enhancer–promoter pairs was significantly higher than that (39.28) of other enhancer–promoter pairs (**Figure 3A**, Student's *t*-test, $p = 1.11 \times 10^{-33}$). Moreover, the expression levels of the target genes of elncRNA-associated enhancers (average FPKM = 58.97) were significantly higher than those of other enhancers (**Figure 3B**, average FPKM = 26.61, Student's *t*-test, $p = 3.77 \times 10^{-27}$), suggesting that the stable interactions of elncRNA-associated enhancer–promoter pairs lead to high expression levels of the target genes.

The elncRNAs were divided into two equal groups with high and low expression levels using the FPKM values; the 50% with the lower FPKM were defined as low expressed elncRNAs, and the 50% with the higher FPKM were defined as high expressed elncRNAs. The interaction frequency (52.88) of high expressed elncRNA-associated enhancer–promoter pairs was significantly higher than that of low expressed elncRNA-associated pairs (48.14, **Figure 3C**, Student's *t*-test, $p = 7.34 \times 10^{-17}$). The expression levels of the target

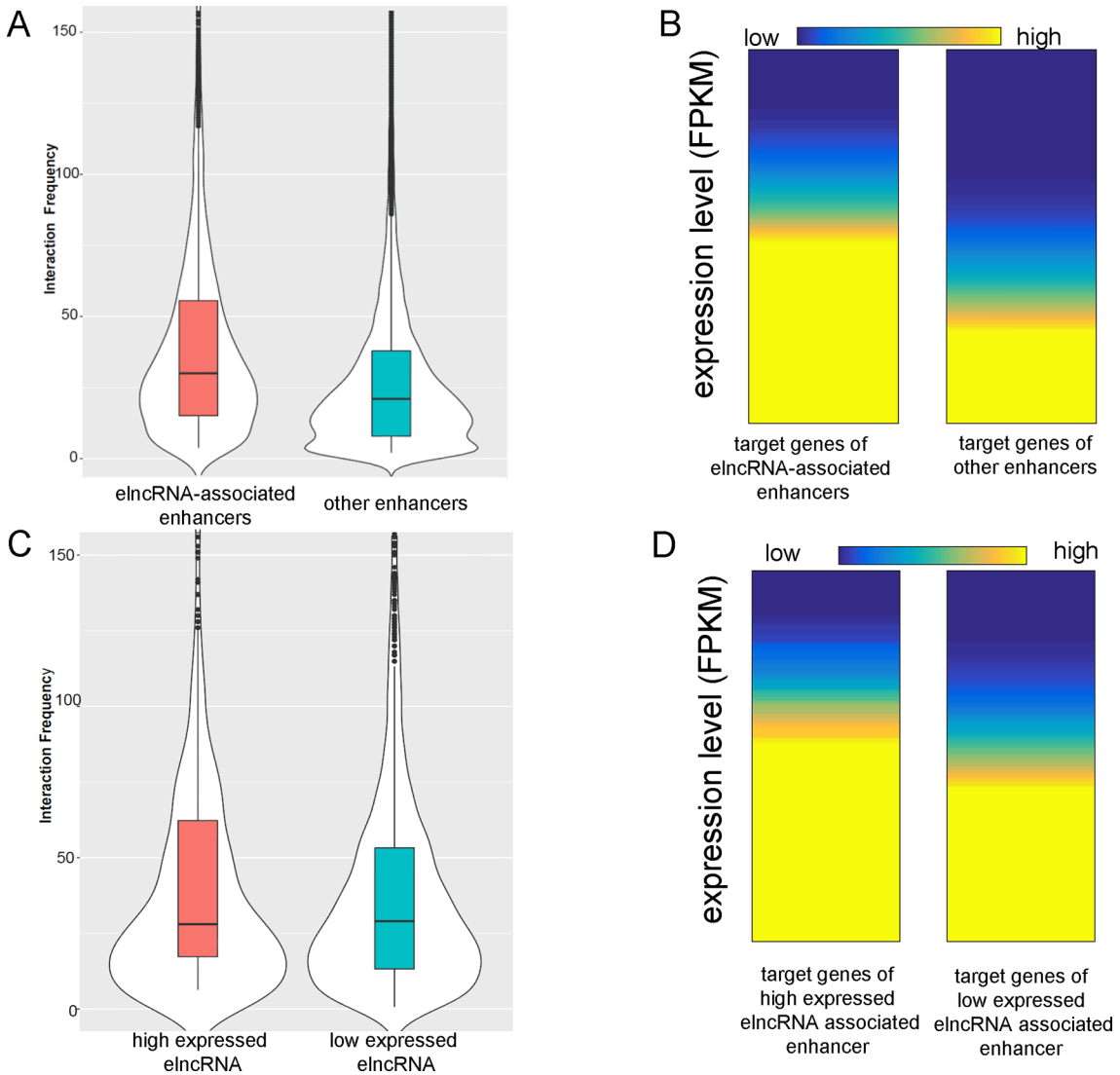


FIGURE 3 | ElncRNAs are associated with enhancer–promoter interactions. **(A)** Interaction frequency of elncRNA-associated enhancer–promoter pairs and other enhancer–promoter pairs. **(B)** Expression levels of target genes of different enhancers. **(C)** Interaction frequency of differently expressed elncRNA-associated enhancer–promoter interactions. **(D)** Expression levels of target genes of elncRNA-associated enhancers.

genes of high expressed elncRNA-associated enhancers (average FPKM = 64.47) were also significantly higher than those of low expressed elncRNA-associated enhancers (**Figure 3D**, average FPKM = 55.79, Student's *t*-test, $p = 4.19 \times 10^{-12}$).

ElncRNAs Are Involved in Enhancer–Promoter Interactions in Different Ways

Using the signal values of structuring factors on enhancer–promoter pairs, we clustered elncRNA-associated enhancer–promoter pairs into 10 groups (**Figure 4A**). In comparison, non-elncRNA enhancer–promoter pairs can be clustered into 6 groups (**Figure 4B**). Although previous research proved that CTCF and cohesin proteins are involved in enhancer–promoter interactions (Li et al., 2015; Tang et al., 2015; Rao et al., 2017), we found that only a small proportion of elncRNA-associated enhancer–promoter interactions (cluster 1 and cluster 10, 20.99%) depended on these architectural proteins. In contrast, most non-elncRNA enhancer–promoter pairs (cluster 1, cluster 2 and cluster 6, 55.92%) were significantly rich in CTCF and cohesin proteins (**Figure 4**, Student's *t-test*, $p = 8.95 \times 10^{-146}$, 1.21×10^{-120}, and 1.94×10^{-102} for CTCF, RAD21, and SMC3, respectively).

To produce elncRNAs, it is reasonable that RNA POLII and YY1 were highly accumulated around elncRNA-associated enhancers. ElncRNA-associated enhancers in clusters 2 and 3 have much higher RNA POLII and YY1 signal values than other enhancers (**Figure 4A**), suggesting that these enhancer–promoter interaction pairs are highly related to RNA POLII and YY1 binding. Intriguingly, the enhancers of non-elncRNA pairs in cluster 3 (**Figure 4B**) also have some RNA POLII signal values, which may contribute to these non-elncRNA enhancer–promoter interactions.

We found that HDGF preferentially localizes at elncRNA-associated enhancers in cluster 4 (**Figure 4A**). HDGF is involved in protein–protein, protein–RNA, and protein–DNA interactions (Zhao et al., 2011; Bao et al., 2014). Our results suggested that elncRNA potentially attracts HDGF to local enhancers and HDGF facilitates enhancer–promoter interactions through protein–protein or protein–DNA interactions. GATAD2B binding sites were abundant on elncRNA-associated enhancers in cluster 5 (**Figure 4A**). Jing et al., (2008) proposed that GATA factors are tightly linked to the chromatin interactions. Our results showed that some enhancer–promoter interaction pairs were associated with GATA factors. GABPA binding sites tend to distribute around elncRNA-associated enhancers in cluster 6. In line with our observations, the binding of GABPA was reported to be capable of mediating long-range chromatin interactions and upregulating transcription (See et al., 2019).

In addition to the TFs discussed above, a large number of elncRNA-associated enhancer–promoter interaction pairs were influenced by other TFs. Compared with non-elncRNA enhancers, most elncRNA-associated enhancers contain enriched TF ChIP-seq peaks, especially in cluster 7. The enhancers and promoters in cluster 7 were brought together by the enriched TFs (**Figure 4A**). Additionally, G-quadruplex sequences were also associated with some enhancer–promoter interaction pairs (**Figure 4A**). In line with our findings, recent research suggested that G-quadruplexes on enhancers and promoters might facilitate enhancer–promoter interactions (Hegyi, 2015; Hou et al., 2019).

Together, these results show that elncRNAs regulate the enhancer–promoter interactions in different ways. Only a fraction (22.9%) of elncRNA enhancer–promoter pairs contained architectural protein binding sites including CTCF, SMC3 and RAD21. Most elncRNA enhancers contained RNA POLII, which

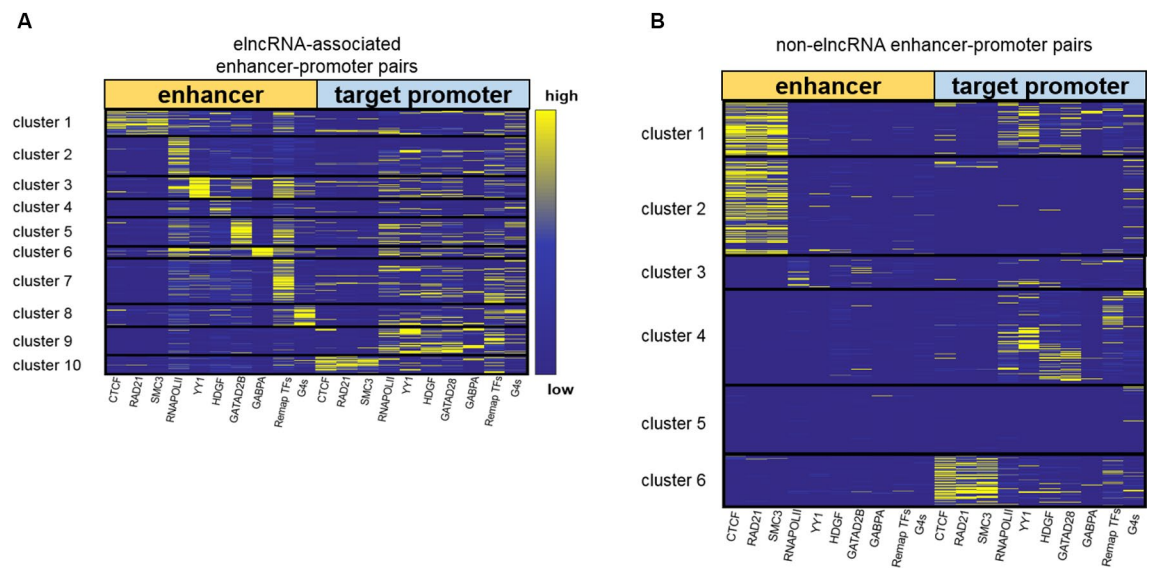


FIGURE 4 | Clustering enhancer–promoter pairs. Heatmap of structuring factor signals on enhancer–promoter pairs. **(A)** ElncRNA-associated enhancer–promoter pairs were clustered into 10 groups using hierarchical clustering according to the various structuring factor signal values. **(B)** Non-elncRNA enhancer–promoter pairs were clustered into 6 groups using hierarchical clustering.

can mediate chromatin interactions and is highly related to elncRNA transcription. YY1, HDGF, GATAD2B, and GABPA were also enriched in parts of elncRNA-associated enhancers. These structuring factors potentially facilitate some elncRNA-associated enhancer–promoter interactions. In addition to the TFs, G-quadruplex sequences, which were highly associated with chromatin structures, were found to be enriched in cluster 8 of elncRNA-associated pairs.

To investigate whether the cluster results were influenced by the number or the choice of structuring factors, we used different numbers of structuring factors to cluster enhancer–promoter interaction pairs (**Supplementary Figures 2A**, **B**). We retained CTCF, RAD21, SMC3, RNA POLII, the merged TFs, and G-quadruplex sequences as the 6 structuring factors. Using these factors, the elncRNA-associated enhancer-promoter pairs can be clustered into 6 groups (**Supplementary Figure 2A**). Because YY1, HDGF, GATAD2B, and GABPA were removed, the pairs in clusters 3–6 of **Figure 4A**, which had enhancers enriched in these TFs, were clustered into different groups according to their structuring factor signal values (**Supplementary Figure 2A**). However, 90.82% of the pairs in the other clusters of **Figure 4A** clustered back into the same groups, in which the enhancers were rich in the architectural protein, RNA POLII, the merged TFs, and G-quadruplex sequences, regardless of whether 6 or 10 structuring factors were used (**Figure 4A** and **Supplementary Figure 2A**). We further added six more structuring factors—NRF1, HSF1, NRSF, MAX, MAZ, and CHD1—to our structuring factor candidates (a total of 16 structuring factors). These TFs are known to be involved in the regulation of chromatin structure (Garriga-Canut et al., 2006; Smolle et al., 2012; Domcke et al., 2015; Sadeghifar et al., 2015; Zhang et al., 2016; Fujimoto et al., 2017). The elncRNA-associated enhancer-promoter pairs were clustered into 11 groups in accordance with the 16 structuring factor signals (**Supplementary Figure 2B**). We found that 82.22% of the elncRNA-associated enhancer-promoter pairs have the same clustering results regardless of whether 10 or 16 structuring factors were used (**Figure 4A** and **Supplementary Figure 2B**). The enhancers in cluster 11 of **Supplementary Figure 2B** were rich in the CHD1 ChIP-seq peaks, but only 1.59% of the elncRNA enhancer-promoter pairs were clustered into cluster 11. Furthermore, the signal values of NRF1, HSF1, NRSF, MAX, and MAZ on elncRNA-associated pairs were quite low and dispersed, indicating that these proteins were only marginally involved in the elncRNA-associated enhancer-promoter interaction pairs. Therefore, only the most commonly used structuring factors (the 10 structuring factors in **Table 1**) were retained.

We calculated the interaction frequency of elncRNA-associated enhancers in the different clusters (**Figure 5A**). Interestingly, elncRNA-associated enhancer–promoter pairs in cluster 8, which were highly associated with G-quadruplex sequences, displayed the highest interaction frequency, suggesting that the enhancer–promoter pairs mediated by G-quadruplexes were quite stable. In addition, elncRNA-associated enhancer–promoter pairs in cluster 7 (**Figure 5A**) also displayed significantly higher interaction frequency than other elncRNA-associated enhancer–promoter pairs (Student's *t*-test, $p = 2.56 \times 10^{-14}$), suggesting a critical role of TFs in enhancer–promoter interactions. However, the elncRNA-associated enhancer–promoter pairs in cluster 6, which were rich in HDGF binding sites, displayed significantly lower interaction frequency (Student's *t*-test, $p = 4.97 \times 10^{-8}$) than other elncRNA-associated enhancer–promoter pairs. It has been reported that the N-terminal PWWP domain of HDGF is required for DNA binding (Yang and Everett, 2007), but PWWP-DNA interactions could be weak and/or unstable (Morchikh et al., 2013). We suspected that the low interaction frequency of enhancer–promoter pairs mediated by HDGF may be explained by the unstable binding of HDGF.

Even though elncRNA-associated enhancer–promoter interaction frequency (cluster 8) was the highest, the elncRNAs in cluster 8 were expressed significantly lower than other elncRNAs (**Figure 5B**, Student's *t*-test, $p = 9.77 \times 10^{-44}$). We suspected that the formation of G-quadruplexes in this cluster serve as a compensation for the low expressed elncRNAs. And the elncRNA-associated enhancer-target genes in cluster 4 and cluster 6 express significantly lower than other enhancer-target genes (Student's *t*-test, $p = 4.49 \times 10^{-22}$ and 1.90×10^{-58} for cluster 4 and cluster 6, respectively), because enhancer–promoter interaction pairs in cluster 4 and cluster 6 were mainly mediated by HDGF and GABPA. HDGF has been reported to function as a transcriptional repressor (Yang and Everett, 2007), suggesting that elncRNAs promote HDGF binding on enhancers which further influence the expression of these enhancer-target genes. GABPA was found to be overrepresented in methylated regions (Hogart et al., 2012). We hypothesized that these interactions mediated by GABPA may be influenced by DNA methylation, which leads to the low expression of these target genes in cluster 6.

We also found that some protein binding sites displayed a strong bias towards the target promoters of elncRNA-associated enhancers, indicating that elncRNA can potentially influence target genes (in *trans*). We showed the ChIP-seq peaks of all merged TFs, YY1, and RNA POLII around the target genes of enhancers in the GM12878 cell line, respectively (**Figures 6A–C**). Compared with the target genes of other enhancers, the target genes of elncRNA-associated enhancers were significantly rich in TFBSs, especially for YY1 and RNA POLII (**Figures 6A–C**, Student's *t*-test, $p = 1.01 \times 10^{-36}$, 1.03×10^{-27}, and 4.42×10^{-29} for all TFs, YY1, and RNA POLII, respectively), suggesting that elncRNAs can influence some proteins, especially for YY1 and RNA POLII, on the target promoters (in *trans*) (**Figures 6A–C**).

G-Quadruplexes Are Associated With the Expression of elncRNAs

It has been reported that G-quadruplexes show hallmarks of dynamic epigenetic features in chromatin primarily found in regulatory, nucleosome-depleted regions and correlate with high expressed genes (Hansel-Hertsch et al., 2016). Because some enhancers can be transcribed to produce elncRNAs, G-quadruplexes on enhancers may also be related to the transcription of enhancers. We suspected that G-quadruplexes on enhancers can facilitate enhancer transcription events. The distribution of G-quadruplex sequences around the enhancers is shown in **Figure 7A**. G-quadruplex sequence counts around elncRNA-associated enhancers were significantly higher than

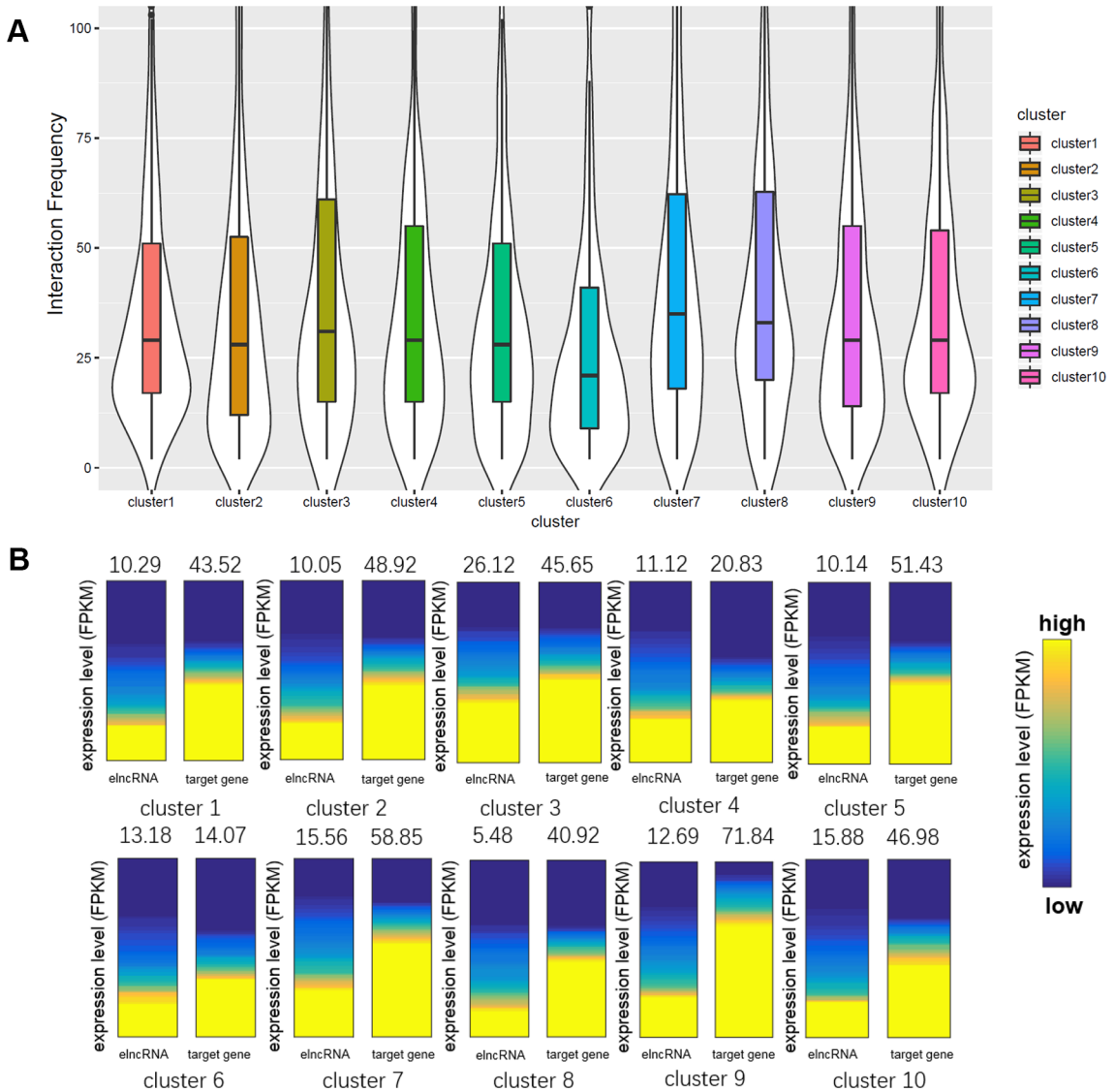


FIGURE 5 | Comparison of the enhancer–promoter pairs in different clusters. **(A)** Boxplot of interaction frequency of different clusters. **(B)** Gene expression analyses of different clusters. The average expression levels were indicated above the heatmaps. The heatmaps represent the expression levels of elncRNAs and their target genes in different clusters. The genes were sorted according to their expression levels.

those around other enhancers (Student's *t*-test, $p = 1.28 \times 10^{-70}$). Moreover, G-quadruplex sequence counts around TSSs of high expressed elncRNAs were significantly higher than those of low expressed elncRNAs (**Figure 7B**, Student's *t*-test, $p = 4.33 \times 10^{-39}$).

We clustered the elncRNAs into two groups (**Figure 7C**). In group 1, G-quadruplex sequences were preferentially localized in promoters of elncRNAs rather than gene body regions. In group 2, G-quadruplex sequences were more likely to be distributed along elncRNA gene body regions. The expression levels of elncRNAs in group 1 were significantly higher than those in group 2 (**Figure 7D**, Student's *t*-test, $p = 7.13 \times 10^{-29}$). Because G-quadruplexes on the gene body may stall elongation of RNAPOLII, high enrichment of G-quadruplex sequences on gene bodies will lead to the low expressed level of elncRNAs. However, G-quadruplex sequences on promoters are highly related to chromatin accessibility, and G-quadruplexes can recruit transcription factors to promoters, which can further promote the expression levels of elncRNAs. We inferred that high abundant G-quadruplex sequences in elncRNA promoters facilitated the steady expression of elncRNA.

DISCUSSION

There is a broad consensus that enhancers can generate non-coding transcripts (Li et al., 2016). Nevertheless, whether these non-coding transcripts are functional or merely a byproduct remains poorly understood. Some studies proved that some specific enhancer RNAs play critical roles in biological processes

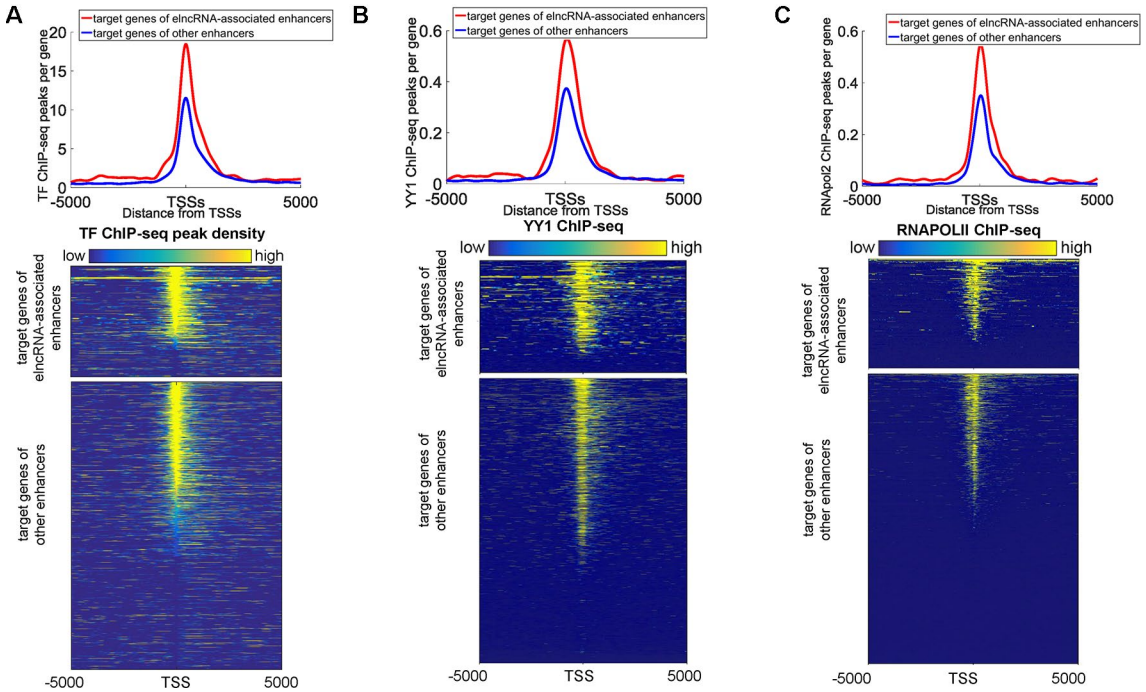


FIGURE 6 | Distribution of ChIP-seq peaks around target genes of elncRNA-associated enhancers and other enhancers, respectively. The red lines and blue lines indicate target genes of elncRNA-associated enhancers and other enhancers, respectively. **(A)** Top panel: the distribution of ChIP-seq peak counts around target genes of elncRNA-associated enhancers and other enhancers. Bottom panel: Heatmap of all TF ChIP-seq peaks around TSSs; each row represents a target gene of enhancers. **(B)** Top panel: Distribution of YY1 ChIP-seq peak counts around target genes of elncRNA-associated enhancers and other enhancers. Bottom panel: Heatmap of YY1 ChIP-seq reads around TSSs. **(C)** Top panel: the distribution of RNA POLII peak counts around target genes of elncRNA-associated enhancers and other enhancers. Bottom panel: Heatmap of RNA POLII ChIP-seq reads around TSSs.

(Lai et al., 2013; Li et al., 2013; Melo et al., 2013; Schaukowitch et al., 2014; Yang et al., 2016a; Tsai et al., 2018; Miao et al., 2018). However, deciphering the function and mechanism of elncRNAs in the genome-wide range remains a challenge. In this study, we characterized elncRNAs in human chromatin structures. Using both GRO-seq and RNA-seq data, we identified active enhancers and elncRNAs of the GM12878 cell line (**Figure 1**). ElncRNAs were significantly enriched in chromatin loop anchors (**Figure 2A**). It is well accepted that loop extrusion should depend on either cohesin slides or ATP-driven motors including transcription and DNA replication (Davidson et al., 2016; Busslinger et al., 2017; Ganji et al., 2018; Vian et al., 2018). We found that chromatin loop anchors are prone to being localized around genomic regions where elncRNAs are expressed (**Figures 2A, B**). Our findings suggested that the transcription of elncRNAs is involved in the formation of chromatin loop structures. Moreover, chromatin loops with anchors containing elncRNAs are more stable than those lacking elncRNAs (**Figure 2C**). The chromatin loop anchors always contain abundant architectural protein binding sites regardless of whether there are elncRNAs on them (**Supplementary Figure 1**). However, the YY1 and RNA POL II ChIP-seq signal values of elncRNA-containing loop anchors were significantly higher than those of other loop anchors. It has been reported that YY1 can be recruited by elncRNAs to active enhancer regions (Sigova et al., 2015) and YY1 can mediate chromatin interactions (Weintraub et al., 2017). Our results suggested that the high enrichment of RNA POLII and TFBSs, especially for YY1 binding sites, promotes the stable interactions between elncRNA-containing anchors (**Figures 2D–G**).

We also found that elncRNAs were potentially involved in maintaining enhancer–promoter interactions in the genome-wide range. The interaction frequency of elncRNA-associated enhancer–promoter pairs was significantly higher than that of other enhancer–promoter pairs (**Figure 3A**). Furthermore, the frequent enhancer–promoter interactions led to significantly higher expression levels of these genes (**Figure 3B**). Additionally, the interaction frequency (52.88) of high expressed elncRNA-associated enhancer–promoter pairs was significantly higher than that of low expressed elncRNA-associated pairs (48.14, **Figure 3C**, Student's *t*-test, $p = 7.34 \times 10^{-17}$). The expression levels of genes interacting with high expressed elncRNA associated enhancers (average FPKM = 64.47) were also significantly higher than those of genes (average FPKM = 55.79) interacting with other enhancers (**Figure 3B**, Student's *t*-test, $p = 4.19 \times 10^{-12}$). Our results suggested that the high expression levels of target genes of elncRNA-associated enhancers might arise from the high enrichment of TFBSs including YY1 on the target promoters (**Figures 6A, B**). As discussed above, YY1 can be recruited by elncRNAs and mediate enhancer–promoter interactions. We inferred that the close association between YY1 and elncRNAs can facilitate the interaction of elncRNA-associated enhancers and their target promoters.

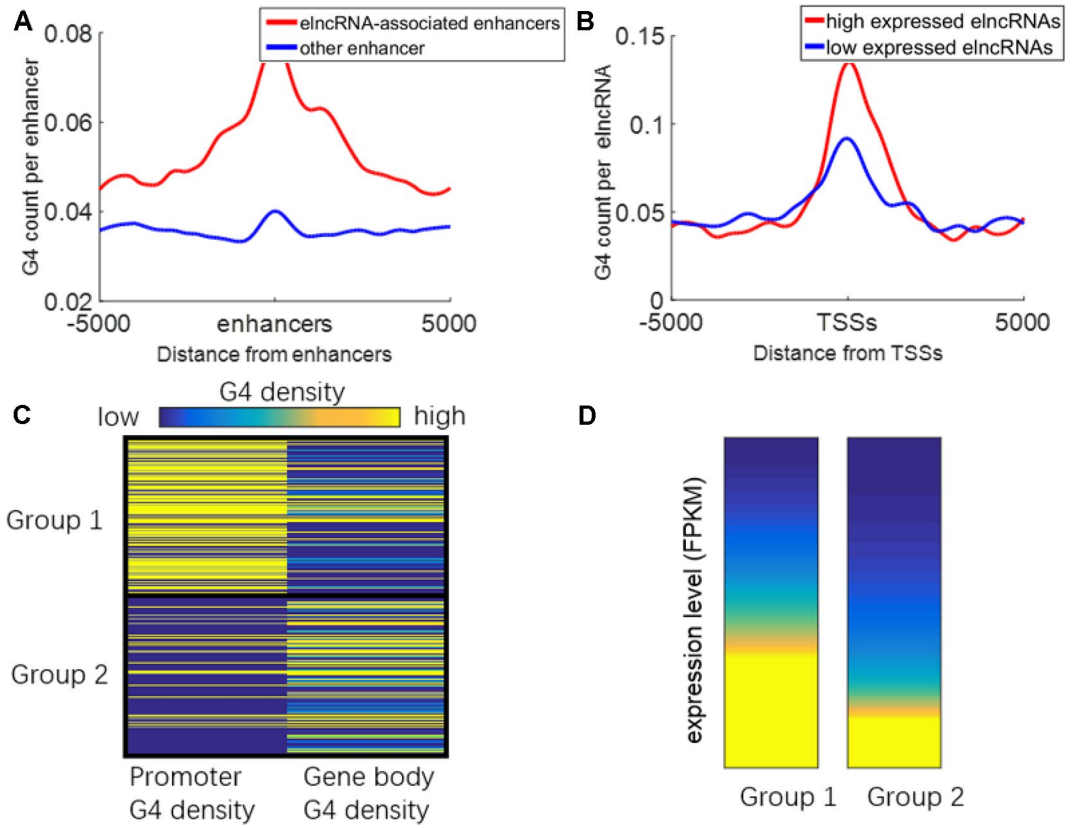


FIGURE 7 | Relationship between G-quadruplexes and elncRNAs. **(A)** Distribution of G-quadruplex sequences around elncRNA-associated enhancers (red) and other enhancers (blue). **(B)** Distribution of G-quadruplex sequence around TSSs of high expressed elncRNAs (red) and low expressed elncRNAs (blue). **(C)** ElncRNAs were clustered into two groups according to the G-quadruplex sequence density of promoters and gene bodies. **(D)** The expression levels of elncRNAs in different groups.

Although our results showed that elncRNAs were highly associated with the high interaction frequency of enhancer-promoter pairs, it remains unclear whether all of these elncRNAs are functional. Because only a few elncRNAs have been proven to be functional with experimental support, further experimental research and more convincing evidence are still needed. In addition, whether elncRNAs have specific distinguishing features compared with other long non-coding transcripts needs further investigation. The causal relationship between enhancer transcripts and enhancer-promoter interactions also requires further study.

To further evaluate the role of elncRNAs in enhancer–promoter interactions, we clustered enhancer–promoter pairs into different groups based on the structuring factor signal values (**Figures 4A**, **B**). The enhancers in elncRNA-associated pairs contained abundant TFBSs. However, the enhancers in non-elncRNA pairs were primarily rich in CTCF and cohesin proteins. Although previous research proved that CTCF and cohesin proteins are important for enhancer–promoter interactions (Li et al., 2015; Tang et al., 2015; Rao et al., 2017), we found that only a small portion of elncRNA-associated enhancer–promoter interactions (cluster 1 and cluster 10, 20.99%) were rich in these architectural proteins (**Figure 4A**). In contrast, more than 55.92% non-elncRNA enhancer–promoter interaction pairs (cluster 1, cluster 2, and cluster 6) are rich in the architectural protein binding sites (**Figures 4A**, **B**, Student's *t*-test, $p = 8.95 \times 10^{-146}$, 1.21×10^{-120}, and 1.94×10^{-102} for CTCF, RAD21, and SMC3, respectively). It was found that elncRNA-associated enhancers in cluster 2 have much higher RNA POLII signal values than other enhancers (**Figure 4A**). Interestingly, the enhancers in cluster 3 of non-elncRNA pairs also contained RNA POLII. We hypothesized that the RNA POLII on the non-elncRNA enhancers was caused by frequent interaction between active genes and these enhancers. Unlike elncRNA-associated pairs, a part of non-elncRNA pairs have almost no structuring factor signal values (cluster 5 in non-elncRNA pairs). In this context, it is possible that these pairs are mainly located in heterochromatin, leading to the lack of TF binding. Compared with other enhancers, elncRNA-associated enhancers contain various TFBSs, suggesting that elncRNAs are involved in enhancer–promoter interactions in different ways. Most elncRNA-associated enhancers contained abundant TF ChIP-seq peaks, which can promote the activity of enhancers and facilitate enhancer–promoter interactions. RNA POLII and YY1, which are able to mediate chromatin interactions and are highly related to elncRNA transcription,

were enriched in parts of elncRNA-associated enhancers. HDGF, GATAD2B, and GABPA also potentially facilitate some enhancer–promoter interactions. In addition, these proteins have been proven to be associated with chromatin interactions by protein-protein interactions or DNA-protein interactions (Jing et al., 2008; Hogart et al., 2012; Bao et al., 2014). G-quadruplexes, the non-canonical secondary structures formed in guanine-rich nucleic acid sequences, are highly associated with gene regulation. We found that G-quadruplex sequences were enriched in cluster 8 of elncRNA-associated pairs. In addition to the identified differences, we also found some similarities. For example, we found that both elncRNA-associated pairs and non-elncRNA pairs include some interaction pairs consisting of the enhancers that lack all of the structuring factors (cluster 9 in elncRNA-associated pairs and cluster 4 in non-elncRNA pairs). However, the target promoters in these pairs contain abundant TFBSs, such as YY1, HDGF, and GATAD2B. The interactions of these pairs may be facilitated by the enriched TFs on the promoters. Our study mainly revealed the association between elncRNAs and the enrichment of TFs on elncRNA-associated pairs. Even though it has been widely accepted that lncRNAs can attract proteins by their specific secondary structure, the causal relationship between elncRNAs and TFs still requires further experimental validation. In addition, the internal mechanism by which different elncRNAs attract different TFs remains unknown.

G-quadruplex sequences, which can form G-quadruplexes in vitro, were significantly accumulated around elncRNA-associated enhancers (**Figure 7A**). Moreover, the levels of G-quadruplex sequences at elncRNA-associated enhancers were significantly higher than other enhancers (**Figure 7A**). The formation of G-quadruplex structures can stabilize the R-loop structures consisting of the nascent RNA and unwound template DNA (Skalska et al., 2017). The high enrichment of G-quadruplexes on elncRNA-associated enhancers is capable of promoting the stability of R-loop structures consisting of elncRNAs and their template DNA. The retained elncRNAs can potentially influence target promoters through enhancer-promoter interactions. Furthermore, G-quadruplex sequence counts around TSSs of high expressed elncRNAs were also significantly higher than those of low expressed elncRNAs (**Figure 7B**). Because G-quadruplexes in promoters are highly associated with elevated transcriptional genes, we hypothesized that high expressed elncRNAs are related to the enrichment of G-quadruplex sequences in their promoters. Although abundant G-quadruplex sequences in elncRNA promoters may be related to the steady expression of elncRNAs, G-quadruplex sequences in gene bodies of elncRNAs may prevent the expression of elncRNAs by stalling the elongation of RNA POLII (**Figures 7C**, **D**).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE60456, GSE90223, GSE63525, GSM935611, GSM935332, GSM935376, GSM803355, GSM803406, GSE91531, GSE105881, GSE63874, GSE96120.

AUTHOR CONTRIBUTIONS

Conceptualization: YH and XS. Methodology and experimentation: YH, RZ, and XS. Writing: YH and XS.

FUNDING

This work was supported by the National Natural Science Foundation of China (61972084, 81830053) and the Key Research and Development Program of Jiangsu Province (BE2016002-3).

ACKNOWLEDGMENTS

We thank Margaret Biswas, PhD, from Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for language editing of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00936/full#supplementary-material

REFERENCES

Ay, F., Bailey, T. L., and Noble, W. S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* 24 (6), 999–1011. doi: 10.1101/gr.160374.113

Bao, C., Wang, J., Ma, W., Wang, X., and Cheng, Y. (2014). HDGF: a novel jack-of-all-trades in cancer. *Future Oncol.* 10 (16), 2675–2685. doi: 10.2217/fon.14.194

Beagrie, R. A., Scialdone, A., Schueler, M., Kraemer, D. C., Chotalia, M., Xie, S. Q., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 543 (7646), 519–524. doi: 10.1038/nature21411

Bianchessi, V., Badi, I., Bertolotti, M., Nigro, P., D'Alessandra, Y., Capogrossi, M. C., et al. (2015). The mitochondrial lncRNA ASncmtRNA-2 is induced in aging and replicative senescence in Endothelial Cells. *J. Mol. Cell Cardiol.* 81, 62–70. doi: 10.1016/j.yjmcc.2015.01.012

Bonnal, R. J., Ranzani, V., Arrigoni, A., Curti, S., Panzeri, I., Gruarin, P., et al. (2015). De novo transcriptome profiling of highly purified human lymphocytes primary cells. *Sci. Data* 2, 150051. doi: 10.1038/sdata.2015.51

Busslinger, G. A., Stocsits, R. R., van der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., et al. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* 544 (7651), 503–507. doi: 10.1038/nature22063

Chambers, V. S., Marsico, G., Boutell, J. M., Di Antonio, M., Smith, G. P., and Balasubramanian, S. (2015). High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* 33 (8), 877–881. doi: 10.1038/nbt.3295

Chen, L., Chen, J. Y., Zhang, X., Gu, Y., Xiao, R., Shao, C., et al. (2017). R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Mol. Cell* 68 (4), 745–757 e5. doi: 10.1016/j.molcel.2017.10.008

Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* 46 (D1), D267–D275. doi: 10.1093/nar/gkx1092

Consortium, Encode Project. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414), 57–74. doi: 10.1038/nature11247

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., and Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46 (12), 1311–1320. doi: 10.1038/ng.3142

Danko, C. G., Hyland, S. L., Core, L. J., Martins, A. L., Waters, C. T., Lee, H. W., et al. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* 12 (5), 433–438. doi: 10.1038/nmeth.3329

Davidson, I. F., Goetz, D., Zaczek, M. P., Molodtsov, M. I., Huis In 't Veld, P. J., Weissmann, F., et al. (2016). Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J.* 35 (24), 2671–2685. doi: 10.15252/embj.201695402

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22 (9), 1775–1789. doi: 10.1101/gr.132159.111

Devaux, Y., Zangrando, J., Schroen, B., Creemers, E. E., Pedrazzini, T., Chang, C. P., et al. (2015). Long noncoding RNAs in cardiac development and ageing. *Nat. Rev. Cardiol.* 12 (7), 415–425. doi: 10.1038/nrcardio.2015.55

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489 (7414), 101–108. doi: 10.1038/nature11233

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L., and Schubeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* 528 (7583), 575–579. doi: 10.1038/nature16462

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., et al. (2010). A three-dimensional model of the yeast genome. *Nature* 465 (7296), 363–367. doi: 10.1038/nature08973

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* 46 (D1), D308–D314. doi: 10.1093/nar/gkx1107

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15 (1), 7–21. doi: 10.1038/nrg3606

Fujimoto, M., Takii, R., Takaki, E., Katiyar, A., Nakato, R., Shirahige, K., et al. (2017). The HSF1-PARP13-PARP1 complex facilitates DNA repair and promotes mammary tumorigenesis. *Nat. Commun.* 8 (1), 1638. doi: 10.1038/s41467-017-01807-7

Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., et al. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science* 360 (6384), 102–105. doi: 10.1126/science.aar7831

Garriga-Canut, M., Schoenike, B., Qazi, R., Bergendahl, K., Daley, T. J., Pfender, R. M., et al. (2006). 2-Deoxy-D-glucose reduces epilepsy progression by NRSF-CtBP-dependent metabolic regulation of chromatin structure. *Nat. Neurosci.* 9 (11), 1382–1387. doi: 10.1038/nn1791

Gertz, J., Savic, D., Varley, K. E., Partridge, E. C., Safi, A., Jain, P., et al. (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol. Cell* 52 (1), 25–36. doi: 10.1016/j.molcel.2013.08.037

Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I., and Chedin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell* 45 (6), 814–825. doi: 10.1016/j.molcel.2012.01.017

Grote, P., Wittler, L., Hendrix, D., Koch, F., Wahrisch, S., Beisaw, A., et al. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev. Cell* 24 (2), 206–214. doi: 10.1016/j.devcel.2012.12.012

Hacisuleyman, E., Goff, L. A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., et al. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.* 21 (2), 198–206. doi: 10.1038/nsmb.2764

Hansel-Hertsch, R., Beraldi, D., Lensing, S. V., Marsico, G., Zyner, K., Parry, A., et al. (2016). G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* 48 (10), 1267–1272. doi: 10.1038/ng.3662

Hao, N., Palmer, A. C., Dodd, I. B., and Shearwin, K. E. (2017). Directing traffic on DNA-How transcription factors relieve or induce transcriptional interference. *Transcription* 8 (2), 120–125. doi: 10.1080/21541264.2017.1285851

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22 (9), 1760–1774. doi: 10.1101/gr.135350.111

Hart, T., Komori, H. K., LaMere, S., Podshivalova, K., and Salomon, D. R. (2013). Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* 14, 778. doi: 10.1186/1471-2164-14-778

He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U. S. A.* 111 (21), E2191–E2199. doi: 10.1073/pnas.1320308111

Hegyi, H. (2015). Enhancer-promoter interaction facilitated by transiently forming G-quadruplexes. *Sci. Rep.* 5, 9165. doi: 10.1038/srep09165

Hogart, A., Lichtenberg, J., Ajay, S. S., Anderson, S., Intramural Sequencing Center, N. I. H., Margulies, E. H., et al. (2012). Genome-wide DNA methylation profiles in hematopoietic stem and progenitor cells reveal overrepresentation of ETS transcription factor binding sites. *Genome Res.* 22 (8), 1407–1418. doi: 10.1101/gr.132878.111

Hong, S., and Kim, D. (2017). Computational characterization of chromatin domain boundary-associated genomic elements. *Nucleic Acids Res.* 45 (18), 10403–10414. doi: 10.1093/nar/gkx738

Hou, Y., Li, F., Zhang, R., Li, S., Liu, H., Qin, Z. S., et al. (2019). Integrative characterization of G-Quadruplexes in the three-dimensional chromatin structure. *Epigenetics*, 1–18. doi: 10.1080/15592294.2019.1621140

Hsieh, C. L., Fei, T., Chen, Y., Li, T., Gao, Y., Wang, X., et al. (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proc. Natl. Acad. Sci. U. S. A.* 111 (20), 7319–7324. doi: 10.1073/pnas.1324151111

Hu, H., Shu, M., He, L., Yu, X., Liu, X., Lu, Y., et al. (2017). Epigenomic landscape of 5-hydroxymethylcytosine reveals its transcriptional regulation of lncRNAs in colorectal cancer. *Br. J. Cancer* 116 (5), 658–668. doi: 10.1038/bjc.2016.457

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47 (3), 199–208. doi: 10.1038/ng.3192

Jing, H., Vakoc, C. R., Ying, L., Mandat, S., Wang, H., Zheng, X., et al. (2008). Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Mol. Cell* 29 (2), 232–242. doi: 10.1016/j.molcel.2007.11.020

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., et al. (2005). Antisense transcription in the mammalian transcriptome. *Science* 309 (5740), 1564–1566. doi: 10.1126/science.1112009

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 (7295), 182–187. doi: 10.1038/nature09033

Kopp, F., and Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* 172 (3), 393–407. doi: 10.1016/j.cell.2018.01.011

Lai, F., Orom, U. A., Cesaroni, M., Beringer, M., Taatjes, D. J., Blobel, G. A., et al. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* 494 (7438), 497–501. doi: 10.1038/nature11884

Lan, X., Witt, H., Katsumura, K., Ye, Z., Wang, Q., Bresnick, E. H., et al. (2012). Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic Acids Res.* 40 (16), 7690–7704. doi: 10.1093/nar/gks501

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9 (4), 357–359. doi: 10.1038/nmeth.1923

Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H. Q., Ong, C. T., et al. (2015). Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell* 58 (2), 216–231. doi: 10.1016/j.molcel.2015.02.023

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., et al. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 498 (7455), 516–520. doi: 10.1038/nature12210

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 (4), 207–223. doi: 10.1038/nrg.2016.4

Li, X., and Fu, X. D. (2019). Chromatin-associated RNAs as facilitators of functional genomic interactions. *Nat. Rev. Genet.* doi: 10.1038/s41576-019-0135-1

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326 (5950), 289–293. doi: 10.1126/science.1181369

Lin, D., Hiron, T. K., and O'Callaghan, C. A. (2018). Intragenic transcriptional interference regulates the human immune ligand MICA. *EMBO J.* 37 (10). doi: 10.15252/embj.201797138

Liu, S. J., Horlbeck, M. A., Cho, S. W., Birk, H. S., Malatesta, M., He, D., et al. (2017). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* 355 (6320). doi: 10.1126/science.aah7111

McHugh, C. A., Chen, C. K., Chow, A., Surka, C. F., Tran, C., McDonel, P., et al. (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* 521 (7551), 232–236. doi: 10.1038/nature14443

Melo, C. A., Drost, J., Wijchers, P. J., van de Werken, H., de Wit, E., Oude Vrielink, J. A., et al. (2013). eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol. Cell* 49 (3), 524–535. doi: 10.1016/j.molcel.2012.11.021

Miao, Y., Ajami, N. E., Huang, T. S., Lin, F. M., Lou, C. H., Wang, Y. T., et al. (2018). Enhancer-associated long non-coding RNA LEENE regulates endothelial nitric oxide synthase and endothelial function. *Nat. Commun.* 9 (1), 292. doi: 10.1038/s41467-017-02113-y

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* 47 (6), 598–606. doi: 10.1038/ng.3286

Morchikh, M., Naughtin, M., Di Nunzio, F., Xavier, J., Charneau, P., Jacob, Y., et al. (2013). TOX4 and NOVA1 proteins are partners of the LEDGF PWWP domain and affect HIV-1 replication. *PLoS One* 8 (11), e81217. doi: 10.1371/journal.pone.0081217

Neppl, R. L., Wu, C. L., and Walsh, K. (2017). lncRNA Chronos is an aging-induced inhibitor of muscle hypertrophy. *J. Cell Biol.* doi: 10.1083/jcb.201612100

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., et al. (2015). RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell* 161 (4), 774–789. doi: 10.1016/j.cell.2015.04.034

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature* 515 (7527), 402–405. doi: 10.1038/nature13986

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159 (7), 1665–1680. doi: 10.1016/j.cell.2014.11.021

Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K. R., et al. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell* 171 (2), 305–320 e24. doi: 10.1016/j.cell.2017.09.026

Ron, G., Globerson, Y., Moran, D., and Kaplan, T. (2017). Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nat. Commun.* 8 (1), 2237. doi: 10.1038/s41467-017-02386-3

Sadeghifar, F., Bohm, S., Vintermist, A., and Ostlund Farrants, A. K. (2015). The B-WICH chromatin-remodelling complex regulates RNA polymerase III transcription by promoting Max-dependent c-Myc binding. *Nucleic Acids Res.* 43 (9), 4477–4490. doi: 10.1093/nar/gkv312

Sallam, T., Sandhu, J., and Tontonoz, P. (2018). Long Noncoding RNA Discovery in Cardiovascular Disease: Decoding Form to Function. *Circ. Res.* 122 (1), 155–166. doi: 10.1161/CIRCRESAHA.117.311802

Schaukowitch, K., Joo, J. Y., Liu, X., Watts, J. K., Martinez, C., and Kim, T. K. (2014). Enhancer RNA facilitates NELF release from immediate early genes. *Mol. Cell* 56 (1), 29–42. doi: 10.1016/j.molcel.2014.08.023

See, Y. X., Wang, B. Z., and Fullwood, M. J. (2019). Chromatin Interactions and Regulatory Elements in Cancer: From Bench to Bedside. *Trends Genet.* 35 (2), 145–158. doi: 10.1016/j.tig.2018.11.007

Shi, X., Sun, M., Liu, H., Yao, Y., and Song, Y. (2013). Long non-coding RNAs: a new frontier in the study of human diseases. *Cancer Lett.* 339 (2), 159–166. doi: 10.1016/j.canlet.2013.06.013

Sigova, A. A., Abraham, B. J., Ji, X., Molinie, B., Hannett, N. M., Guo, Y. E., et al. (2015). Transcription factor trapping by RNA in gene regulatory elements. *Science* 350 (6263), 978–981. doi: 10.1126/science.aad3346

Skalska, L., Beltran-Nebot, M., Ule, J., and Jenner, R. G. (2017). Regulatory feedback from nascent RNA to chromatin and transcription. *Nat. Rev. Mol. Cell Biol.* 18 (5), 331–337. doi: 10.1038/nrm.2017.12

Skourti-Stathaki, K., Proudfoot, N. J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol. Cell* 42 (6), 794–805. doi: 10.1016/j.molcel.2011.04.026

Smolle, M., Venkatesh, S., Gogol, M. M., Li, H., Zhang, Y., Florens, L., et al. (2012). Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange. *Nat. Struct. Mol. Biol.* 19 (9), 884–892. doi: 10.1038/nsmb.2312

St Laurent, G., Wahlestedt, C., and Kapranov, P. (2015). The Landscape of long noncoding RNA classification. *Trends Genet.* 31 (5), 239–251. doi: 10.1016/j.tig.2015.03.007

Stork, C. T., Bocek, M., Crossley, M. P., Sollier, J., Sanz, L. A., Chedin, F., et al. (2016). Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *Elife* 5. doi: 10.7554/eLife.17548

Sun, C., Luan, S., Zhang, G., Wang, N., Shao, H., and Luan, C. (2017). CEBPA-mediated upregulation of the lncRNA PLIN2 promotes the development of chronic myelogenous leukemia *via* the GSK3 and Wnt/beta-catenin signaling pathways. *Am. J. Cancer Res.* 7 (5), 1054–1067.

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., et al. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* 163 (7), 1611–1627. doi: 10.1016/j.cell.2015.11.024

Tehrani, S. S., Karimian, A., Parsian, H., Majidinia, M., and Yousefi, B. (2018). Multiple functions of long non-coding rnas in oxidative stress, dna damage response and cancer progression. *J. Cell Biochem.* 119 (1), 223–236. doi: 10.1002/jcb.26217

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* 31 (1), 46–53. doi: 10.1038/nbt.2450

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 (3), 562–578. doi: 10.1038/nprot.2012.016

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

Tsai, P. F., Dell'Orso, S., Rodriguez, J., Vivanco, K. O., Ko, K. D., Jiang, K., et al. (2018). A Muscle-Specific Enhancer RNA Mediates Cohesin Recruitment and Regulates Transcription In trans. *Mol. Cell* 71 (1), 129–141 e8. doi: 10.1016/j.molcel.2018.06.008

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigo, R., and Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* 19 (9), 535–548. doi: 10.1038/s41576-018-0017-y

Vian, L., Pekowska, A., Rao, S. S. P., Kieffer-Kwon, K. R., Jung, S., Baranello, L., et al. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* 175 (1), 292–294. doi: 10.1016/j.cell.2018.09.002

Wan, G., Zhou, W., Hu, Y., Ma, R., Jin, S., Liu, G., et al. (2016). Transcriptional Regulation of lncRNA Genes by Histone Modification in Alzheimer's Disease. *Biomed. Res. Int.* 2016, 3164238. doi: 10.1155/2016/3164238

Wang, Z., Chu, T., Choate, L. A., and Danko, C. G. (2019). Identification of regulatory elements from nascent transcription using dREG. *Genome Res.* 29 (2), 293–303. doi: 10.1101/gr.238279.118

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

Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., et al. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 171 (7), 1573–1588 e28. doi: 10.1016/j.cell.2017.11.008

Whalen, S., Truty, R. M., and Pollard, K. S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* 48 (5), 488–496. doi: 10.1038/ng.3539

Yan, X., Hu, Z., Feng, Y., Hu, X., Yuan, J., Zhao, S. D., et al. (2015). Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell.* 28 (4), 529–540. doi: 10.1016/j.ccell.2015.09.006

Yang, D., Lian, T., Tu, J., Gaur, U., Mao, X., Fan, X., et al. (2016a). LncRNA mediated regulation of aging pathways in Drosophila melanogaster during dietary restriction. *Aging (Albany NY)* 8 (9), 2182–2203. doi: 10.18632/aging.101062

Yang, F., Deng, X., Ma, W., Berletch, J. B., Rabaia, N., Wei, G., et al. (2015). The lncRNA Firre anchors the inactive X chromosome to the nucleolus by

binding CTCF and maintains H3K27me3 methylation. *Genome Biol.* 16, 52. doi: 10.1186/s13059-015-0618-0

Yang, J., and Everett, A. D. (2007). Hepatoma-derived growth factor binds DNA through the N-terminal PWWP domain. *BMC Mol. Biol.* 8, 101. doi: 10.1186/1471-2199-8-101

Yang, Y., Su, Z., Song, X., Liang, B., Zeng, F., Chang, X., et al. (2016b). Enhancer RNA-driven looping enhances the transcription of the long noncoding RNA DHRS4-AS1, a controller of the DHRS4 gene cluster. *Sci. Rep.* 6, 20961. doi: 10.1038/srep20961

Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., et al. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* 13 (9), R48. doi: 10.1186/gb-2012-13-9-r48

Yu, F., Zhang, G., Shi, A., Hu, J., Li, F., Zhang, X., et al. (2018). LnChrom: a resource of experimentally validated lncRNA-chromatin interactions in human and mouse. *Database (Oxford)* 2018. doi: 10.1093/database/bay039

Zhang, J., Zhang, Z., Chen, Z., and Deng, L. (2017). Integrating Multiple Heterogeneous Networks for Novel LncRNA-disease Association Inference. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2017.2701379

Zhang, K., Li, N., Ainsworth, R. I., and Wang, W. (2016). Systematic identification of protein combinations mediating chromatin looping. *Nat. Commun.* 7, 12249. doi: 10.1038/ncomms12249

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9 (9), R137. doi: 10.1186/gb-2008-9-9-r137

Zhao, J., Yu, H., Lin, L., Tu, J., Cai, L., Chen, Y., et al. (2011). Interactome study suggests multiple cellular functions of hepatoma-derived growth factor (HDGF). *J. Proteomics* 75 (2), 588–602. doi: 10.1016/j.jprot.2011.08.021

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 © 2019 Hou, Zhang and Sun. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 18 October 2019
doi: 10.3389/fbioe.2019.00257

The Construction and Analysis of ceRNA Network and Patterns of Immune Infiltration in Mesothelioma With Bone Metastasis

Runzhi Huang[1,2,3†], Jiawen Wu[1,3†], Zixuan Zheng[3†], Guanghua Wang[1,3], Dianwen Song[4], Penghui Yan[5], Huabin Yin[4], Peng Hu[5], Xiaolong Zhu[5], Haiyun Wang[6], Qi Lv[6], Tong Meng[2,3,4], Zongqiang Huang[5*] and Jie Zhang[1,3*]*

[1] Shanghai First Maternity and Infant Hospital, Tongji University School of Medicine, Shanghai, China, [2] Division of Spine, Department of Orthopedics, Tongji Hospital Affiliated to Tongji University School of Medicine, Shanghai, China, [3] Tongji University School of Medicine, Tongji University, Shanghai, China, [4] Department of Orthopaedics, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [5] Department of Orthopedics, School of Medicine, Shanghai General Hospital, Shanghai Jiaotong University, Shanghai, China, [6] Tongji University School of Life Sciences and Technology, Shanghai, China

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Limei Zhao,
University of Washington,
United States
Juan Xu,
Harbin Medical University, China

***Correspondence:**
Tong Meng
mengtong@medmail.com.cn
Zongqiang Huang
gzhuangzq@163.com
Jie Zhang
jiezhang@tongji.edu.cn

[†]These authors have contributed equally to this work as first authors

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 30 August 2019
Accepted: 24 September 2019
Published: 18 October 2019

Citation:
Huang R, Wu J, Zheng Z, Wang G, Song D, Yan P, Yin H, Hu P, Zhu X, Wang H, Lv Q, Meng T, Huang Z and Zhang J (2019) The Construction and Analysis of ceRNA Network and Patterns of Immune Infiltration in Mesothelioma With Bone Metastasis. Front. Bioeng. Biotechnol. 7:257. doi: 10.3389/fbioe.2019.00257

Background: Mesothelioma is a rare and aggressive tumor. Bone metastasis often occurs in the later stages of this disease along with poor quality of life. Thus, it is important to explore the tumorigenesis and bone metastasis mechanism of invasive mesothelioma. For this purpose, we established two nomograms based on tumor-infiltrating immune cells and ceRNA networks to describe the molecular immunity and the clinical prediction of mesothelioma patients with bone metastasis.

Method: The expression profiles of mRNAs, lncRNAs, and miRNAs of 87 primary mesotheliomas were obtained from the TCGA database; there were four patients with bone metastasis and 83 patients without. We constructed a ceRNAs network based on the differentially expressed RNAs between mesothelioma and bone metastasis. CIBERSORT was used to distinguish 22 immune cell types from the tumor transcriptomes. Kaplan–Meier survival analysis and the Cox proportional hazards model were used to evaluate the prognostic value of each factor. Prognosis-associated immune cells and ceRNAs were applied to establish prediction nomograms. The receiver operating characteristic curves (ROC) and calibration curves were utilized to assess the discrimination and accuracy of the nomogram.

Results: Differential analysis revealed that 20 lncRNAs, 15 miRNAs, and 230 mRNAs were significantly different in mesothelioma samples vs. bone metastasis samples. We constructed the ceRNA network to include 10 protein-coding mRNAs, 8 lncRNAs, and 10 miRNAs. Nine of 28 ceRNAs were found to be significant in the Kaplan–Meier analysis. Out of the 22 cell types, the fraction of dendritic cells resting ($P = 0.018$) was significantly different between the bone metastasis group and the non-bone metastasis group. The ROC and the calibration curves, based on ceRNA networks and tumor-infiltrating immune cells, respectively, suggested acceptable accuracy (AUC of 3-year survival: 0.827, AUC of 5-year survival: 0.840; AUC of 3-year survival: 0.730; AUC of 5-year survival: 0.753). Notably, based on the co-expression patterns between ceRNAs and Immune cells, we

found that the hsa-miR-582-5p, CASP9, dendritic cells resting, ANIX2, T cells CD8, and T cells CD4 memory resting might be associated with the mesothelioma bone metastasis.

Conclusion: Based on ceRNA networks and patterns of immune infiltration, our study provided a valid bioinformatics basis in order to explore the molecular mechanism and predict the possibility of mesothelioma bone metastasis.

Keywords: mesothelioma, bone metastasis, ceRNA network, immune infiltration, prognosis, nomogram

INTRODUCTION

Mesothelioma (MESO) is a rare and aggressive tumor with male predominance (Musk et al., 2017). It is derived from mesothelial cells existing in various organs. The incidence is higher in men rather than women, with the major cause being asbestos exposure (Dalsgaard et al., 2019; van Gerwen et al., 2019). Patients with mesothelioma do not have specific symptoms, which makes early diagnosis difficult. Thus, many mesothelioma patients have distant metastases when the primary tumor is discovered. Apart from pleural, mesothelioma can develop in many other serosal surfaces, including the peritoneum, liver capsule, pericardium, and vagina tunica. Distant metastases are located mostly from local drainage lymph nodes to the serosal surface of multiple organs, such as the lungs, brain, liver, and bone (King et al., 1997).

Traditionally, treatment strategies for mesothelioma are determined by the stage of cancer and the patients' physical condition. Cancer-directed surgery in early-stage mesothelioma can clearly improve survival (Kim et al., 2019). Radiotherapy and chemotherapy are applied to patients individually or are combined with surgery to prevent local recurrence. However, these treatments cannot improve survival dramatically, and no standard second-line therapy can be selected (Vogelzang et al., 2003; Infante et al., 2016; Trovo et al., 2019). Recently, immunotherapy is widely used for many kinds of tumors, mesothelioma included, and immunotherapy does show apparent survival benefits for mesothelioma patients (Brower, 2016). However, it is difficult to predict the benefits of this therapeutic method. Additionally, there is no prediction model for patients with mesothelioma that can be used to predict the possibility of distant bone metastases (Bertoglio et al., 2017).

In this study, we try to solve this issue and explore the molecular mechanism of mesothelioma bone metastasis. As both the competing endogenous RNA (ceRNA) network and immune cell subtypes may predict the prognosis and bone metastasis, we applied the CIBERSORT algorithm to gene expression profiles acquired from the cancer genome atlas (TCGA) to evaluate the cell fraction and establish a ceRNA network. Then, prognosis-associated immune cells and ceRNAs were applied to establish prediction nomograms. Moreover, we assessed the connection between immune cells and ceRNA networks to provide a bioinformatics basis for the discovery of possible molecular pathways.

Abbreviations: TIIC, Tumor infiltrates immune cells; ceRNAs, Competing endogenous RNAs; LncRNA, Long non-coding RNA; MiRNA, microRNA CCLE: Cancer Cell Line Encyclopedia; SD, Standard deviation.

MATERIALS AND METHODS

Data Collection and Differential Gene Expression Analysis

Expression profiles of the primary mesothelioma and of bone metastasis were downloaded from TCGA's (https://tcga-data.nci.nih.gov/tcga/) database, including mRNA, lncRNA, and miRNA. We collected both HTseq-count and Fragments per kilobase of exon per million reads mapped (FPKM) profiles of 87 samples, comprising 83 primary mesotheliomas without bone metastasis and four primary tumors with bone metastasis. Demographic information and the survival endpoints of patients were also retrieved. After filtering non-mesothelioma specific expression genes (no expression was detected in neither the experimental group nor the control group), differences in the expression of each RNA between mesothelioma and bone metastasis were analyzed using the DEseq2. With a false discovery rate (FDR) adjusted P value < 0.05, the log(fold-change) > 1.0 or < -1.0 was defined as a downregulated or upregulated gene, respectively. The relevant data provided by TCGA are publicly available.

Construction of the ceRNA Network

Before primary statistical analysis, the miRNA–mRNA interaction information based on experimental verification was predicted using the miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/) (Chou et al., 2018), while the lncbase v.2 Experimental Module (http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=lncbasev2%2Findex-experimental) was used to predict lncRNA–miRNA interaction (Paraskevopoulou et al., 2016). Then, miRNAs regulated for both lncRNAs and mRNAs, showing significant results in hypergeometric testing and correlation analysis, were selected for visualization of the ceRNA network using Cytoscape v.3.5.1 (Shannon et al., 2003).

Survival Analysis and Nomograms of Key Members in the ceRNA Network

We used Kaplan–Meier survival analysis curves and the Cox proportional hazards model to evaluate the prognostic value of all biomarkers. All significant biomarkers were integrated into the Cox regression model and the Lasso regression was employed to ensure that the multivariate models were not overfitting. Eventually, we built a nomogram based on the multifactor models to predict the prognosis of patients with mesothelioma. The receiver operating characteristic curves (ROC) and calibration curves were utilized to assess the discrimination and accuracy of the nomogram.

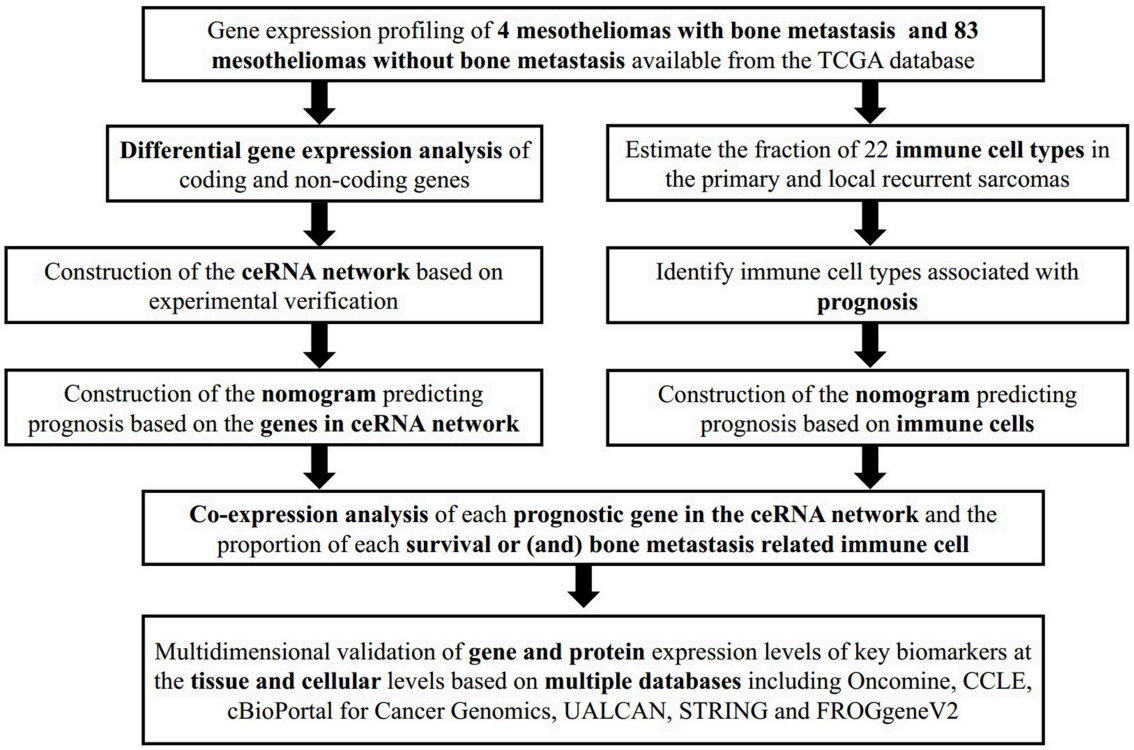


FIGURE 1 | The flow chart of the analytical process.

CIBERSORT Estimation

To estimate the proportion of infiltrating immune cells, standard annotated gene expression data were uploaded to the CIBERSORT web portal (http://cibersort.stanford.edu/), with LM22 signature and 1,000 permutation applied to the algorithm. Only cases with CIBERSORT $P < 0.05$ were considered eligible for subsequent Kaplan-Meier analysis.

Survival Analysis and Nomograms of Key Members of the Immune Cells

We used Kaplan–Meier survival analysis and Cox regression to detect the prognosis-associated cell types. The Wilcoxon rank-sum test was used to evaluate the association between cell subtypes and metastasis, with the outcome determined as the TNM stage. Immune cells showing significant association with prognosis in the initial Cox model were selected to build the nomogram. The receiver operating characteristic (ROC) curve was used to assess the sensitivity and specificity of the diagnostic and prognostic model, and these were quantified by the area under the ROC curve (AUC). The predictive accuracy of the nomogram was assessed by calibration curve and the concordance index (C-index). The relationships among 22 immune cells and between ceRNAs and immune cells were calculated using Pearson correlation coefficients.

Only the two-sided P value < 0.05 was considered to be of statistical significance. All statistical analysis was implemented with R version 3.5.1 software (Institute for Statistics and Mathematics, Vienna, Austria; www.r-project.org) [Package: GDCRNATools (Li et al., 2018), edgeR, ggplot2, rms, glmnet, preprocessCore, survminer, timeROC].

Multidimensional Validation

To minimize bias caused by small study cohorts, multiple databases were used to explore the gene and protein expression levels of key biomarkers at tissue and cellular levels on the Oncomine (Contributors, 2006), cBioPortal for Cancer Genomics (Cerami et al., 2012; Gao et al., 2013), UALCAN (Chandrashekar et al., 2017), STRING (Szklarczyk et al., 2019), PROGgeneV2 (Goswami and Nakshatri, 2014) databases. The whole analytical process of this study is shown in **Figure 1**.

RESULTS

Identification of Significant Differential Genes

We used the log(fold-change) > 1.0 or < -1.0 and FDR < 0.05 as cutoffs to identify differential RNA profiles. The baseline characteristics of all the patients available from the TCGA are described in **Table S1**. In a total of 14,447 lncRNAs, 2,588 miRNAs and 19,660 mRNAs from the TCGA database (**Figures 2A,B**), there are 15 differentially expressed miRNAs (2 downregulated and 13 upregulated). **Figure 2** shows 230 differentially expressed protein-coding genes (55 downregulated and 175 upregulated) (**Figures 2E,F**) and 20 differentially

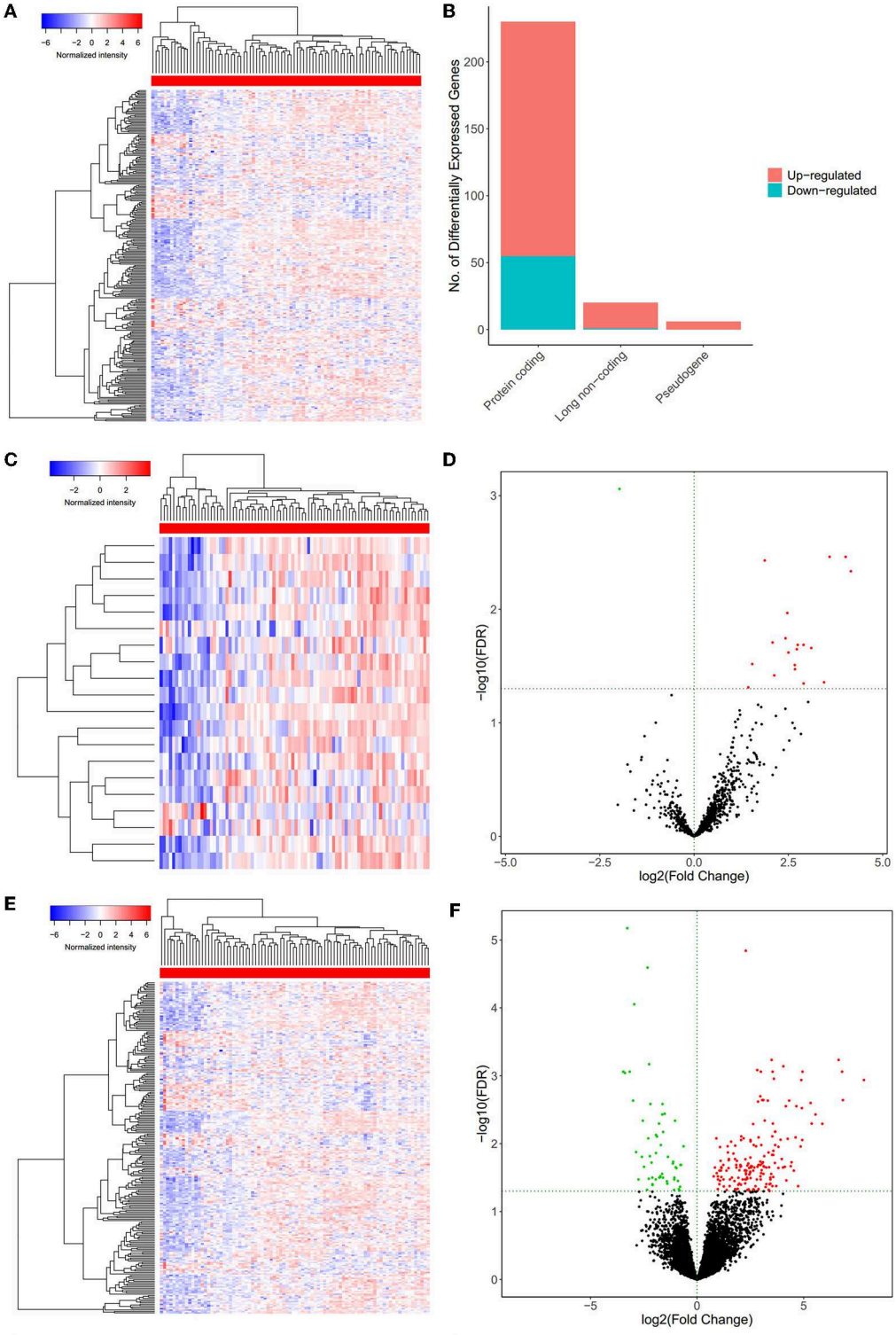


FIGURE 2 | The differentially expressed messenger RNAs (mRNAs) between primary mesothelioma and bone metastasis mesothelioma. The heatmap of differentially expressed mRNAs between primary mesothelioma and bone metastasis **(A)**; the composition of differentially expressed genes **(B)**; the heatmap **(C);** and the volcano Plot **(D)** of differentially expressed protein-coding genes between primary mesothelioma and bone metastasis; the heatmap **(E);** and the volcano Plot **(F)** of differentially lncRNAs between primary mesothelioma and bone metastasis.

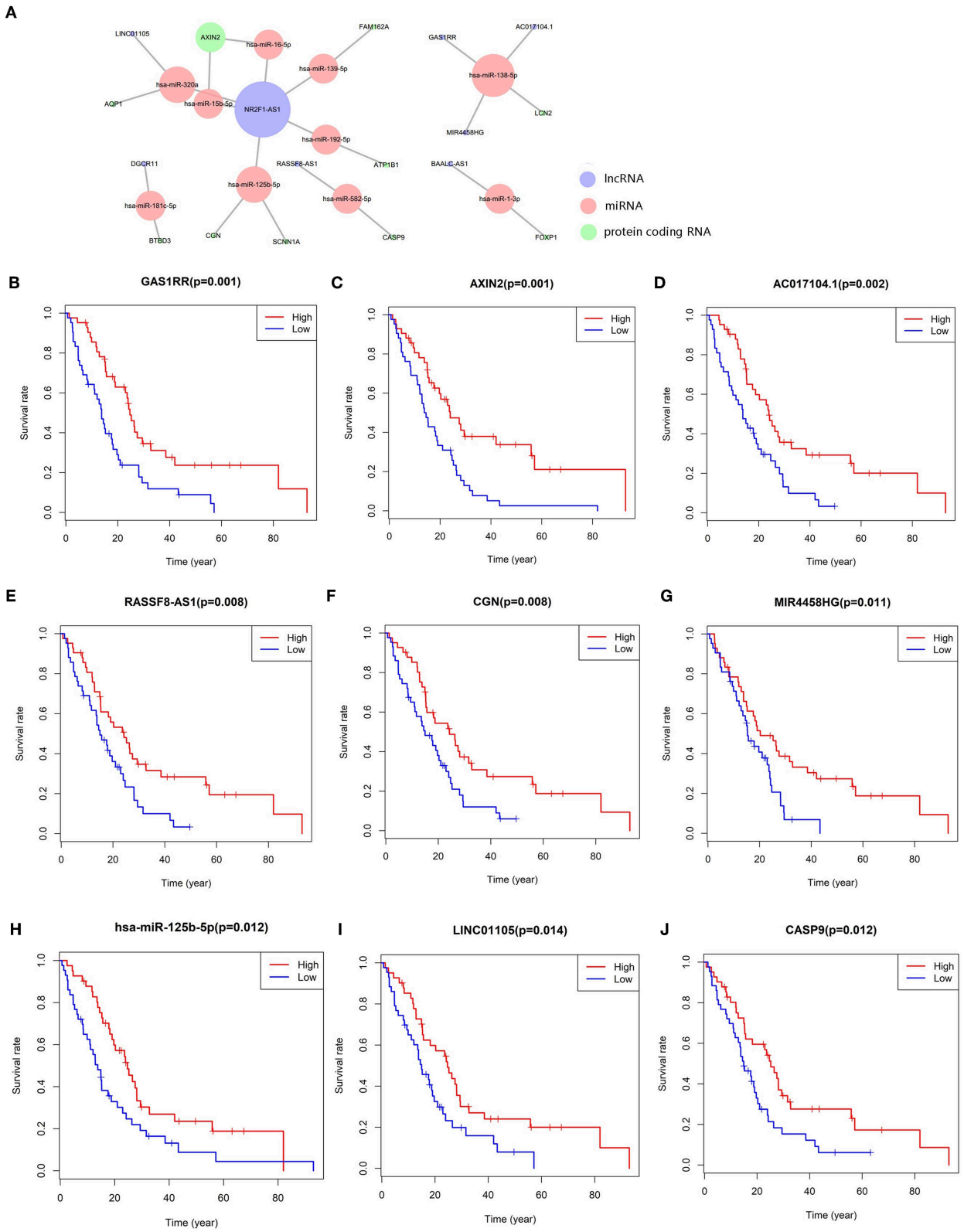


FIGURE 3 | The meso-bone metastasis related ceRNA network **(A)**; the Kaplan–Meier survival curves of the key members of the ceRNA network: GAS1RR **(B)**, AXIN2 **(C)**, AC017104.1 **(D)**, RASSF8-AS1 **(E)**, CGN **(F)**, MIR4458HG **(G)**, hsa-miR-125b-5p **(H)**, LINC01105 **(I)**, and CASP9 **(J)**.

TABLE 1 | Hypergeometric testing and correlation analysis results of ceRNAs network.

LncRNA	Protein-coding RNA	MiRNA	Correlation *P*	Hypergeometric test *P*
DGCR11	BTBD3	hsa-miR-181c-5p	8.11E-03	6.43E-03
RASSF8-AS1	CASP9	hsa-miR-582-5p	4.25E-02	1.05E-06
MIR4458HG	LCN2	hsa-miR-138-5p	1.44E-02	9.65E-03
GAS1RR	LCN2	hsa-miR-138-5p	4.37E-03	8.04E-03
BAALC-AS1	FOXP1	hsa-miR-1-3p	1.24E-02	9.92E-06
AC017104.1	LCN2	hsa-miR-138-5p	6.24E-04	1.69E-10
LINC01105	AQP1	hsa-miR-320a	1.50E-02	6.52E-06
NR2F1-AS1	SCNN1A	hsa-miR-125b-5p	1.94E-02	1.80E-03
NR2F1-AS1	CGN	hsa-miR-125b-5p	1.94E-02	1.37E-02
NR2F1-AS1	AQP1	hsa-miR-320a	1.94E-02	4.03E-02
NR2F1-AS1	FAM162A	hsa-miR-139-5p	1.94E-02	2.05E-02
NR2F1-AS1	ATP1B1	hsa-miR-192-5p	1.94E-02	1.52E-03
NR2F1-AS1	AXIN2	hsa-miR-15b-5p,hsa-miR-16-5p	1.48E-02	4.61E-05

ceRNAs, Competing endogenous RNAs; LncRNA, Long non-coding RNA; MiRNA, microRNA. Pearson correlation analysis was performed for lncRNA in the upstream and mRNA in the downstream, and the P value <0.05 was shown in Table.

expressed lncRNAs (19 downregulated and 1 upregulated) (**Figures 2C,D**).

The Construction of the ceRNA Network and Survival Analysis

We constructed the ceRNA network, which includes 28 genes based on the interactions of 13 lncRNA–miRNA pairs and 10 miRNA–mRNA pairs (**Figure 3A**; **Table 1**). We used Cox regression, Kaplan–Meier and the log-rank test to examine the relationship between the biomarkers in the bone metastasis ceRNA network and the prognosis. GAS1RR ($P = 0.001$), AXIN2 ($P = 0.001$), AC017104.1 ($P = 0.002$), RASSF8-AS1 ($P = 0.008$), CGN ($P = 0.008$), MIR4458HG ($P = 0.011$), hsa-miR-125b-5p ($P = 0.012$), linc01105 ($P = 0.014$), and CASP9 ($P = 0.012$) were found to be significant in the Kaplan–Meier analysis (**Figures 3B–J**). Six potential prognosis-related biomarkers were regarded as key members of the ceRNA network and were integrated into a new multivariable model (**Figure 4A**). The nomogram was constructed based on the model (**Figure 4E**). The results of the Lasso regression revealed that all six genes were essential for modeling (**Figures 4B,C**). Furthermore, the ROC and the calibration curves suggested an acceptable accuracy (AUC of 3-year survival: 0.827; AUC of 5-year survival: 0.840) and discrimination of the nomogram (**Figures 4D,F**).

The Composition of Immune Cells in MESO

Figures 5A,B displayed the proportion of 22 immune cells detected by the CIBERSORT algorithm. The violin plot (**Figure 5C**) depicted results of the Wilcoxon rank-sum test, which showed that the fraction of dendritic cells resting ($P = 0.018$) was significantly different between the bone metastasis group and the non-bone metastasis group.

Clinic Correlation and Nomogram of Immune Cells

We used the non-parameter test and Kaplan–Meier survival analysis to examine the association between the fraction of different immune cell subtypes and the prognosis. The fraction of T cells CD4 memory resting was significantly different among four tumor stages of cancer ($P = 0.027$, **Figure 6A**). The fraction of eosinophils ($P = 0.017$, **Figure 6B**) and mast cells activated ($P = 0.044$, **Figure 6C**) was significantly different among T stages. The fraction of T cells CD8 ($P = 0.080$, **Figure 6D**) and Dendritic cells activated ($P = 0.003$) was found to be significantly associated with overall survival (**Figure 6E**).

Six potential prognosis-related biomarkers were regarded as key members among 22 immune cell subtypes and were integrated into a new multivariable model (**Figure 7A**). The nomogram was constructed based on this model (**Figure 7B**). All the cases were identified as high or low risk groups according to the nomogram model. The proportion of immune cells and the survival of each group are depicted in **Figure 7C**. The ROC and the calibration curve demonstrated the nomogram's good accuracy (AUC of 3-year survival: 0.730; AUC of 5-year survival: 0.753) and concordance (**Figures 7D,E**). The result of Kaplan–Meier analysis showed the significant difference between the high and low risk group (**Figure 7F**).

The Co-expression Analysis

Significant co-expression patterns between proportions of immune cells (**Figure 8A**) and ceRNA-immune cells (**Figure 8B**) were analyzed via Pearson correlation analysis. The fraction of T cells CD4 memory resting is negatively associated with CD8 T cells ($R = -0.62$, $P < 0.001$) and positively associated with AXIN2 expression ($R = 0.36$, $P < 0.001$). The fraction of dendritic cells activated is positively associated with the expression of CASP9 ($R = 0.28$, $P = 0.010$) and RASSF8-AS1 ($R = 0.26$, $P = 0.020$). The fraction of T cells CD8 is negatively associated with the expression of AXIN2 ($R = -0.32$, $P < 0.001$) and RASSF8-AS1 ($R = -0.26$, $P = 0.020$) (**Figures 8C–H**).

DISCUSSION

Malignant mesothelioma, deriving from mesothelial cells, is a tumor type with aggressive invasion and poor prognosis. Distant metastasis, especially for bone metastasis, is often found at the late stage, but its molecular mechanism is still unclear. During tumor initiation and metastasis, molecular, and cellular characteristics play a crucial part and are regarded as prognostic factors (Rodina et al., 2016). Among them, ceRNAs network and tumor-infiltrating immune cells attract our attention. In this study, we first discovered the significant tumor-infiltrating immune cells and ceRNAs between primary mesothelioma and bone metastasis and established two prediction models. The high AUC values of both nomograms proved their clinical application.

The ceRNA networks consist of protein-coding mRNAs and ncRNAs, such as miRNAs and lncRNAs (Salmena et al., 2011). In the present study, we utilized bioinformatics analysis to

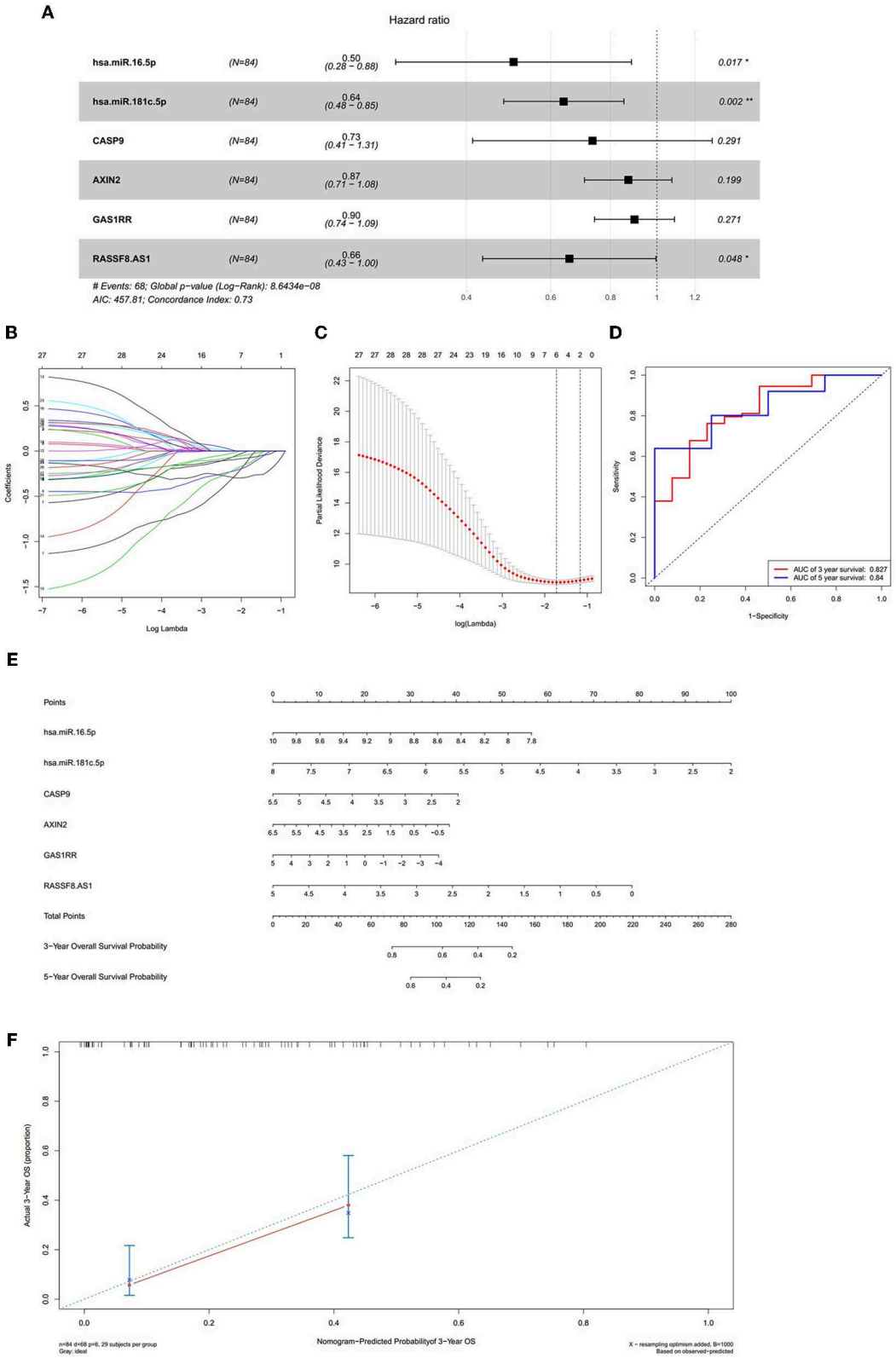


FIGURE 4 | The results of the multivariate Cox regression **(A)**, the results of the Lasso regression **(B,C)**; the ROC curves **(D)**; the nomogram **(E)**; the discrimination of nomogram **(F)**. The results of the Lasso regression **(B,C)** suggested that all six genes were essential for modeling. The nomogram **(E)** was constructed based on the model. The ROC and the calibration **(D,F)** indicated the acceptable accuracy [Area Under Curve (AUC) of 3-year survival: 0.827; AUC of 5-year survival: 0.84] and discrimination of the nomogram.

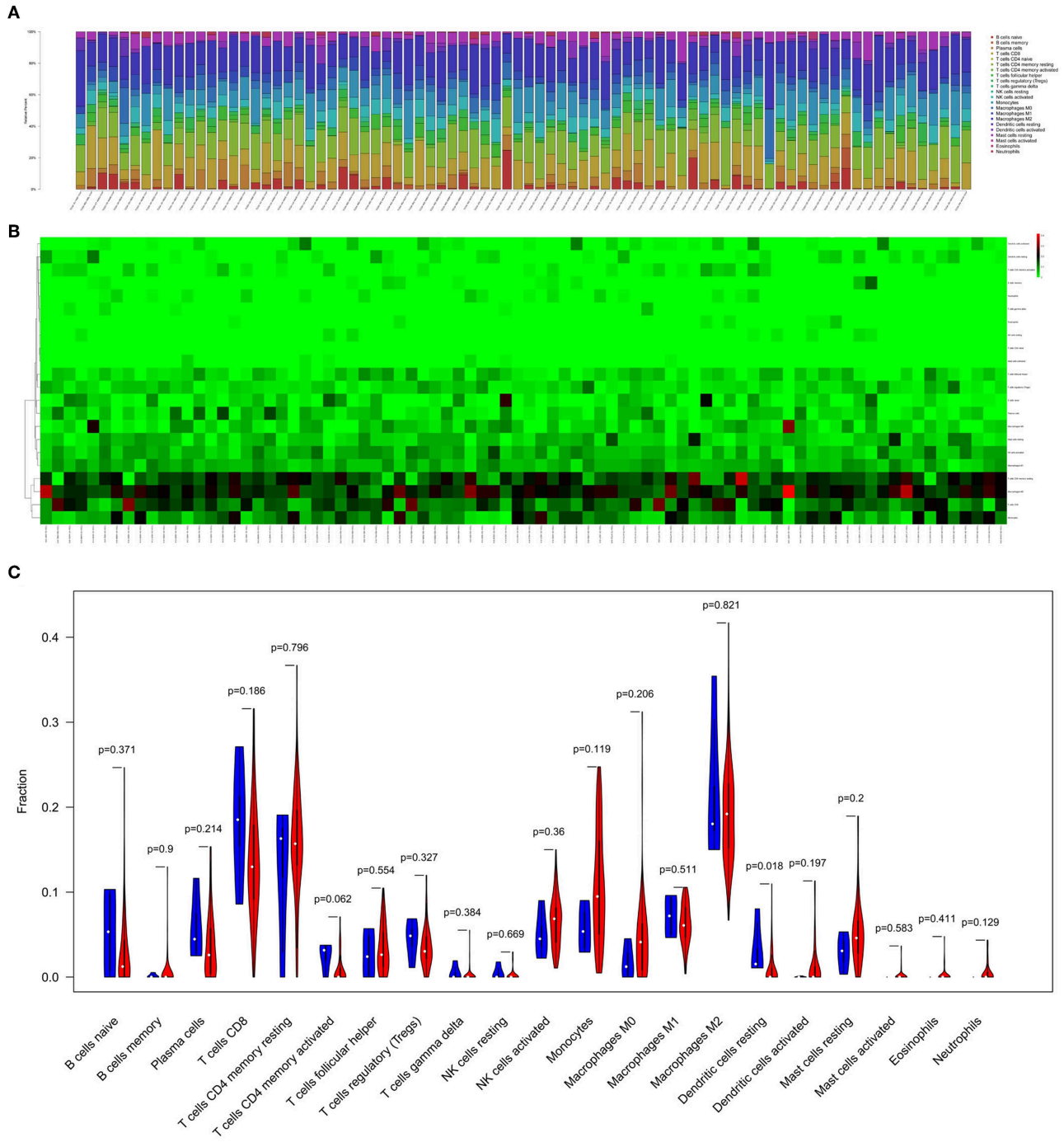


FIGURE 5 | The composition **(A)** and heatmap **(B)** of immune cells estimated by CIBERSORT algorithm in mesothelioma. The violin plot of immune cells **(C)** and the blue and red bar represent recurrent the tumor group and primary tumor group, respectively.

identify the ceRNA networks that regulate bone metastases from mesothelioma with 10 protein-coding mRNAs, 8 lncRNAs, and 10 miRNAs. In the ceRNA networks, GAS1RR, AXIN2, AC017104.1, RASSF8-AS1, CGN, MIR4458HG, has-miR-125b-5p, LINC01105, and CASP9 were significantly associated with overall survival in mesothelioma. The prediction nomogram was constructed and the AUC value of 3-year survival and 5-year survival was 0.827 and 0.840, respectively.

According to the hypergeometric testing and correlation analysis, the results of the ceRNAs network revealed that hsa-miR-582-5p (miRNA), CASP9 (protein-coding RNA), and RASSF8-AS1 (LncRNA) were correlated ($P = 0.486$).

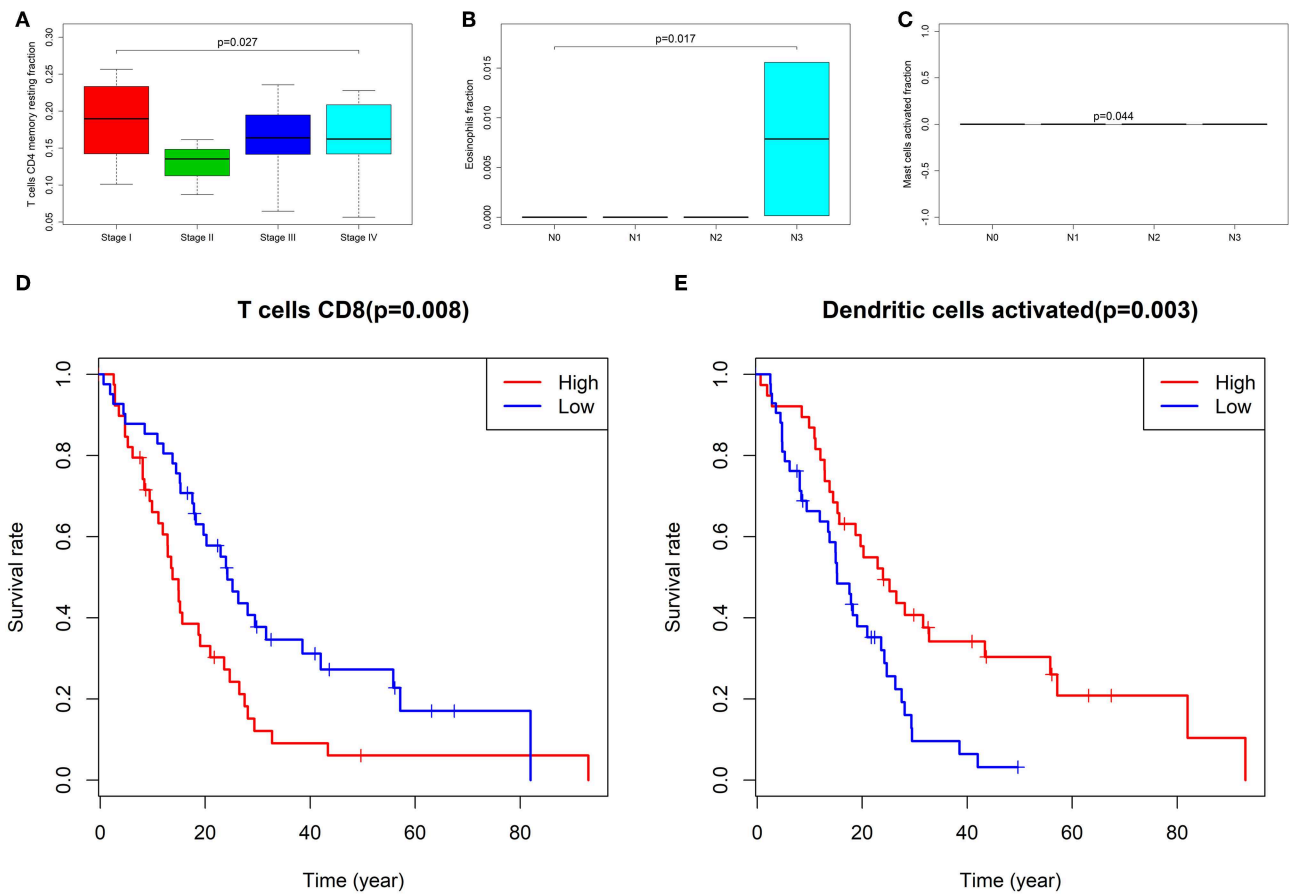


FIGURE 6 | The fraction of T cells CD4 memory resting between four stages of cancer **(A)**; the fraction of eosinophils **(B)** and Mast cells activated **(C)** between T groups; the Kaplan–Meier survival curves of Fraction of T cells CD8 **(D)** and Dendritic cells activated **(E)**.

Subsequently, the correlation analysis between ceRNAs and tumor-infiltrating immune cells discovered that both CASP9 and RASSF8-AS1 were positively correlated with Dendritic cells activated. Thus, we focused on CASP9 and RASSF8-AS1 in our study.

RASSF8-AS1 is an endogenous, unspliced long noncoding RNA (lncRNA) transcribed from the opposite strand of the RASSF8 gene. Divergent lncRNAs can perform biological processes related to the protein through the regulation of the transcription of adjacent protein-coding genes, which is a widely existing new model of gene expression regulation (Luo et al., 2016). RASSF8 is widely expressed in all major organs and tissues, and endogenous RASSF8 is expressed in both the cell membrane and the nucleus. It promotes cell–cell adhesion by maintaining the stability of the adhesion junction (Lock et al., 2010). Zhang et al. discovered that RASSF8 downregulation promoted lymph angiogenesis and metastasis in esophageal squamous-cell carcinomas. Tumor cells with low RASSF8 expression had higher migratory ability and promoted lymph angiogenesis both *in vitro* and *in vivo* (Zhang et al., 2015). They were also associated with tumorigenesis and metastasis in gastric cancer and malignant thyroid neoplasms (Li et al., 2013; He et al., 2017). RASSF8-AS1 may be involve in mesothelioma metastasis through the downregulation of RASSF8 as a cis-acting element. CASP9 is thought to play a central role in apoptosis and to be a tumor suppressor. Dendritic cells (DCs), known as major antigen-presenting cells, can initiate and direct adaptive immune responses with their surface expression of pattern recognition receptors. DC vaccination acts as a promising approach to further promote cancer immunotherapy. However, some studies demonstrated that cancer cells might still influence DCs to improve an immunosuppressive phenotype. It has been reported that pDCs could recruit other immunosuppressive immune cells comprising myeloid-derived suppressor cells (MDSCs) and Tregs, thereby promoting tumor progression and metastasis (Shurin et al., 2011).

MiRNAs are able to bind to the 3′ untranslated region (3′ UTR) of the target mRNAs in a complementary base-pairing manner, which has been demonstrated to contribute to the suppression and induction of oncogenesis (Huang et al., 2014). It has been reported that miR-582-5p interacted with the CASP9 mRNA 3′ UTR and induced a downregulation of CASP9 expression, which was consistent with our present study (Floyd et al., 2014). One of the main functions of

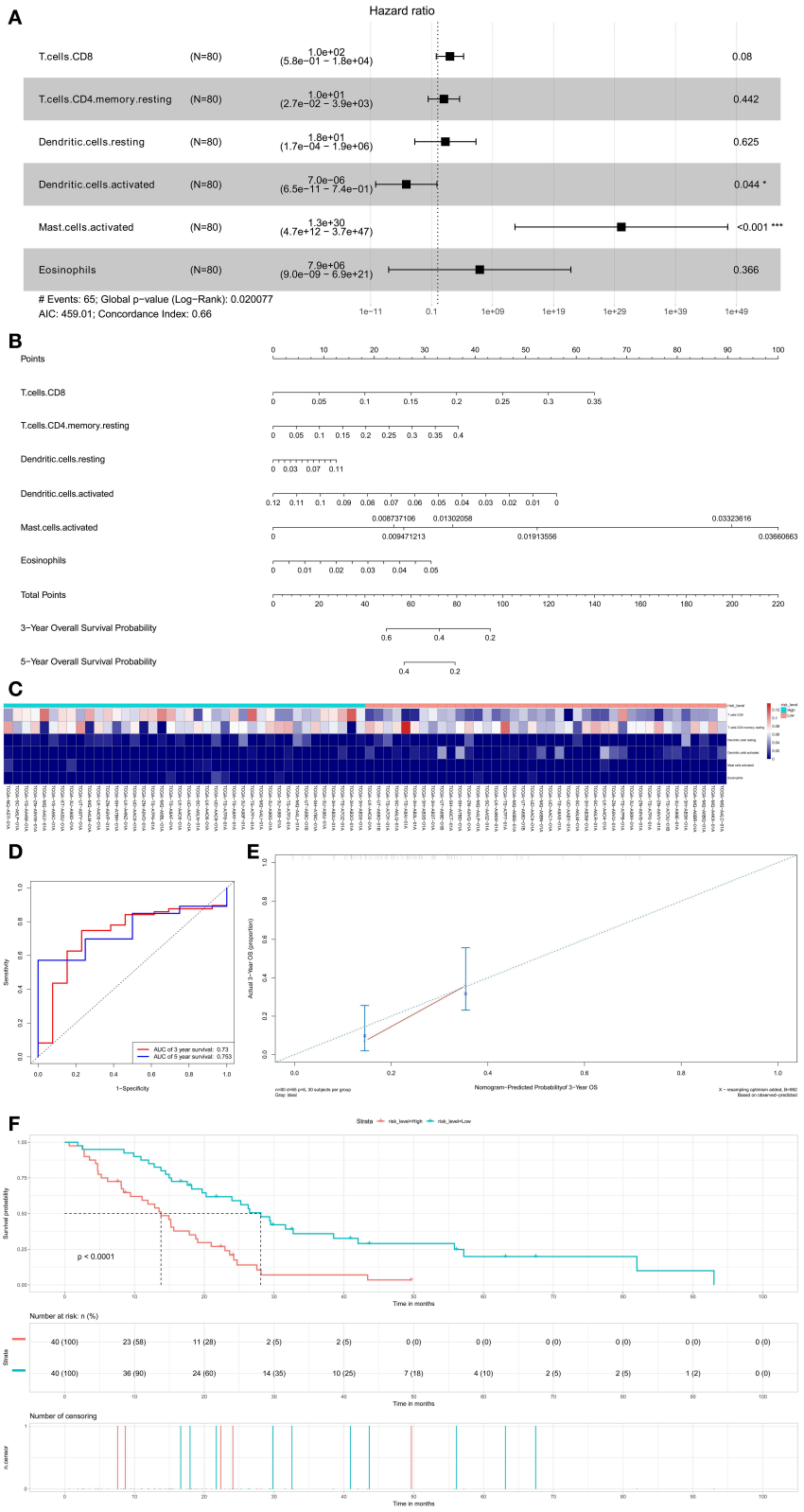


FIGURE 7 | The results of the multivariate Cox regression **(A)** based on prognosis-related immune cells nomogram **(B)**; the ROC curves **(C)**; the heatmap of the six immune cells in Cox regression model **(D)**; Nomogram-Predicted probabilitu of 3-year overall survival **(E)**; the Kaplan–Meier survival curve **(F)**.

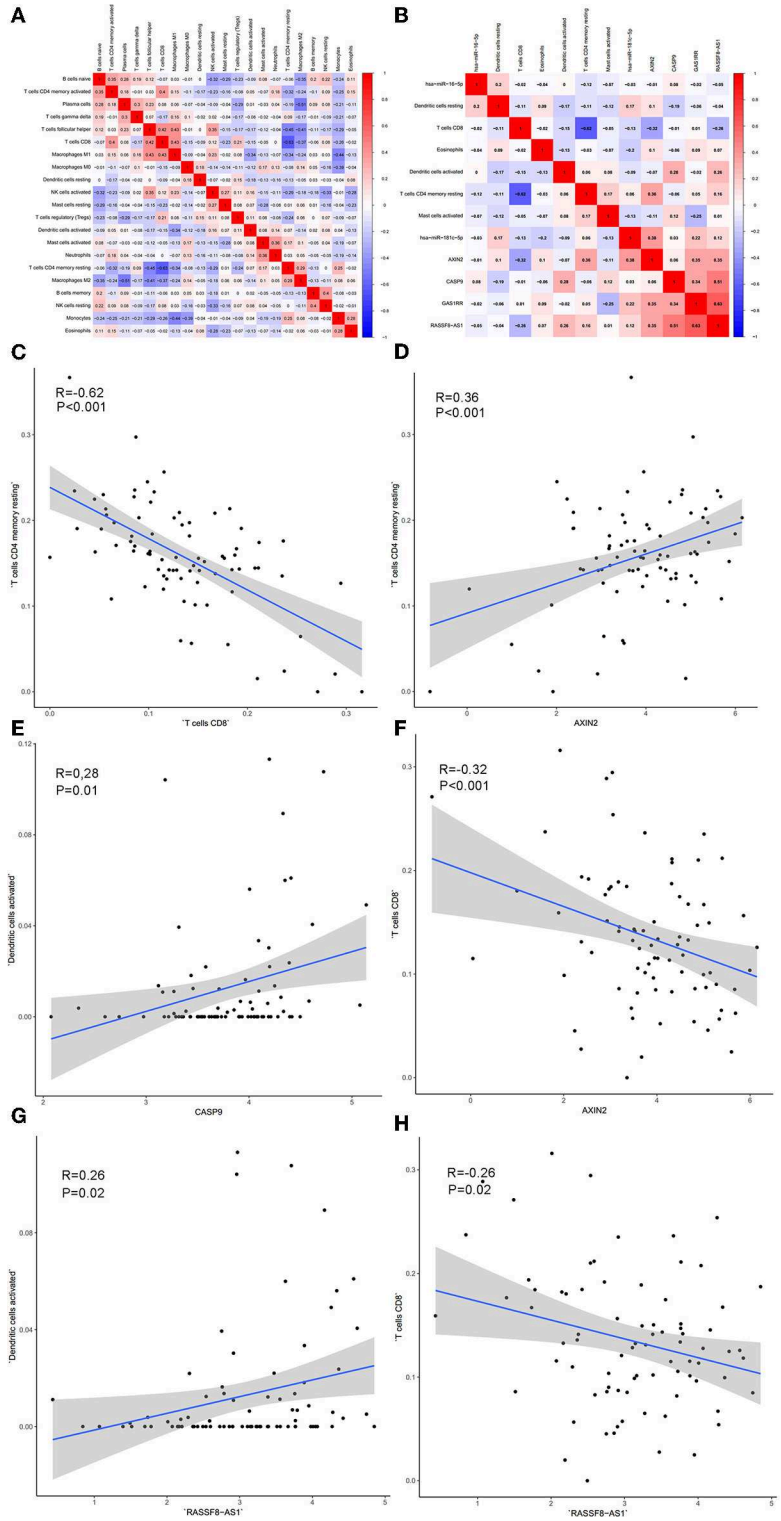


FIGURE 8 | The co-expression patterns among fractions of immune cells **(A)**; the co-expression patterns among fractions of immune cells and key members in the ceRNA network **(B)**; the relationships among immune cells and between ceRNAs and immune cells were calculated using Pearson correlation coefficients: T cells CD4 memory resting and T cells CD4 **(C)**; T cell CD4 memory resting and AXIN2 **(D)**; Dendritic cells activated and CASP9 **(E)**; T cells CD8 and AXIN2 **(F)**; Dendritic cells activated and RASSF8-AS1 **(G)**; T cells CD8 and RASSF8-AS1 **(H)**.

TABLE 2 | Summary of multidimensional external validation results based on multiple databases.

Database	AXIN2		CASP9		CGN		RASSF8-AS1		MIR4458HG		Results
	Cancer	Normal	Cancer	Normal	Cancer	Normal	Cancer	Normal	Cancer	Normal	
Oncomine	↑	↓	↑	↓	-	-	-	-	-	-	According to 17 analyses, AXIN was highly expressed in various tumors compared to normal tissue, while CASP9 was highly expressed only in medulloblastoma across two analyses (**Figure S1**).
cBioPortal	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	AXIN2, CASP9, CGN, RASSF8-AS1, and MIR4458HG were highly expressed in mesothelioma compared to some other types of cancer (**Figure S2**).
CCLE	-	-	-	-	-	-	-	-	-	-	AXIN2, CASP9, CGN, RASSF8-AS1, and MIR4458HG were expressed in various mesothelioma cell lines but not in immune cells biomarkers (**Figure S3**).
UALCAN	-	-	-	-	-	-	-	-	-	-	High expression of CGN indicated higher survival probability (**Figure S6**).
PROGgeneV2	-	-	-	-	-	-	-	-	-	-	High expression of AXIN2, CASP9, and CGN indicated higher overall survival (**Figure S4**).
STRING	-	-	-	-	-	-	-	-	-	-	AXIN2, CASP9, and CGN had significant Protein-Protein interaction network (**Figure S5**).

"↑" was defined as a significantly upregulated gene; "↓" was defined as a significantly downregulated gene; "-" was defined as a gene with no significant difference in expression. CCLE, Cancer Cell Line Encyclopedia.

IncRNAs, indirectly regulating mRNA, DNA, and protein expressions, was implemented through miRNA binding to silence the miRNA. The targeting relationship between hsa-mir-582-5p and RASSF8-AS1 was also verified by IP experiments in another study (Kameswaran et al., 2014). We inferred that the mechanism of hsa-miR-582-5p regulating CASP9 and dendritic cells resting might play a critical role in bone metastasis in the correlation analysis and RASSF8-AS1 might indirectly regulate CASP9 through interaction with hsa-miR-582-5p.

We also found T cells CD8, T cells CD4 memory resting and Axin2 were all associated with metastasis prognosis. The proportion of T cells CD8 and expression of Axin2 were found to be significantly associated with overall survival. The proportion of T cells CD4 memory resting was significantly different among four tumor stages of cancer. Moreover, co-expression shows that the fraction of T cells CD8 is negatively associated with the expression of AXIN2. The fraction of T cells CD4 memory resting is negatively associated with CD8 T cells and positively associated with AXIN2 expression. Thus, we presumed that the mechanism of AXIN2 regulating T cells CD8 and T cells CD4 memory resting might have an influence on tumor progression and metastasis.

Axin2 is a protein coding gene, identified as a regulator of the Wnt signaling pathway (Zeng et al., 1997). Axin2 can form a compound with adenomatous polyposis coli (APC) and glycogen synthase kinase-3β (GSK-3β) and then negatively regulate the Wnt/β-catenin signaling pathway, which has been considered as one of the primary pathways participates in cell proliferation, differentiation and migration (Ikeda et al., 1998; Sharpe et al., 2001). Axin2 inhibits the translocation of β-catenin into the nucleus to inhibit it from subsequently binding to transcription factors, and it also downregulates several target genes such as the vascular endothelial growth factor (VEGF), cyclin D1 and matrix metalloproteinases (MMP) (Cadigan and Nusse, 1997; Tortelote et al., 2017). In this way, axin2 is identified as a tumor suppressor (Behrens et al., 1998).

Previous study indicated that WNT activation downregulates the proportion of T cells memory by decreasing the critical transcription factor for the generation of T cells memory (Shen et al., 2013). That is consistent with our result that AXIN2 is positively associated with CD4 T cells memory resting. The WNT/β-catenin pathway also has an impact on thymocyte development, especially in the double-negative to double-positive transition and positive/negative selection progress (Pongracz et al., 2006; Yu et al., 2007; Kovalovsky et al., 2009; Shen et al., 2013). Yu et al. confirmed the expression of stabilized β-catenin accelerated the production of CD8 single-positive thymocytes using a transgenic mouse strain (Yu et al., 2007), which indicated that AXIN2 may be negatively associated with CD8 T cells, and our study indeed proved that. Thus, we inferred that the WNT/β-catenin pathway may be the potential

mechanism behind the association between AXIN2 and the fraction of immune cells.

In order to explore the gene and protein expressions of key biomarkers in the mesothelioma, normal tissue and cell lines we performed a dimensional validation (**Table 2**), applying multiple databases. First of all, across 17 analyses, AXIN2 (Median rank 178, $P < 0.05$) was highly expressed in various tumors compared to normal tissue, while CASP9 (Median rank 28, $P < 0.05$) showed higher expression only in medulloblastoma in the Oncomine (**Figure S1**). The integrative analysis of genomics and the clinical profiles using the cBioPortal suggested that AXIN2, CASP9, CGN, RASSF8-AS1, and MIR4458HG were highly expressed in mesothelioma compared to some other types of cancer (**Figure S2**). At the cellular level, the expression of ceRNA biomarkers was detected in various mesothelioma cell lines using the CCLE database (**Figure S3**). In clinical analysis, high expression of CGN showed higher survival probability in the UALCAN database (**Figure S6**), and high expression of AXIN2, CASP9, and CGN indicated higher overall survival in the PROGgeneV2 database (**Figure S4**). What's more, AXIN2, CASP9, and CGN had a significant Protein-Protein interaction network, according to the STRING database (**Figure S5**).

There are inevitably several limitations to our study that should be acknowledged. All data series in our study were from Western countries; therefore, one should be cautious in applying this conclusion to patients from Asian countries. Due to the limited amount of data available in public datasets, the clinicopathological parameters analyzed in this study are not comprehensive, which may cause analysis deviation. Thus, we explored multiple databases to observe gene and protein expression of key biomarkers at the cellular and tissue levels to minimize bias. The results showed that key biomarkers in our nomogram model were significantly changed in mesothelioma (**Figures S1–S6**). But, ignoring these limitations, our study first applied the nomograms model to predict prognoses of mesothelioma patients on the basis of ceRNA networks and tumor-infiltrating immune cells. We also inferred that AXIN2 probably regulates the cell differentiation of T cells CD8 and T cells CD4 memory resting through the WNT pathway, subsequently influencing cancer prognoses.

The tumor microenvironment often affects the invasive processes. The extracellular matrix molecules and secreted growth factors are involved in the transition of tumor cells into an invasive phenotype. It is noteworthy that the invasion and metastasis of tumor cells may have nothing to do with the proliferation of tumor cells but have occurred already at the early developmental stage of the tumor (Hosseini et al., 2016). Therefore, it is essential to identify molecules that may lead to mesothelioma invasion and metastasis. We established two nomograms to predict survival and bone metastasis of mesothelioma patients based on tumor-infiltrating immune cells and ceRNA networks, and we determined their utility by the high AUC values. The prediction nomograms proposed in this study might provide further comprehensive clinical information for furthering the personalized management of mesothelioma patients. Our study is a correlation study from multiple dimensions rather than a biological mechanism study. Based on the results of this study, we will conduct biological experiments to further validate our conclusion. A luciferase reporter assay will be using to prove the direct interaction mechanism of ceRNAs and Chromatin Immunoprecipitation (ChIP) to verify the molecular expression level of protein product. Next, we would like to demonstrate a molecular crosstalk between cancer cells and immune cells. We wish to consider if an exosome secreted by tumor cells contains ceRNAs, which act on immune cells and mediate the metastasis of Mesothelioma.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The relevant data provided by TCGA (https://tcga-data.nci.nih.gov/tcga/).

AUTHOR CONTRIBUTIONS

TM, ZH, and JZ conceived and designed the study. RH performed the bioinformatic analysis. GW, DS, and PY performed the data interpretation. JW and ZZ wrote the manuscript. HY and PH conducted the literature search. XZ, HW, and QL contributed to the revision of the manuscript draft. All authors read and approved the final manuscript.

FUNDING

This work was supported by grants from the Natural Science Foundation of China (Grant Nos. 81702659; 81772856; 81501203), the Youth Fund of Shanghai Municipal Health Planning Commission (No. 2017YQ054), and the Henan medical science and technology research project (Grant No. 201602031).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00257/full#supplementary-material

Figure S1 | Validation of AXIN2 **(A,B)** and CASP9 **(C,D)** on a transcriptional level in multiple cancer types and multiple studies using the Oncomine database.

Figure S2 | Integrative analysis of genomics using the cBioPortal suggested.

Figure S3 | Expression of ceRNA biomarkers in mesothelioma cell lines using CCLE database.

Figure S4 | Validation of expression levels of AXIN2 **(A)**, CASP9 **(B)**, and CGN **(C)** with overall survival of mesothelioma patients using P.

Figure S5 | Protein-Protein interaction network (PPI) of AXIN2 **(A)**, CASP9 **(B)**, and CGN **(C)** using the String database.

Figure S6 | Validation of CGN expression level on mesothelioma patient survival using the ULCAN database.

Table S1 | Baseline information of 87 patients diagnosed with mesothelioma, available from the TCGA. SD, Standard deviation.

REFERENCES

Behrens, J., Jerchow, B. A., Würtele, M., Grimm, J., Asbrand, C., Wirtz, R., et al. (1998). Functional interaction of an axin homolog, conductin, with beta-catenin, APC, and GSK3 beta. *Science* 280, 596–599. doi: 10.1126/science.280.5363.596

Bertoglio, P., Ambrogi, M. C., Chella, A., Aprile, V., Dini, P., Korasidis, S., et al. (2017). Is less also better? A single-institution experience on treatment of early stage malignant pleural mesothelioma. *EJSO* 43, 1365–1371. doi: 10.1016/j.ejso.2017.02.010

Brower, V. (2016). Tackling mesothelioma with immunotherapies. *JNCI* 108:djw198. doi: 10.1093/jnci/djw198

Cadigan, K. M., and Nusse, R. (1997). Wnt signaling: a common theme in animal development. *Genes Dev.* 11, 3286–3305. doi: 10.1101/gad.11.24.3286

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

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B. V. S. K., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

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, D296–302. doi: 10.1093/nar/gkx1067

Contributors, O. (2006). *Oncomine*. Available online at: https://openwetware.org/mediawiki/index.php?title=Oncomine&oldid=51984. (accessed July 19, 2006).

Dalsgaard, S. B., Würtz, E. T., Hansen, J., Røe, O. D., and Omland, Ø. (2019). Environmental asbestos exposure in childhood and risk of mesothelioma later in life: a long-term follow-up register-based cohort study. *Occup. Environ. Med.* 76, 407–413. doi: 10.1136/oemed-2018-105392

Floyd, D. H., Zhang, Y., Dey, B. K., Kefas, B., Breit, H., Marks, K., et al. (2014). Novel anti-apoptotic microRNAs 582–5p and 363 promote human glioblastoma stem cell survival via direct inhibition of caspase 3, caspase 9, and bim. *PLoS ONE* 9:e96239. doi: 10.1371/journal.pone.0096239

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:pl1. doi: 10.1126/scisignal.2004088

Goswami, C. P., and Nakshatri, H. (2014). PROGgeneV2: enhancements on the existing database. *BMC Cancer*. 14:970. doi: 10.1186/1471-2407-14-970

He, C., Wang, L., Zhang, J., and Xu, H. (2017). Hypoxia-inducible microRNA-224 promotes the cell growth, migration and invasion by directly targeting RASSF8 in gastric cancer. *Mol. Cancer* 16:35. doi: 10.1186/s12943-017-0603-1

Hosseini, H., Obradović, M. M. S., Hoffmann, M., Harper, K. L., Sosa, M. S., Werner-Klein, M., et al. (2016). Early dissemination seeds metastasis in breast cancer. *Nature* 540, 552–558. doi: 10.1038/nature20785

Huang, Q., Jiang, Z., Meng, T., Yin, H., Wang, J., Wan, W., et al. (2014). MiR-30a inhibits osteolysis by targeting RunX2 in giant cell tumor of bone. *Biochem. Biophys. Res. Commun.* 453, 160–165. doi: 10.1016/j.bbrc.2014.09.076

Ikeda, S., Kishida, S., Yamamoto, H., Murai, H., Koyama, S., Kikuchi, A. (1998). Axin, a negative regulator of the Wnt signaling pathway, forms a complex with GSK-3 beta and beta-catenin and promotes GSK-3 beta-dependent phosphorylation of beta-catenin. *EMBO J.* 17, 1371–1384. doi: 10.1093/emboj/17.5.1371

Infante, M., Morenghi, E., Bottoni, E., Zucali, P., Rahal, D., Morlacchi, A., et al. (2016). Comorbidity, postoperative morbidity and survival in patients undergoing radical surgery for malignant pleural mesothelioma. *Eur. J. Cardio Thorac. Surg.* 50, 1077–1082. doi: 10.1093/ejcts/ezw215

Kameswaran, V., Bramswig, N. C., McKenna, L. B., Penn, M., Schug, J., Hand, N. J., et al. (2014). Epigenetic regulation of the DLK1-MEG3 MicroRNA cluster in human type 2 diabetic islets. *Cell Metab.* 19, 135–145. doi: 10.1016/j.cmet.2013.11.016

Kim, S., Bull, D. A., Garland, L., Khalpey, Z., Stea, B., Yi, S., et al. (2019). Is there a role for cancer-directed surgery in early-stage sarcomatoid or biphasic mesothelioma? *Ann. Thorac. Surg.* 107, 194–201. doi: 10.1016/j.athoracsur.2018.07.081

King, J. A. C., Tucker, J. A., and Wong, S. W. (1997). Mesothelioma: a study of 22 cases. *South. Med. J.* 90, 199–205. doi: 10.1097/00007611-199702000-00006

Kovalovsky, D., Yu, Y., Dose, M., Emmanouilidou, A., Konstantinou, T., Germar, K., et al. (2009). Beta-catenin/Tcf determines the outcome of thymic selection in response to alpha beta TCR signaling. *J. Immunol.* 183, 3873–3884. doi: 10.4049/jimmunol.0901369

Li, R., Qu, H., Wang, S., Wei, J., Zhang, L., Ma, R., et al. (2018). GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. *Bioinformatics* 34, 2515–2517. doi: 10.1093/bioinformatics/bty124

Li, X., Zhao, G., Wang, Y., Zhang, J., Duan, Z., and Xin, S. (2013). RASSF7 and RASSF8 proteins are predictive factors for development and metastasis in malignant thyroid neoplasms. *J. Cancer Res. Ther.* 9, S173–S177. doi: 10.4103/0973-1482.122519

Lock, F. E., Underhill-Day, N., Dunwell, T., Matallanas, D., Cooper, W., Hesson, L., et al. (2010). The RASSF8 candidate tumor suppressor inhibits cell growth and regulates the Wnt and NF-kappa B signaling pathways. *Oncogene* 29, 4307–4316. doi: 10.1038/onc.2010.192

Luo, S., Lu, J. Y., Liu, L., Yin, Y., Chen, C., Han, X., et al. (2016). Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. *Cell Stem Cell* 18, 637–652. doi: 10.1016/j.stem.2016.01.024

Musk, A. B. W., de Klerk, N., and Brims, F. J. (2017). Mesothelioma in Australia: a review. *Med. J. Aust.* 207, 449–452. doi: 10.5694/mja17.00397

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

Pongracz, J. E., Parnell, S. M., Jones, T., Anderson, G., and Jenkinson, E. J. (2006). Overexpression of ICAT highlights a role for catenin-mediated canonical Wnt signalling in early T cell development. *Eur. J. Immunol.* 36, 2376–2383. doi: 10.1002/eji.200535721

Rodina, A., Wang, T., Yan, P., Gomes, E. D., Dunphy, M. P., Pillarsetty, N., et al. (2016). The epichaperome is an integrated chaperome network that facilitates tumour survival. *Nature* 538, 397. doi: 10.1038/nature19807

Salmena, L., Poliseno, L., Tay, Y., Kats, L., Pandolfi, P. P. (2011). A ceRNA hypothesis: the rosetta stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

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

Sharpe, C., Lawrence, N., and Martinez Arias, A. (2001). Wnt signalling: a theme with nuclear variations. *Bioessays* 23, 311–318. doi: 10.1002/bies.1045

Shen, S., Klamer, G., Xu, N., O'Brien, T. A., Dolnikov, A. (2013). GSK-3β inhibition preserves naive T cell phenotype in bone marrow reconstituted mice. *Exp. Hematol.* 41, 1016–1027. doi: 10.1016/j.exphem.2013.08.006

Shurin, M. R., Naiditch, H., Zhong, H., and Shurin, G. V. (2011). Regulatory dendritic cells New targets for cancer immunotherapy. *Cancer Biol. Ther.* 11, 988–992. doi: 10.4161/cbt.11.11.15543

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

Tortelote, G. G., Reis, R. R., de Almeida Mendes, F., Abreu, J. G. (2017). Complexity of the Wnt/β-catenin pathway: Searching for an activation model. *Cell. Signal.* 40, 30–43. doi: 10.1016/j.cellsig.2017.08.008

Trovo, M., Franceschini, D., Furlan, C., Pietrobon, F., Vagge, S., Farina, E., et al. (2019). Extra-pleural pneumonectomy in the era of image-guided intensity-modulated radiotherapy. *Radiol. Med.* 124, 854–859. doi: 10.1007/s11547-019-01030-y

van Gerwen, M., Alpert, N., Wolf, A., Ohri, N., Lewis, E., Rosenzweig, K. E., et al. (2019). Prognostic factors of survival in patients with malignant pleural mesothelioma; an analysis of the national cancer data base. *Carcinogenesis* 40, 529–536. doi: 10.1093/carcin/bgz004

Vogelzang, N. J., Rusthoven, J. J., Symanowski, J., Denham, C., Kaukel, E., Ruffie, P., et al. (2003). Phase III study of pemetrexed in combination with cisplatin versus cisplatin alone in patients with malignant pleural mesothelioma. *J. Clin. Oncol.* 21, 2636–2644. doi: 10.1200/JCO.2003.11.136

Yu, Q., Xu, M., and Sen, J. M. (2007). β-catenin expression enhances IL-7 receptor signaling in thymocytes during positive selection. *J. Immunol.* 179, 126–131. doi: 10.4049/jimmunol.179.1.126

Zeng, L., Fagotto, F., Zhang, T., Hsu, W., Vasicek, T. J., Perry, W. L., et al. (1997). The mouse fused locus encodes Axin, an inhibitor of the Wnt signaling pathway that regulates embryonic axis formation. *Cell* 90, 181–192. doi: 10.1016/S0092-8674(00)80324-4

Zhang, L., Wang, J. H., Liang, R. X., Huang, S. T., Xu, J., Yuan, L. J., et al. (2015). RASSF8 downregulation promotes lymphangiogenesis and metastasis in esophageal squamous cell carcinoma. *Oncotarget* 6, 34510–34524. doi: 10.18632/oncotarget.5923

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 © 2019 Huang, Wu, Zheng, Wang, Song, Yan, Yin, Hu, Zhu, Wang, Lv, Meng, Huang 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 26 November 2019
doi: 10.3389/fbioe.2019.00365

Prediction of circRNAs Based on the DNA Methylation-Mediated Feature Sponge Function in Breast Cancer

Yue Gu[1†], Ce Ci[1†], Xingda Zhang[2], Mu Su[3], Wenhua Lv[1], Chuangeng Chen[3], Hui Liu[1], Dongwei Zhang[4], Shumei Zhang[1,5*] and Yan Zhang[1,3,6*]*

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] Department of Breast Surgery, Harbin Medical University Cancer Hospital, Harbin, China, [3] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China, [4] Department of General Surgery, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [5] College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [6] State Key Laboratory of Respiratory Disease, Guangzhou Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
Xiaotian Zhang,
Van Andel Research Institute (VARI), United States

Reviewed by:
Nitish Kumar Mishra, University of Nebraska Medical Center, United States
Yanfeng Zhang, HudsonAlpha Institute for Biotechnology, United States

***Correspondence:**
Dongwei Zhang
mugong0@163.com
Shumei Zhang
zhangshumei@hrbmu.edu.cn
Yan Zhang
zhangtyo@hit.edu.cn

[†]*These authors share first authorship*

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *19 July 2019*
Accepted: *12 November 2019*
Published: *26 November 2019*

Citation:
Gu Y, Ci C, Zhang X, Su M, Lv W, Chen C, Liu H, Zhang D, Zhang S and Zhang Y (2019) Prediction of circRNAs Based on the DNA Methylation-Mediated Feature Sponge Function in Breast Cancer. Front. Bioeng. Biotechnol. 7:365. doi: 10.3389/fbioe.2019.00365

Several studies have found that DNA methylation is associated with transcriptional regulation and affect sponge regulation of non-coding RNAs in cancer. The integration of circRNA, miRNA, DNA methylation and gene expression data to identify sponge circRNAs is important for revealing the role of DNA methylation-mediated regulation of sponge circRNAs in cancer progression. We established a DNA methylation-mediated circRNA crosstalk network by integrating gene expression, DNA methylation and non-coding RNA data of breast cancer in TCGA. Four modules (26 candidate circRNAs) were mined. Next, 10 DNA methylation-mediated sponge circRNAs (sp_circRNAs) and five sponge driver genes (sp_driver genes) in breast cancer were identified in the CMD network using a computational process. Among the identified genes, *ERBB2* was associated with six sponge circRNAs, which illustrates its better sponge regulatory function. Survival analysis showed that DNA methylations of 10 sponge circRNA host genes are potential prognostic biomarkers in the TCGA dataset ($p = 0.0239$) and GSE78754 dataset ($p = 0.0377$). In addition, the DNA methylation of two sponge circRNA host genes showed a significant negative correlation with their driver gene expressions. We developed a strategy to predict sponge circRNAs by DNA methylation mediated with playing the role of regulating breast cancer sponge driver genes.

Keywords: circular RNA, DNA methylation, miRNA sponge, breast cancer, network

INTRODUCTION

Breast cancer is one of the most common malignant tumors in women (Koboldt et al., 2012). The HER2 (encoded by *ERBB2*) amplification group has achieved great clinical success because HER2 is an effective therapeutic target for breast cancer (Slamon et al., 1987). However, the current targeted therapy for breast cancer has not been fully and effectively implemented (Ayca and Traina, 2011). Studies have shown that dysregulated gene expression in breast cancer is affected by various genetic and epigenetic factors, including frequent somatic mutations (Nik-Zainal et al., 2016), copy number variations (Nik-Zainal et al., 2016), single nucleotide polymorphisms (SNPs) (Michailidou et al., 2017), non-coding RNA (Iorio et al., 2015), and DNA methylation (Yang et al., 2001). Abnormal regulation of epigenetic modifications leads to aberrant gene expression and function

(Berger et al., 2009; Pan et al., 2017). Epigenetic modifications thus play an important role in the development and progression of cancer (Esteller, 2008).

DNA methylation is a common epigenetic modification that regulates gene expression at the transcriptional level (Lindahl, 1981; Bird, 1984; Baylin and Jones, 2011; Wise and Charchar, 2016). Abnormal DNA methylation, including global hypomethylation and aberrant methylation status on key regulatory elements, has been observed in multiple human cancers (Trimarchi et al., 2011; Aran et al., 2013). Many studies have found significant differences in DNA methylation between breast cancer samples and normal samples. In general, cancer cells show hypomethylation of the entire genome, while promoter sequences of tumor suppressor genes show hypermethylation. Hypomethylation of the whole genome is caused by genomic instability and loss of gene imprinting, which causes overexpression of oncogenes (Trimarchi et al., 2011). In contrast, hypermethylation of tumor suppressor gene promoters causes genetic silencing of tumor suppressor genes (Bird, 1996; Aran et al., 2013). The hypomethylation of the gene promoter CpG island promotes its downstream gene transcription, whereas its hypermethylation inhibits the transcription of the downstream genes (Deaton and Bird, 2011).

Circular RNAs (circRNAs) form a covalently linked continuous loop and lack 5′ and 3′ ends and polyadenylation tail structures (Diener, 1989; Chen and Yang, 2015; Song et al., 2016). CircRNAs exhibit important effects on RNA-related protein binding, gene splicing regulation and transcription, as well as the modification of parental gene expression (Li et al., 2015; Peng et al., 2017). Recent studies have shown that circRNAs are more stable than their linear counterparts and some function as biological markers for disease treatment (Weil et al., 2015; Zhang et al., 2015; Abu and Jamal, 2016; Kulcheski et al., 2016). Many studies have shown that circRNAs can also function as miRNA sponges in the regulation of gene expression (Hansen et al., 2013; Jens, 2014; Qiang et al., 2016). It can be involved in the development of a variety of diseases, such as atherosclerotic disease, prion disease, and myotonic dystrophy (Qu et al., 2015). For example, the circular transcript CDR1as (or ciRS-7), which is encoded by the reverse genome of the human CDR1 gene in human and mouse brain tissues, functions as an miRNA sponge to miR-7(Hansen et al., 2011). ciRS-7 inhibits miR-7 activity and competes with miR-7 in binding to other RNAs to modulate target gene expression (Hansen et al., 2013; Peng et al., 2015). Several studies have demonstrated that circRNAs also exhibit functions in cancer. For instance, circ-Amotl1 is highly expressed in breast cancer samples and many cancer cell lines (Yang et al., 2017). Interactions between circ-Amotl1 and c-myc lead to increased tumorigenicity (Yang et al., 2017). The potential clinical value of circRNAs in cancer diagnosis, therapy and prognosis has been the subject of recent investigation (Abu and Jamal, 2016; Kulcheski et al., 2016; Yang et al., 2017).

circRNAs are weakly expressed and have a long half-life (Enuka et al., 2016). The detection process of circRNA expression is technically challenging (Enuka et al., 2016). Here we integrated genome, epigenome and non-coding RNA data of breast cancer samples and used a bioinformatics approach to identify potential important sponge circRNAs in breast cancer. This study demonstrates the important role of circRNA sponge regulation associated with DNA methylation in breast cancer, which suggests a therapeutic strategy for manipulating the driver gene function in breast cancer through circRNA sponge regulation.

MATERIALS AND METHODS

Data Sources

miRNASeq data (1,102 breast cancer samples, 104 normal samples), Illumina Infinium Human Methylation 450 BeadChip level 3 data (684 cancer samples, 96 normal samples), and RNASeqV2 data (1,102 breast cancer samples, 113 normal samples) were downloaded from The Cancer Genome Atlas (TCGA) (https://gdc.cancer.gov/). Clinical data were downloaded from TCGA (https://gdc.cancer.gov/) and GSE78754. Among them, DNA methylation and gene expression matching data were found in 781 cancer samples and 84 normal samples. CircRNA-miRNA interaction data were downloaded from the Starbase v2.0 (http://starbase.sysu.edu.cn/). MiRNA-target gene interaction data were downloaded from miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/), which stores experimentally verified miRNA-target gene interaction information. CircRNA location information was obtained from Circbase (http://www.circbase.org/). The protein coding driver gene list was derived from Vogelstein et al. (Vogelstein et al., 2013). The reference genome (hg19) was obtained from UCSC (http://genome.ucsc.edu/).

Data Preprocessing

We integrated miRNASeq data from 1,102 breast cancer samples and 104 normal samples to construct miRNA expression profiles between breast cancer samples and normal samples. If there was a missing value for miRNA expression, the miRNA was removed.

We used Illumina Infinium Human Methylation 450 BeadChip data from 684 breast cancer samples and 96 normal samples and circRNA position information to construct the DNA methylation profile of circRNA host genes. We performed data preprocessing to remove probes containing missing values in the sample and probes for multiple genes, CpG sites on the sex chromosomes and SNPs. Previous studies showed that multiple adjacent CpG sites have the same DNA methylation pattern (Lehmann-Werman et al., 2016); if a circRNA host gene contains multiple CG loci, the DNA methylation level of the circRNA host gene is quantified as the average methylation level of these CG loci.

We used the same samples to construct the DNA methylation profile of target genes interacting with differentially expressed miRNAs. We performed data preprocessing to remove probes containing missing values in the sample and probes for multiple genes, CpG sites on the sex chromosomes and SNPs. The DNA methylation of target gene was quantified as the average methylation level of CG loci located in the target gene.

We use RNASeqV2 data of breast cancer samples and normal samples to construct target gene expression profile interacting

with differentially expressed miRNAs. We removed target genes with missing values.

Establishment of the DNA Methylation-Mediated circRNA Crosstalk (DMCC) Network and Identification of Candidate circRNAs

First, differentially expressed miRNAs (fold change $>$ 2 or fold change $<$ 0.5, q value $<$ 1) were screened between breast cancer samples and normal samples using SAMR package. CircRNAs that interacted with differentially expressed miRNAs were matched. If two circRNAs shared the same miRNA (two circRNAs have sequence and function similarity), then both circRNAs form a connection. A circRNA crosstalk (CC) network was constructed.

In addition, the DNA methylation profile of circRNA host genes in the CC network was constructed according to the data preprocessing description. Differential methylation circRNA host genes between breast cancer samples and normal samples were screened by SAMR package (fold change $>$ 1.5 or fold change $<$ 2/3, q value $<$ 5), and Pearson correlation of DNA methylation of two circRNA host genes was calculated. If the Pearson correlation of two circRNA host genes was greater than random (permutation $>$ 1,000) even $>$0.6, these two circRNA host genes will be two nodes of one edge in the CC network. The DMCC network was constructed with Pearson correlation as the weight. The circRNA host genes are co-methylated in the DMCC network.

Module-mining was performed in the DMCC network by using MCODE plugin in the Cytoscape, and circRNAs in the modules were selected as the candidate circRNAs.

miRNA Target Prediction and circRNA-miRNA-Driver (CMD) Gene Network Establishment

All human miRNAs and target information were derived from the miRTarBase database, which stores the experimental validated miRNA targets and is prevalent in the target prediction. We used SAMR package to screen differentially expressed genes between breast cancer samples and normal samples (fold change $>$ 2 or fold change $<$ 0.5, q value $<$ 1). We obtained breast cancer driver genes by taking the intersection of breast cancer differentially expressed genes and cancer driver genes. We used candidate circRNAs, breast cancer driver genes and all shared miRNAs to construct the CMD network. Each pair of circRNA-driver gene in this network is a candidate circRNA-driver gene pair. All implementations of network diagrams and module mining are realized by Cytoscape.

Predicting the Sponge circRNAs of Protein-Coding Driver Genes

Studies have shown that circRNAs are almost co-expressed with their linear transcripts (Enuka et al., 2016). Therefore, DNA methylation of the circRNA host genes may regulate the expression of circRNAs, which influences the competitive regulation of circRNAs as miRNA sponge. Similarly, DNA methylation of breast cancer driver genes can also regulate the expression of driver genes. Sponge circRNAs positively regulate the expression of their targets, and their regulation depends on the miRNA stoichiometry (Du et al., 2016). Therefore, we used the following hypothesis to predict sponge circRNA-driver gene pairs: sponge circRNA host genes and target driver genes are co-methylated and they interact with the same miRNAs.

A complete computational analysis process was designed to predict sponge circRNA-driver gene pairs. First, for each candidate circRNA-driver gene pair, the significance (P_1: P-value of Fisher's exact test) of shared miRNA with same seeds and the significance (P_2: P-value of Pearson's correlation coefficient test) of the DNA methylation correlation between a circRNA host gene and driver gene promoter region (-1.5 to $+0.5$ kb relative to transcription start site) in breast cancer samples were calculated. We computed P_1 using formula (1):

$$P_1 = \frac{C_{a+b}^{a} C_{c+d}^{c}}{C_n^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \quad (1)$$

where a is the number of shared miRNAs in a candidate circRNA-driver gene pair; b is the number of shared miRNAs in a circRNA-other driver genes; c is the number of shared miRNAs in a driver gene-other circRNAs; d is the number of remaining miRNAs; and n represents the number of total miRNAs.

We computed P_2 using formula (2):

$$P_2 = P(r) \quad (2)$$

with

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{Y_i - \overline{Y}}{s_Y}\right) \quad (3)$$

where X is the DNA methylation profile of circRNA host genes in breast samples, and Y is the DNA methylation profile of driver genes in breast cancer samples.

The following conditions were fulfilled for the candidate circRNA-driver gene pairs as predicted sponge circRNA-driver gene pairs: (1) P_1 and P_2 are no larger than a threshold of 0.05 ($P_1 \leq$ 0.05 and $P_2 \leq$ 0.05); (2) the Pearson correlation of DNA methylation between circRNA host genes and driver gene promoter region shows a positive correlation and greater than random (permutation 1000 times, $p < 0.05$); (3) circRNA-driver gene pairs that shared at least eight unique targeting sites for these shared miRNAs (different miRNAs sharing the same seed sequence could target the same site) are selected (Du et al., 2016); and (4) there was a significant negative correlation between DNA methylation of driver gene promoter region and the expression of driver genes ($p < 0.05$).

Functional Enrichment Analysis of circRNA Host Genes Associated With DNA Methylation

The circRNA host genes in the DMCC network are associated with DNA methylation and are co-methylated. We next analyzed the function of circRNA host genes in the DMCC network.

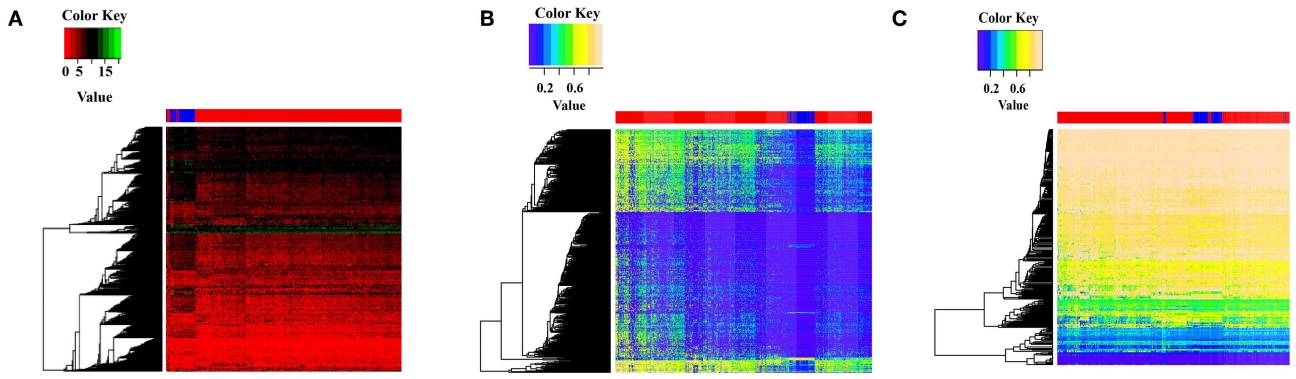


FIGURE 1 | Feature display of target genes and circRNAs that interact with differentially expressed miRNAs. **(A)** The target genes expression of differential expression miRNAs between breast cancer samples and normal samples. Gene expression values were log2 transformed after adding a pseudo-value of 1 to avoid infinite values. Red indicates breast cancer samples, blue indicates normal samples. **(B)** DNA methylation cluster analysis of target genes bound by differentially expressed miRNAs. Red represents breast cancer samples, blue represents normal samples. **(C)** DNA methylation cluster analysis of circRNA host genes that interact with differentially expressed miRNAs. Red indicates breast cancer samples, and blue indicates normal samples.

Functional and pathway enrichment of the DNA methylation-mediated circRNA host genes was achieved using DAVID (https://david.ncifcrf.gov/). Rich factor is the ratio of the number of circRNA host genes mapped to this GO term and the number of annotated genes in this term; the higher rich factor means the more significant enrichment. *P*-value was corrected by Benjamini. Enrichment analysis. The graph was performed using the ggplot2 package in R 3.2.1.

Survival Analysis

The clinical survival data for breast cancer were downloaded from TCGA and GEO (GSE78754). Survival data of breast cancer patients were integrated (missing data were removed). Finally, we obtained 656 breast cancer samples in TCGA and 70 breast cancer samples in GSE78754. We assessed the predictive effect of sponge circRNAs on overall survival in breast cancer by survival analysis. Survival analysis was performed using Cox proportional hazards model. DNA methylation of the sponge circRNA host gene was used as a covariate to assess the independent contribution of each circRNA host gene to prognosis. The DNA methylation value of the circRNA host gene was multiplied by the linear sum of the regression coefficients in the multivariate Cox regression to assign a prognostic index (PI). PI for each patient was calculated as shown in formula 4.

$$\mathrm{PI} = \beta_1 x_1 + \beta_1 x_1 + \ldots + \beta_n x_n \quad (4)$$

where x_i represents DNA methylation of circRNA host gene i; β_i represents regression coefficient of gene i in multivariate Cox regression; and n represents the number of genes. The median of PI was used as a threshold to classify breast cancer patients into the high risk group and low risk group. We used R V3.2.1 for the survival analysis and generating survival plot.

RESULTS

Differential Target Gene Expression and DNA Methylation of circRNA Host Genes and Target Genes in Breast Cancer

miRNAs are abundantly present in many organisms and can selectively interact with complementary mRNAs to reduce protein production. We identified 192 differentially expressed miRNAs between breast cancer samples and normal samples in TCGA datasets. To determine potential targets of these miRNAs, we next used miRTarBase, which revealed 4,546 potential target genes of the differentially expressed miRNAs. Notably, the target genes also showed differential expression between breast cancer and normal samples and target gene expression can separate breast cancer samples from normal samples (**Figure 1A**). This suggests that these differentially expressed miRNAs in breast cancer may be targeting specific mRNAs to regulate target gene expression, leading to the differential gene expression in breast cancer samples.

We next constructed a DNA methylation profile of the 4,546 target genes between 684 breast cancer samples and 96 normal samples. The DNA methylation heat map showed that the target genes showed differential methylation between breast cancer and normal samples, with higher DNA methylation in breast cancer compared with normal samples. Furthermore, DNA methylation of the target genes could separate breast cancer samples and normal samples (**Figure 1B**). These findings suggest that the differential levels of DNA methylation may impact target gene expression in breast cancer.

circRNA can bind to miRNAs through miRNA response elements (MREs) and regulate miRNA functions. We found that 747 circRNAs showed potential interaction with the 192 differentially expressed miRNAs. DNA methylation of circRNA host genes may affect their miRNA sponge function and impact miRNA target gene expression. To explore the potential role of DNA methylation of circRNA host genes, we constructed

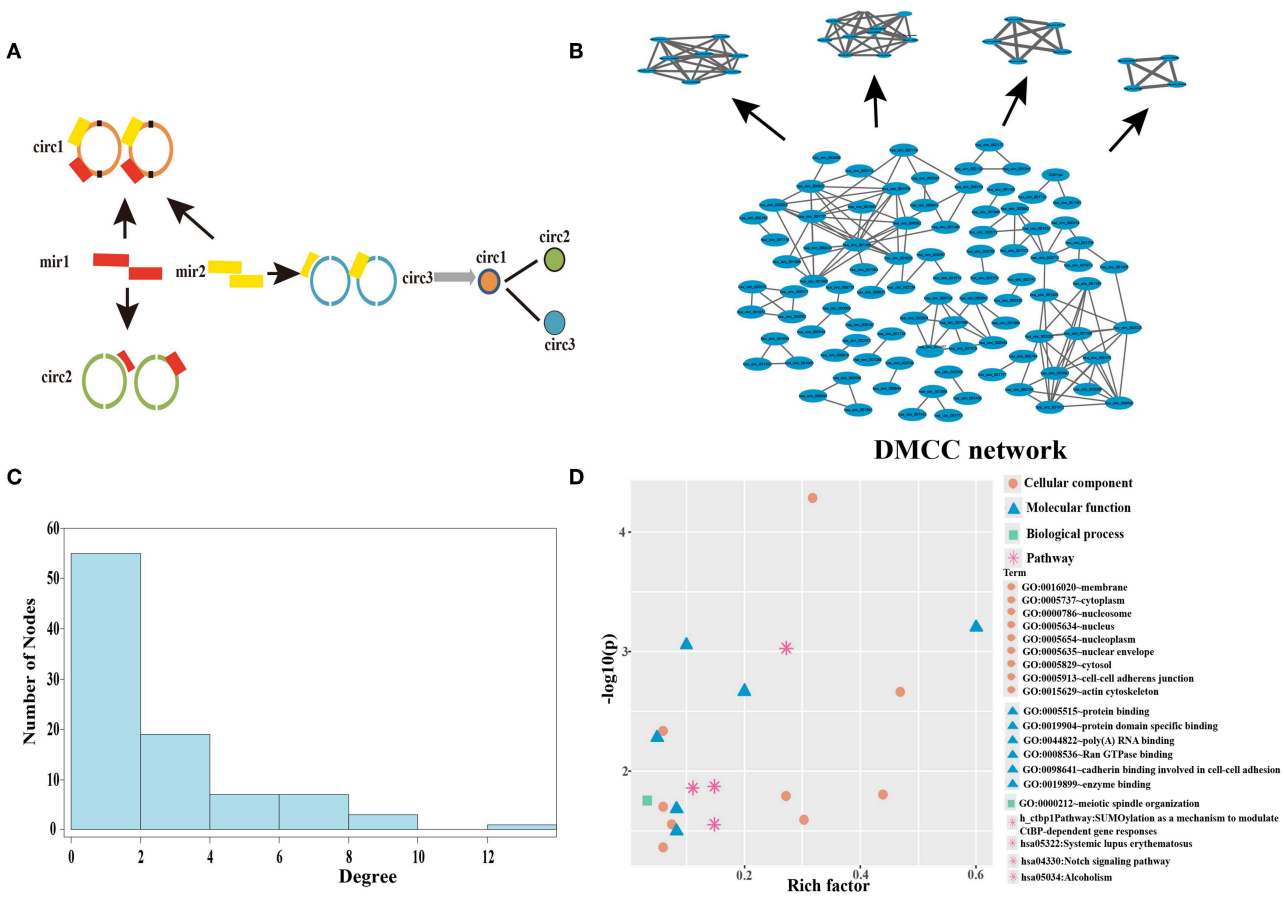


FIGURE 2 | Establishment and analysis of DMCC network in the breast cancer. **(A)** The construction of the circRNA crosstalk (CC) network. The orange, green, and blue represent three different circRNAs, and the red and yellow rectangles are two miRNAs. **(B)** The DNA methylation mediated circRNA crosstalk (DMCC) network and candidate circRNAs. The blue nodes represent circRNAs. **(C)** Degree distribution of DMCC network. **(D)** The function and pathway enrichment analysis of circRNA host genes in the DMCC network. The abscissa represents the rich factor, and the ordinate represents the inverse of the logarithm of the *P*-value corrected by Benjamini. The larger value represents the significant of GO terms and pathway.

a DNA methylation profile of the 747 circRNA host genes and generated a DNA methylation heat map. The heat map showed that circRNA host genes showed differential methylation between breast cancer samples and normal samples, and the DNA methylation of circRNA host genes could separate breast cancer samples from normal samples (**Figure 1C**). These results confirm differential DNA methylation of circRNA host genes in breast cancer and suggests that methylation of circRNA host genes may indirectly regulate target gene expression.

Construction of the DMCC Network and Identification of Candidate circRNAs

We next aimed to predict sponge circRNAs from the perspective of epigenetics regulation. To first identify circRNAs that highly correlated with DNA methylation, we established a DMCC network. We screened the 192 miRNAs differentially expressed in breast cancer and found that they formed 2,419 circRNA-miRNA interactions. A circRNA crosstalk (CC) network (**Supplementary Figure 1**) was established by sharing same miRNAs (**Figure 2A**), thus two circRNAs have sequence similarity. The CC network contained 747 circRNAs and 48,492 circRNA-circRNA connections. The circRNA-circRNA network represents the interactions between circRNAs, in which two circRNAs may have sequence functional similarities. The maximum degree of the CC network is 938. The degree distribution of the CC network is subject to the power-law distribution (**Supplementary Figure 2**), and therefore the CC network is a biologically significant network.

We next constructed a DNA methylation profile of the 747 circRNA host genes according to the circRNA location information and identified 201 differentially methylated circRNA host genes ($P < 0.05$). DNA methylation of these circRNA host genes could separate cancer samples from normal samples (**Supplementary Figure 3**). We calculated the Pearson correlation coefficient of DNA methylation between every two circRNA host genes. By connecting the two circRNAs with a Pearson correlation coefficient greater than random even >0.6 ($P < 0.05$), a DMCC network was

constructed (**Figure 2B**). The circRNA host genes in the DMCC network are characterized by co-methylation. The DMCC network contains 92 circRNAs and 138 circRNA-circRNA interactions. The degree distribution of the DMCC network is also subject to the power-law distribution (**Figure 2C**), and therefore the DMCC network is also biologically significant.

Our results above show that the 92 circRNAs in the DMCC network are differentially methylated in breast cancer samples, with sequence similarity and co-methylation characteristics. We next evaluated the biological function of these circRNA host genes by DAVID and found that the circRNA host genes are mainly enriched in the cell membrane, cytoplasm, protein binding, and protein domain specific binding (**Figure 2D**). These genes are mainly enriched in SUMOylation as a mechanism to modulate CtBP-dependent gene responses, systemic lupus erythematosus, and the Notch signaling pathway (**Figure 2D**).

To identify novel circRNAs associated with DNA methylation, module mining in the DMCC network was performed. Four network modules (**Figure 2B**) were obtained in the DMCC network according to the module score. Four circRNAs (hsa_circ_000582, hsa_circ_002025, hsa_circ_002024, hsa_circ_001851) are seed nodes of four modules. A total of 26 circRNAs in the four modules were used as candidate circRNAs for subsequent studies.

Establishment and Global Properties of the CMD Network

miRNAs interact with circRNAs as well as cancer driver genes through MREs (Bartel, 2009; Lü et al., 2017). Thus, some circRNAs may act as sponges by binding to miRNAs and preventing their interactions with cancer driver genes. These circRNAs therefore regulate the expression of driver genes and play a very important role in disease regulation.

To examine the regulatory interactions among potential circRNAs, their miRNA targets and putative cancer driver genes, we established a CMD gene network. We first screened 4,541 differentially expressed genes between breast cancer and normal samples (fold change > 2 or fold change < 0.5, $p < 0.01$). Differentially expressed genes and 125 cancer driver genes were selected and 25 breast cancer driver genes were obtained. According to rules (**Figure 3A**), we constructed a CMD network (**Figure 3B**), containing 23 circRNAs, 23 cancer driver genes and 45 miRNAs. The CMD network is divided into three layers: driver genes act as functional layer in the outermost layer, and the internal circRNAs indirectly regulate driver genes through miRNAs in the middle layer.

The miRNA-mediated complex CC network was established by the circRNA-miRNA-driver gene interaction. The crosstalk type between the circRNA-miRNA-driver genes represents a complex transcriptional regulatory network. This provides a perspective on how to indicate the intermolecular relationship of cell behavior. Each circRNA-driver gene pair in the circRNA-miRNA-driver gene network served as candidate circRNA-driver gene pair for subsequent study.

Prediction of Sponge circRNAs Regulating Cancer Driver Genes Based on the CMD Network

We next used a bioinformatics approach to predict sponge circRNA-driver gene pairs in breast cancer affected by DNA methylation.

Sponge circRNAs have a positive regulatory effect on their target gene expression (Hansen et al., 2011; Peng et al., 2015; Zheng et al., 2016), and the strength of their regulation depends on the stoichiometry of the involved miRNAs (Du et al., 2016). If the host gene of the sponge circRNA is hypermethylated, the circRNA host gene will be silenced, eliminating its ability to bind miRNAs, leading to downregulation of the target gene. This suggests that high methylation of host genes may correlate with downregulated target gene expression and vice versa. Therefore, the sponge circRNA may have the same methylation pattern as the sponge driver gene (**Figure 3C**).

Based on this assumption, we developed a computational process to predict the sponge circRNA and gene pairs in breast cancer. CircRNA-breast cancer driver gene pairs that share at least one miRNA were selected as candidate circRNA-driver gene pairs. Using the prediction algorithm in the flow (**Figure 3D**), we identified 10 sponge circRNA-driver gene pairs comprising 10 sponge circRNAs and five protein-coding driver genes (**Figures 4A,B**).

We found that most of the sponge circRNAs and driver genes are located on the different chromosomes (even when located on the same chromosome, the distance was >2 Mb). This illustrates that the regulation between sponge circRNAs and sponge driver genes is a trans-acting relationship, supporting the mechanism by which sponge circRNAs indirectly regulate sponge driver gene expression by interacting with miRNAs.

Characteristics of Sponge circRNAs Based on DNA Methylation

Many studies have shown that circRNAs can act as sponge circRNAs to competitively bind to miRNAs and regulate gene expression (Leonardo Salmena et al., 2011; Hansen et al., 2013; Zheng et al., 2016; Peng et al., 2017). We next analyzed the predicted sponge circRNA-driver gene pairs that showed the following characteristics: (1) DNA methylation of the sponge circRNA host gene was positively correlated with DNA methylation of the sponge driver gene promoter region in breast cancer samples (**Figures 5A,B**) and (2) DNA methylation of the driver gene promoter region was negatively correlated with driver gene expression in breast cancer samples (**Figures 5C,D**). We found that the *ERBB2* gene was associated with a large number of predicted sponge circRNAs compared with other examined genes (**Supplementary Table 1**), which suggests that it may have the better sponge regulation. We also examined DNA methylation of sponge circRNA host genes, DNA methylation of sponge driver genes and sponge driver gene expression patterns in breast cancer samples and normal samples (**Supplementary Figure 5**). The box plot showed that the DNA methylation of the sponge circRNA host genes and the DNA methylation of the sponge driver genes have the

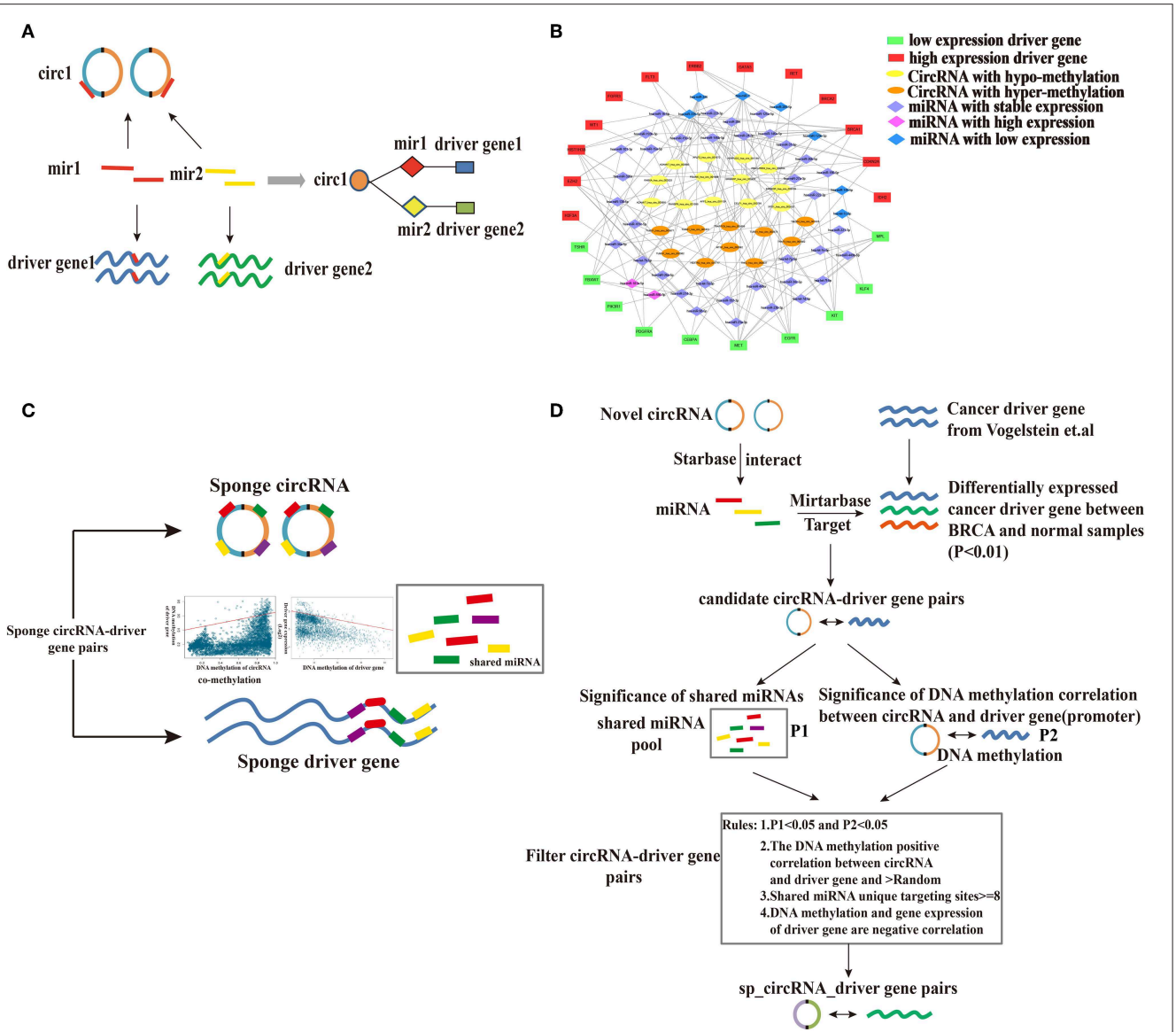


FIGURE 3 | Establishment of the CMD gene network and prediction process of sponge circRNA-driver gene pairs. **(A)** The construction of the CMD gene network. **(B)** The CMD network. The ellipse represents circRNAs, the diamond represents miRNAs, and the rectangle represents driver genes. **(C)** Construction of sponge regulatory mechanism for circRNA and driver gene by targeting the same miRNA and prediction of sponge circRNAs. **(D)** Computational strategy of predicting sponge circRNA-protein coding driver genes in breast cancer.

same tendency, while the DNA methylation of the sponge driver gene and sponge driver gene expression showed the opposite trend.

We further analyzed pathways of the predicted five cancer driver genes and found that *KIT*, *ERBB2* and *CEBPA* are related to PI3K and RAS pathways (Vogelstein et al., 2013). *GATA3* is associated with transcriptional regulatory core pathways (Vogelstein et al., 2013). *WT1* is associated with the chromosome modification core pathway (Vogelstein et al., 2013). These genes are enriched in the process of cell survival and cell fate (**Supplementary Table 2**) (Vogelstein et al., 2013). Together these results suggest that the driver genes regulated by sponge circRNAs may be associated with cancer. These results are consistent with previous studies that showed that sponge RNAs may have the same carcinogenic or tumor suppressor function as their regulatory genes (Du et al., 2016).

Sponge circRNAs Are Potential Prognostic Biomarkers for Breast Cancer

To determine whether the sponge circRNAs are associated with the prognosis of breast cancer, we performed survival analysis using Infinium Human Methylation 450 BeadChip

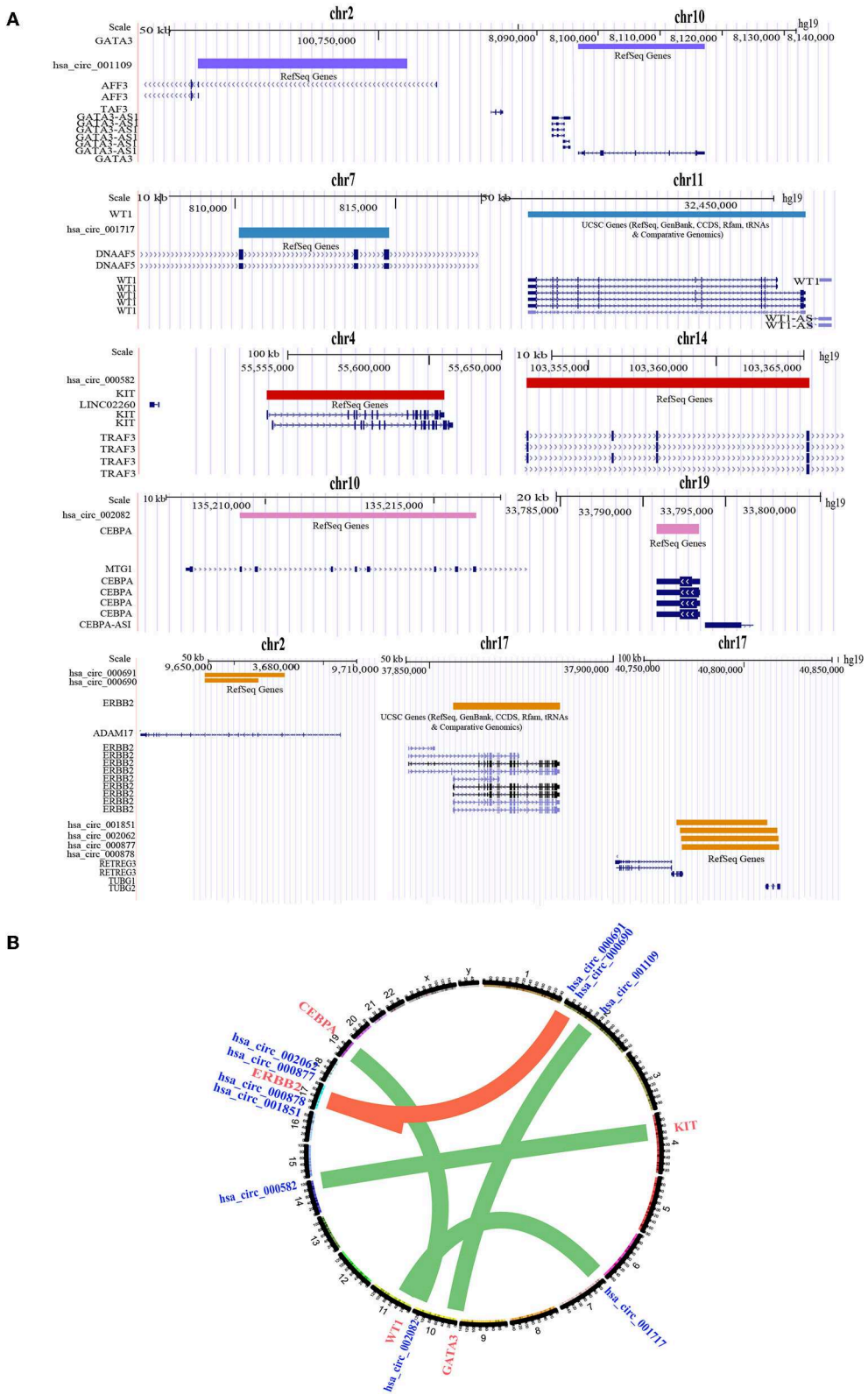


FIGURE 4 | Location information of sponge circRNA-driver gene pairs. **(A)** Genome position information of 10 sponge circRNA-driver gene pairs. Purple bar represents the genome location of *GATA3* and its sponge circRNA. Blue bar represents the genome location of *WT1* and its sponge circRNA. Red bar represents the genome location of *KIT* and its sponge circRNA. Pink bar represents the genome location of *CEBPA* and its sponge circRNA. Yellow bar represents the genome location of *ERBB2* and its sponge circRNAs. **(B)** Sponge circRNA-driver gene pairs in breast cancer. The nodes represent circRNAs and driver genes (blue represents circRNAs, red represents driver genes) and the edges represent the predicted regulation between sponge circRNA and the corresponding protein-coding driver gene.

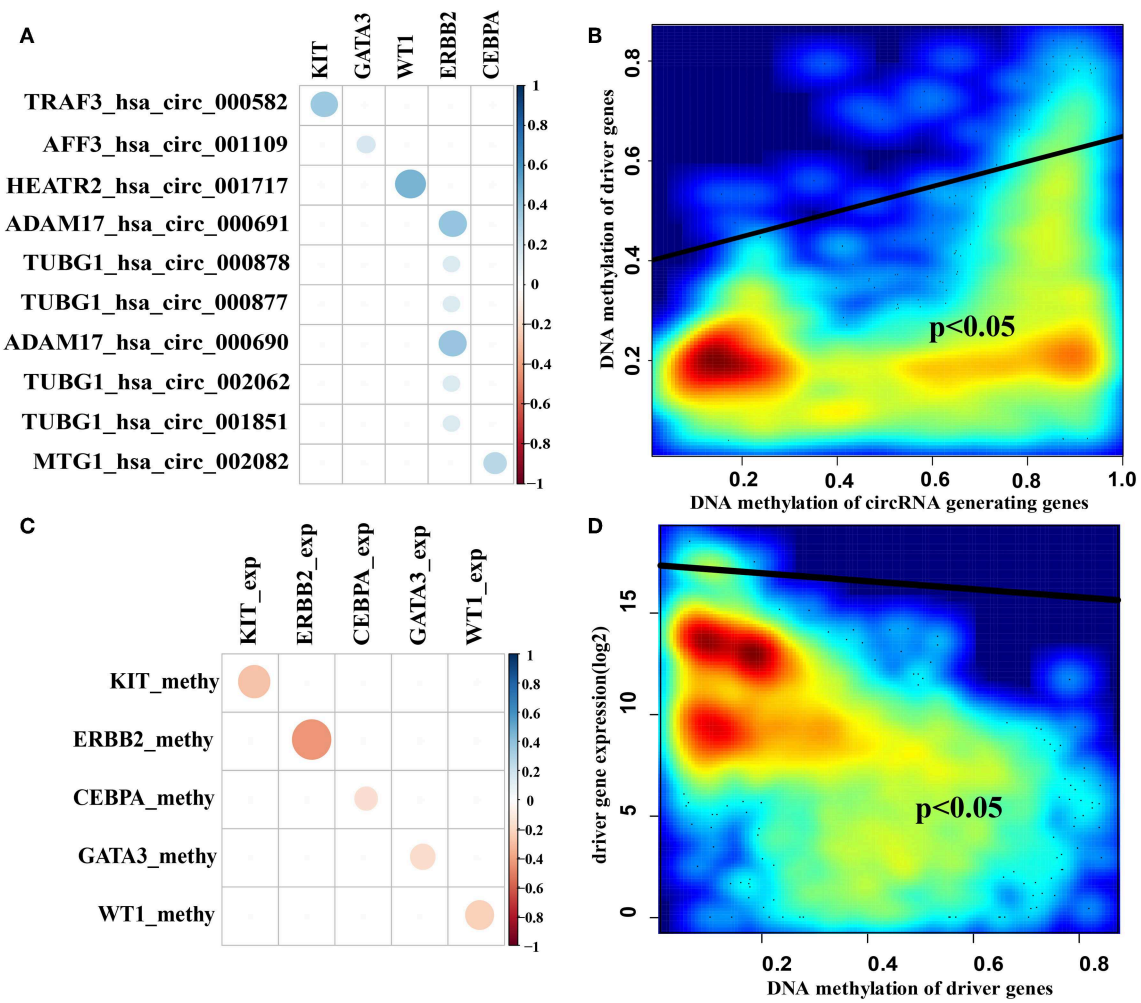


FIGURE 5 | Correlation of sponge circRNAs and target driver genes. **(A)** The correlation coefficient diagram of DNA methylation of sponge circRNA host genes and target driver genes in breast cancer samples. **(B)** The smoothscatter plot of DNA methylation of sponge circRNA host genes and DNA methylation of target driver genes in breast cancer samples. **(C)** The correlation coefficient diagram of DNA methylation of target driver genes and sponge driver genes expression in breast cancer samples. **(D)** The smoothscatter plot of DNA methylation of sponge driver genes and sponge driver genes expression in breast cancer samples. Gene expression values were log2 transformed after adding a pseudo-value of 1 to avoid infinite values. The scalable size of circles in correlation coefficient diagram represent the absolute value of the correlation coefficient.

data and clinical survival data of breast cancer in TCGA and GSE78754.

Survival analysis shows that patients in the high risk group have worse overall survival than those in the low risk group in the TCGA dataset (**Figure 6A**, $p = 0.0239$). DNA methylation of the circRNA host genes clearly separated the high and low risk groups in breast cancer patients in the first 10 years. Similarly, patients in the high risk group showed worse survival compared with those in the low risk group in the GSE78764 dataset (**Figure 6B**, $p = 0.0377$), and DNA methylation of circRNA host genes separated the high and low risk groups of breast cancer patients. These results indicate that the 10 sponge circRNAs can not only be used as potential diagnostic markers for breast cancer, but also may reflect the prognosis of patients.

DNA Methylation of circRNA Host Gene and Driver Gene Synergistically Affect Driver Gene Expression

We also calculated the correlation between DNA methylation of circRNA host genes and driver genes with driver gene expression. The results indicated that DNA methylation of sponge circRNA host gene and DNA methylation of driver gene synergistically affect driver gene expression in breast cancer (**Supplementary Figure 4**). Moreover, DNA methylation of the predicted two sponge circRNAs (hsa_circ_000582, hsa_circ_001109) showed a significant negative correlation with the expression of their sponge driver genes (*KIT*, *GATA3*) (**Figures 7A,B**). The relationship between the expression of *ERBB2*, *CEBPA*, and *WT1* and DNA methylation of their sponge circRNAs is shown in **Supplementary Figure 6**.

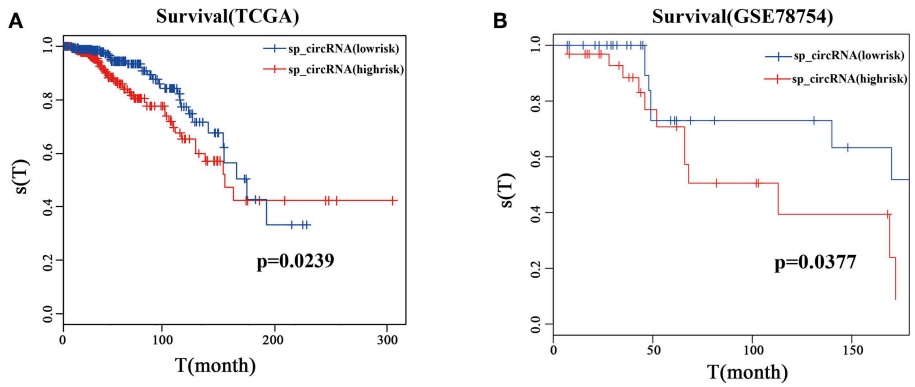


FIGURE 6 | Survival analysis of 10 sponge circRNAs. **(A)** Survival analysis curve of 10 sponge circRNAs using TCGA clinical data (breast cancer). Red curve represents high risk patients, and blue curve represents low risk patients. **(B)** Survival analysis curve of 10 sponge circRNAs using GSE78754. Red curve represents high risk patients, and blue curve represents low risk patients.

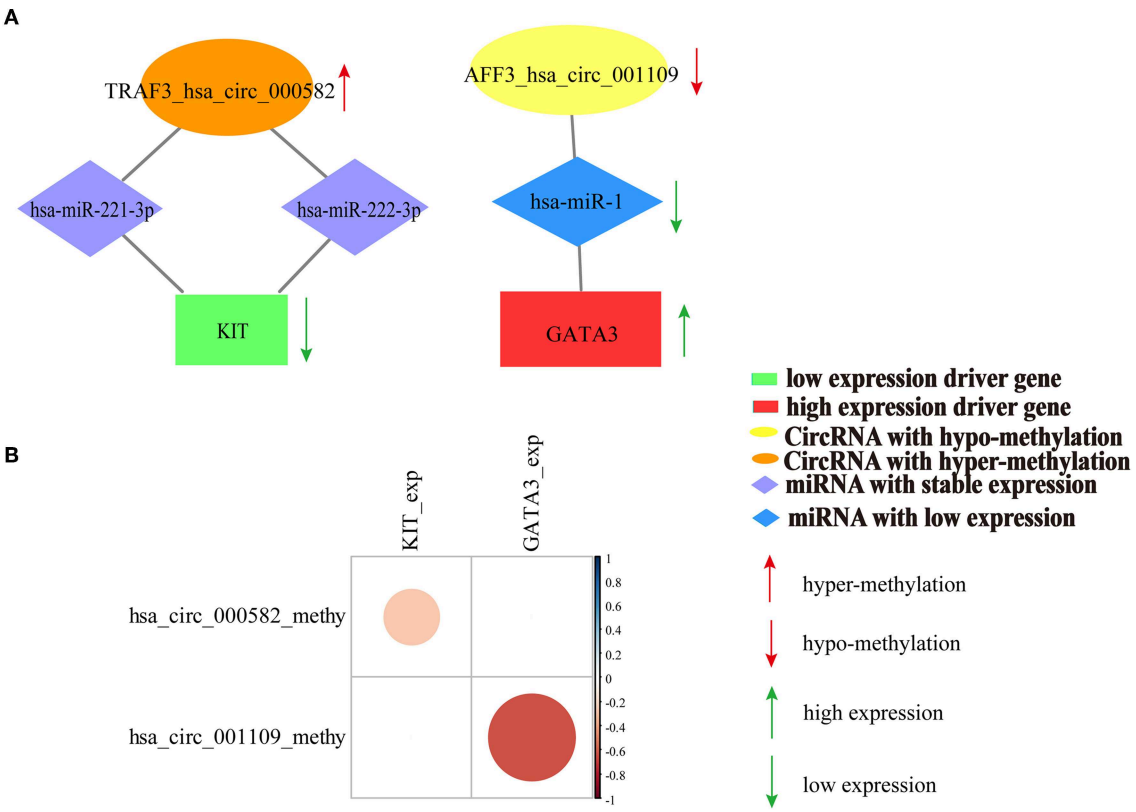


FIGURE 7 | Relationship between the DNA methylation of the predicted two sponge circRNAs (hsa_circ_000582, hsa_circ_001109) and expression of their sponge driver genes (*KIT*, *GATA3*). **(A)** Competition interaction diagram of the predicted two sponge circRNAs (hsa_circ_000582, hsa_circ_001109) and their sponge driver genes (*KIT*, *GATA3*). **(B)** The correlation coefficient diagram of DNA methylation of the predicted two sponge circRNAs (hsa_circ_000582, hsa_circ_001109) and expression of their sponge driver genes (*KIT*, *GATA3*). The scalable size of circles in correlation coefficient diagram represent the absolute value of the correlation coefficient.

DISCUSSION

CircRNAs are a class of non-coding RNAs that regulate gene expression at both transcriptional and post-transcriptional levels. Recent studies have shown that circRNAs can also function as miRNA sponges (Ebert and Sharp, 2010; Leonardo Salmena et al., 2011). CircRNAs play an important role in cancer and can be used as biological markers for prognosis.

In our study, we consider the effect of transcriptional and post-transcriptional levels on breast cancer. First, genomic, epigenetic genomics, and non-coding RNA data were integrated. Next, we constructed a DMCC network and a CMD gene network. We designed a computational process to predict sponge circRNA-driver gene pairs in breast cancer to evaluate the effect of DNA methylation of sponge circRNA host genes on circRNA sponge function. The regulation of sponge circRNA on the protein coding driver genes is not one to one, but there exists a mixed regulation. *ERBB2* showed a large number of predicted sponge circRNAs compared with other examined genes, indicating that *ERBB2* may have better circRNA sponge regulation function. ERBB2 is a 185 kDa cell membrane receptor encoded by the oncogene erbB-2, a member of the epidermal growth factor receptor family. Higher Ras-MAPK and PI3K-Akt signaling activity is detected in ERBB2 overexpressing tumor cells, which show stronger cell proliferation ability. Survival analysis showed that DNA methylation of sponge circRNA host genes could significantly differentiate breast cancer samples in two data sets. These results suggest that the predicted sponge circRNAs can be used as a prognostic biomarker for breast cancer.

The experimental methods that are currently used to verify sponge circRNAs require extensive time and are costly. Therefore, using bioinformatics calculation methods to identify sponge circRNAs may be more advantageous. While our study focused on predicting sponge circRNAs in breast cancer, these methods can be extended to other cancer types. We provide a new approach to study the regulatory effects of sponge circRNAs associated with DNA methylation in cancer, which may be helpful for understanding of the mechanism of competitive endogenous RNAs (ceRNA).

Our research reveals a complex DNA methylation-mediated circRNA sponge regulatory mechanism (**Supplementary Figure 7**). Interestingly, four of the 10 Sponge circRNAs (hsa_circ_000582, hsa_circ_000691, hsa_circ_000877, hsa_circ_000690) were identical to the breast cancer-specific circRNAs from previous studies (Coscujuela Tarrero et al., 2018). Sponge circRNAs may lead to abnormal expression of important protein coding driver genes in breast cancer. Furthermore, the DNA methylation-mediated disruption of circRNA sponge regulation may be a useful target for cancer treatment.

The quantification of the DNA methylation of circRNA host genes in this study is based on the Illumina Infinium HumanMethylation 450 BeadChip of breast cancer in the TCGA database, which covered only 482,421 CpG sites, whereas WGBS covers more CpG sites. Therefore, using high throughput sequencing data such as WGBS to quantify the DNA methylation of circRNA host genes may be more accurate, but still shows the problem of insufficient sample size.

Our research reveals a complex DNA methylation-mediated regulation of circRNA sponges. Sponge circRNAs bind to miRNAs to prevent their interactions with target genes, and sponge circRNAs showed the same DNA methylation pattern and expression pattern as the sponge-driven genes. Sponge circRNAs may significantly contribute to the abnormal expression of important protein-coding genes in breast cancer. The DNA methylation-mediated circRNA sponge regulation may be a potential target for cancer therapy. Our findings also show a new perspective of circRNA as a miRNA sponge in the pathogenesis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://gdc.cancer.gov/GSE78754.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the working institution and in accordance with the Helsinki Declaration (revised in Fortaleza, Brazil, October 2013). 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

YG, CCi, and SZ contributed to the study design, identification of the level of DNA methylation of circular RNA, analysis of data and manuscript draft. YZ contributed to the study design and supervision. XZ provided help about the mechanism of circRNA during the manuscript revision. MS, WL, CCh, and HL were responsible for data search and figure beautification. DZ contributed to the study supervision.

FUNDING

This work was supported by National Natural Science Foundation of China (grant number 61972116).

ACKNOWLEDGMENTS

We thank Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00365/full#supplementary-material

REFERENCES

Abu, N., and Jamal, R. (2016). Circular RNAs as promising biomarkers: a mini-review. *Front. Physiol.* 7:355. doi: 10.3389/fphys.2016.00355

Aran, D., Sabato, S., and Hellman, A. (2013). DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* 14:R21. doi: 10.1186/gb-2013-14-3-r21

Ayca, G., and Traina, T. A. (2011). Triple-negative breast cancer: adjuvant therapeutic options. *Chemother. Res. Pract*. 2011:696208. doi: 10.1155/2011/696208

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215–233. doi: 10.1016/j.cell.2009.01.002

Baylin, S. B., and Jones, P. A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. *Nat. Rev. Cancer* 11, 726–734. doi: 10.1038/nrc3130

Berger, S. L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes Dev*. 23:781–783. doi: 10.1101/gad.1787609

Bird, A. P. (1984). DNA methylation–how important in gene control? *Nature* 307, 503–504. doi: 10.1038/307503a0

Bird, A. P. (1996). The relationship of DNA methylation to cancer. *Cancer Surv*. 28, 87–101.

Chen, L. L., and Yang, L. (2015). Regulation of circRNA biogenesis. *RNA Biol.* 12, 381–388. doi: 10.1080/15476286.2015.1020271

Coscujuela Tarrero, L., Ferrero, G., Miano, V., De Intinis, C., Ricci, L., Arigoni, M., et al. (2018). Luminal breast cancer-specific circular RNAs uncovered by a novel tool for data analysis. *Oncotarget* 9, 14580–14596. doi: 10.18632/oncotarget.24522

Deaton, A. M., Bird, A. (2011). CpG islands and regulation of transcription. *Genes Dev*. 25, 1010–1022. doi: 10.1101/gad.2037511

Diener, T. O. (1989). Circular RNAs: relics of precellular evolution? *Proc. Natl. Acad. Sci. U.S.A*. 86:9370. doi: 10.1073/pnas.86.23.9370

Du, Z., Sun, T., Hacisuleyman, E., Fei, T., Wang, X., Brown, M., et al. (2016). Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer. *Nat. Commun.* 7:10982. doi: 10.1038/ncomms10982

Ebert, M. S., and Sharp, P. A. (2010). Emerging roles for natural microRNA sponges. *Curr. Biol.* 20, R858–R861. doi: 10.1016/j.cub.2010.08.052

Enuka, Y., Lauriola, M., Feldman, M. E., Sas-Chen, A., Ulitsky, I., Yarden, Y., et al. (2016). Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. *Nucleic Acids Res.* 44, 1370–1383. doi: 10.1093/nar/gkv1367

Esteller, M. (2008). Epigenetics in cancer. *N. Engl. J. Med*. 358, 1148–1159. doi: 10.1056/NEJMra072067

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Hansen, T. B., Wiklund, E. D., Bramsen, J. B., Villadsen, S. B., Statham, A. L., Clark, S. J., et al. (2011). miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J.* 30, 4414–4422. doi: 10.1038/emboj.2011.359

Iorio, M. V., Ferracin, M., Liu, C. G., Veronese, A., Spizzo, R., Sabbioni, S., et al. (2015). MicroRNA gene expression deregulation in human breast cancer. *Cancer Res*. 65, 7065–7070. doi: 10.1158/0008-5472.CAN-05-1783

Jens, M. (2014). "Circular RNAs are a large class of animal RNAs with regulatory, potency," in *Dissecting Regulatory Interactions of RNA and Protein* (Switzerland: Springer International Publishing), 333.

Koboldt, D. C., Fulton, R., Mclellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumors. *Nature* 490, 61–70. doi: 10.1038/nature11412

Kulcheski, F. R., Christoff, A. P., and Margis, R. (2016). Circular RNAs are miRNA sponges and can be used as a new class of biomarker. *J. Biotechnol.* 238, 42–51. doi: 10.1016/j.jbiotec.2016.09.011

Lehmann-Werman, R., Neiman, D., Zemmour, H., Moss, J., Magenheim, J., Vaknin-Dembinsky, A., et al. (2016). Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci. U.S.A*. 113, E1826–E1834. doi: 10.1073/pnas.1519286113

Leonardo Salmena, L. P., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

Li, J., Yang, J., Zhou, P., Le, Y., Zhou, C., Wang, S., et al. (2015). Circular RNAs in cancer: novel insights into origins, properties, functions and implications. *Am. J. Cancer Res*. 5, 472–480.

Lindahl, T. (1981). DNA methylation and control of gene expression. *Nature* 290, 363–364. doi: 10.1038/290363b0

Lü, L., Sun, J., Shi, P., Kong, W., Xu, K., He, B., et al. (2017). Identification of circular RNAs as a promising new class of diagnostic biomarkers for human breast cancer. *Oncotarget* 8, 44096–44107. doi: 10.18632/oncotarget.17307

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92–94. doi: 10.1038/nature24284

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 47–54. doi: 10.1038/nature17676

Pan, Y., Liu, G., Zhou, F., Su, B., and Li, Y. (2017). DNA methylation profiles in cancer diagnosis and therapeutics. *Clin. Exp. Med*. 18, 1–14. doi: 10.1007/s10238-017-0467-0

Peng, L., Chen, G., Zhu, Z., Shen, Z., Du, C., Zang, R., et al. (2017). Circular RNA ZNF609 functions as a competitive endogenous RNA to regulate AKT3 expression by sponging miR-150-5p in Hirschsprung's disease. *Oncotarget* 8, 808–818. doi: 10.18632/oncotarget.13656

Peng, L., Yuan, X. Q., and Li, G. C. (2015). The emerging landscape of circular RNA ciRS-7 in cancer (Review). *Oncol. Rep*. 33, 2669–2674. doi: 10.3892/or.2015.3904

Qiang, L., Zhang, X., Hu, X., Dai, L., Fu, X., Zhang, J., et al. (2016). Circular RNA related to the chondrocyte ECM regulates MMP13 expression by functioning as a MiR-136 'Sponge' in human cartilage degradation. *Sci. Rep*. 6:22572. doi: 10.1038/srep22572

Qu, S., Yang, X., Li, X., Wang, J., Gao, Y., Shang, R., et al. (2015). Circular RNA: a new star of noncoding RNAs. *Cancer Lett*. 365, 141–148. doi: 10.1016/j.canlet.2015.06.003

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., and McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235, 177–182. doi: 10.1126/science.3798106

Song, X., Zhang, N., Han, P., Moon, B. S., Lai, R. K., Wang, K., et al. (2016). Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res*. 44, e87–e87. doi: 10.1093/nar/gkw075

Trimarchi, M. P., Mouangsavanh, M., and Huang, T. H. (2011). Cancer epigenetics: a perspective on the role of DNA methylation in acquired endocrine resistance. *Chin. J. Cancer* 30, 749–756. doi: 10.5732/cjc.011.10128

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Kinzler, K. W., et al. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Weil, P. P., Hensel, K. O., Weber, D., and Postberg, J. (2015). An expanding universe of noncoding RNAs between the poles of basic science and clinical investigations. *Epigenomics* 8, 317–319. doi: 10.2217/epi.15.65

Wise, I. A., and Charchar, F. J. (2016). Epigenetic modifications in essential hypertension. *Int. J. Mol. Sci*. 17:451. doi: 10.3390/ijms17040451

Yang, Q., Du, W. W., Wu, N., Yang, W., Awan, F. M., Fang, L., et al. (2017). A circular RNA promotes tumorigenesis by inducing c-myc nuclear translocation. *Cell DeathDiff*. 24, 1609–1620. doi: 10.1038/cdd.2017.86

Yang, X., Yan, L., and Davidson, N. E. (2001). DNA methylation in breast cancer. *Endocr. Relat. Cancer* 8:115. doi: 10.1677/erc.0.0080115

Zhang, T., Li, L., Shang, Q., Lv, C., Wang, C., and Su, B. (2015). Circulating miR-126 is a potential biomarker to predict the onset of type 2 diabetes mellitus in susceptible individuals. *Biochem. Biophys. Res. Commun.* 463, 60–63. doi: 10.1016/j.bbrc.2015.05.017

Zheng, Q., Bao, C., Guo, W., Li, S., Chen, J., Chen, B., et al. (2016). Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nat. Commun.* 7:11215. doi: 10.1038/ncomms11215

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 © 2019 Gu, Ci, Zhang, Su, Lv, Chen, Liu, Zhang, Zhang 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 Genetics

ORIGINAL RESEARCH
published: 28 November 2019
doi: 10.3389/fgene.2019.01228

Identification of Common and Subtype-Specific Mutated Sub-Pathways for a Cancer

Haidan Yan[1,2†], Xusheng Deng[1,2†], Haifeng Chen[3], Jun Cheng[1,2], Jun He[1,2], Qingzhou Guan[1,2], Meifeng Li[1,2], Jiajing Xie[1,2], Jie Xia[1,2], Yunyan Gu[4] and Zheng Guo[1,2]*

[1] Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Department of Bioinformatics, The School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China, [2] Key Laboratory of Medical Bioinformatics, Fujian Medical University, Fuzhou, China, [3] Department of General Surgery, Fuzhou Second Hospital Affiliated to Xiamen University, Xiamen, China, [4] Department of Systems Biology, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

The heterogeneity of cancer is a big obstacle for cancer diagnosis and treatment. Prioritizing combinations of driver genes that mutate in most patients of a specific cancer or a subtype of this cancer is a promising way to tackle this problem. Here, we developed an empirical algorithm, named PathMG, to identify common and subtype-specific mutated sub-pathways for a cancer. By analyzing mutation data of 408 samples (Lung-data1) for lung cancer, three sub-pathways each covering at least 90% of samples were identified as the common sub-pathways of lung cancer. These sub-pathways were enriched with mutated cancer genes and drug targets and were validated in two independent datasets (Lung-data2 and Lung-data3). Especially, applying PathMG to analyze two major subtypes of lung cancer, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LSCC), we identified 13 subtype-specific sub-pathways with at least 0.25 mutation frequency difference between LUAD and LSCC samples in Lung-data1, and 12 of the 13 sub-pathways were reproducible in Lung-data2 and Lung-data3. Similar analyses were done for colorectal cancer. Together, PathMG provides us a novel tool to identify potential common and subtype-specific sub-pathways for a cancer, which can provide candidates for cancer diagnoses and sub-pathway targeted treatments.

Keywords: mutation, common sub-pathways, subtype-specific sub-pathways, cancer genes, drug targets

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Yang Wang,
Hubei University, China
Hao Lin,
University of Electronic Science and Technology of China, China

****Correspondence:***
Zheng Guo
guoz@ems.hrbmu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *04 September 2019*
Accepted: *06 November 2019*
Published: *28 November 2019*

Citation:
Yan H, Deng X, Chen H, Cheng J, He J, Guan Q, Li M, Xie J, Xia J, Gu Y and Guo Z (2019) Identification of Common and Subtype-Specific Mutated Sub-Pathways for a Cancer. Front. Genet. 10:1228. doi: 10.3389/fgene.2019.01228

INTRODUCTION

Thousands of mutations are detected for a cancer with the advances of DNA sequencing technologies. The mutation frequencies of most genes are very low (<5%) in all patients of a cancer (Ciriello et al., 2013; Kandoth et al., 2013; Tamborero et al., 2013). Therefore, many algorithms have been developed to identify a panel of genes or pathways that mutate in a significantly high fraction of patients in a particular type of cancer. These identified mutation genes or pathways might be drivers contributing to cancer (Youn and Simon, 2011; Dees et al., 2012; Hua et al., 2013; Merid et al., 2014; Leiserson et al., 2015) or potential diagnosis biomarkers for a cancer (Ece Solmaz et al., 2015; Clifford et al., 2016; Li et al., 2016; Sato et al., 2016). For example, Clifford et al. identified a panel of 400 mutations covering more than 80% of the lung adenocarcinoma (LUAD) patients from The Cancer Genome Atlas (TCGA) database (Clifford et al., 2016). However, in an independent validation dataset, this panel of mutations only covered 55% of 183 patients (Clifford et al., 2016). It is not surprising that

the coverage drops so much in the validation dataset because the distribution of somatic mutations is highly heterogeneous (N. Cancer Genome Atlas Research, 2011; N. Cancer Genome Atlas, 2012; Hofree et al., 2013). Thus, the panel of mutated genes, identified only by the mutation information of individual genes, may vary across different independent datasets.

It has been reported that certain pathways are frequently altered across patients of a cancer by mutations in different genes of the pathways (N. Cancer Genome Atlas Research, 2008; Gu et al., 2011; N. Cancer Genome Atlas, 2012). Therefore, a combination of the individual mutations within a pathway (Vaske et al., 2010; Bertrand et al., 2015; Cho et al., 2016; Hristov and Singh, 2017; Shrestha et al., 2017) or a molecular network (Hristov and Singh, 2017) is a preferred method to deal with inter-tumor heterogeneity. To obtain a small subset genes that was more relevant to disease, many methods have been developed to identify sub-pathways or sub-networks. For sub-pathway analysis, most methods are based on the enrichment analysis of differentially expressed genes, such as topology enrichment analysis framework (Judeh et al., 2013), pathway and transcriptome information (Nam et al., 2014), and Subpathway-GM (Li et al., 2013). The methods to extract sub-networks mainly combined mutations with copy number variations to identify modules related with diseases, such as HotNet (Leiserson et al., 2015) and MEMo (Ciriello et al., 2012). Panels of mutation genes have been reported to be a promising way to diagnose a specific cancer (Ece Solmaz et al., 2015; Clifford et al., 2016; Li et al., 2016; Sato et al., 2016). It would be of great significance if we could find sub-pathways mutated in almost all patients of a cancer. Here, we think that the panel of mutation genes within a common sub-pathway will be a reliable diagnostic marker for a cancer when the common mutated sub-pathway is reproducible in different independent datasets of this cancer. However, current methods didn't consider this application of sub-pathways. On the other hand, a cancer may have different subtypes with different causes and clinical outcomes. Thus, it is also important to obtain subtype-specific biomarkers to guide subtype diagnoses and treatments.

In this study, we developed an empirical algorithm, called PathMG, to identify common and subtype-specific mutated sub-pathways for a cancer. By analyzing multiple mutation profiles of lung cancer, three reproducible common mutated sub-pathways were identified. PathMG was also used to identify LUAD-specific and lung squamous cell carcinoma (LSCC)-specific sub-pathways, respectively. Based on the subtype-specific sub-pathways, we further identified the sub-pathways related to the prognosis of lung cancer. Similarly, we also identified common and subtype-specific sub-pathways for colorectal cancer (CRC). PathMG is available on the web at https://github.com/dxsbiocc/C-Sub.

MATERIALS AND METHODS

Data and Preprocessing

As described in **Table 1**, the public available somatic mutation profiles, measured by whole-exome sequencing for lung cancer and CRC from six different studies (N. Cancer Genome Atlas, 2012; N. Cancer Genome Atlas Research, 2012; N. Cancer Genome Atlas Research, 2014; Campbell et al., 2016; Giannakis et al., 2016; Ellrott et al., 2018), were downloaded from the cBioPortal (http://www.cbioportal.org/) database. The mutation profiles of 230 LUAD samples (N. Cancer Genome Atlas Research, 2014) and 178 LSCC samples from Lung-data1 (N. Cancer Genome Atlas Research, 2012) were integrated to identify commonly mutated sub-pathways for lung cancer. The identified common sub-pathways were validated in two independent datasets (Lung-data2 and Lung-data3). For CRC, the mutation profiles of 619 samples from CRC-data1 were used to identify commonly mutated sub-pathways, while two publicly available independent datasets (CRC-data2 and CRC-data3) and one dataset (CRC-data4) measured by our laboratory were used for validation. We measured 13 samples of CRC from five different patients by whole-exome sequencing. For each patient, three specimens were sampled in three different locations. Two specimens with poor DNA quality were excluded from the analysis. The proportion of the tumor epithelial cell was measured by pathological section analysis, ranging from 40 to 100% (**Supplementary Table 1**). This study was approved by the institutional review boards of all participating institutions, and written consent forms were obtained from all participants. All cancer samples were collected from the operating room immediately after surgical resection and were fresh frozen for subsequent DNA extraction. The quantity and quality of extracted DNA was estimated with Qubit 2.0 Fluorometer (Life Technologies, Foster City, CA) by using 2 μl of undiluted DNA solution. The resulting raw whole-exome sequencing files (.fastq) were preprocessed using Trimmomatic (version 0.38) (Bolger et al., 2014), and reads were aligned to the reference genome (GRCh37) using Burrows-Wheeler aligner [BWA; version 0.7.1 (Li and Durbin, 2009)]. Finally, the variant calling was done with variant caller Mutect2 algorithm in GATK4 with high stringency parameters (Cibulskis et al., 2013).

TABLE 1 | Description of mutation data used in this study.

Data	Cancer type	Samples	References
Lung-data1	LUAD	230	(N. Cancer Genome Atlas Research, 2014)
	LSCC	178	(N. Cancer Genome Atlas Research, 2012)
Lung-data2	LUAD	562	(Ellrott et al., 2018)
	LSCC	469	
Lung-data3	LUAD	438	(Campbell et al., 2016)
	LSCC	308	
CRC-data1	CRC	619	(Giannakis et al., 2016)
CRC-data2		536	(Ellrott et al., 2018)
CRC-data3		224	(N. Cancer Genome Atlas, 2012)
CRC-data4		13	–

Kyoto Encyclopedia of Genes and Genomes Pathways

The 239 pathways covering 6,688 unique genes were downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2010) on October 21, 2018. Here, the human disease pathways were excluded from this study. For each pathway, the interactions between genes were also collected for the following analysis.

Identify Significantly Mutated Pathways

For a pathway in a given sample, we assume that the pathway is mutated in the sample if at least one gene within the pathway is mutated (Gu et al., 2011; Vandin et al., 2011). Then, for a given pathway *i*, we calculated the number of samples mutated in this pathway for a dataset with *N* samples, denoted as M_i. To test whether the number of mutated samples in the pathway *i* was significantly more than expected by chance, random experiments were performed. For a cancer sample, we calculated the number of mutated genes from its real mutation profile. Simultaneously, to produce a simulated mutation profile for the cancer sample, we randomly selected the same number of genes from the background genes as mutated genes. The total Refseq genes were defined as the background genes. After a random experiment, a random mutation dataset with *N* samples was produced. For a pathway *i*, we could calculate the number of randomly mutated samples after a random experiment, denoted as R_i. The random experiment was repeated *n* (default 1,000) times, which may be adjusted by users. Then, the probability p_i, that the number of randomly mutated samples (R_i) of the pathway *i* is greater than the number of real mutated samples (M_i), is calculated as follows:

$$p_i = \frac{\sum_{r=1}^{n} H_r}{n} \quad (1)$$

In a random experiment, if $R_i > M_i$, then $H_r = 1$; Otherwise, $H_r = 0$. The formula was used to calculate the *p* values of all pathways, and the *p* values were adjusted using the Benjamini-Hochberg (BH) method to control the false discovery rate (FDR).

Identification of Common Sub-Pathways

After identifying the significantly mutated pathways for a cancer, we further extracted common sub-pathways in each of the significant pathways by the greedy search algorithm. Here, the sub-pathways, covering more than 90% (a default parameter) of samples of this cancer, were defined as common sub-pathways. The detailed algorithm to identify common sub-pathways in a given pathway is shown as follows (**Figure 1**).

Step 1: Based on the interactions between genes in a pathway annotated by KEGG database, convert a significant pathway to an undirected graph, and obtain all the connected components on the graph.
Step 2: For a certain connected component, the genes within the connected component are ranked according to the mutation frequency (denoted as *f*) of each gene. Select the gene with the largest *f* value as a seed for an initial sub-pathway.
Step 3: For each of the direct interaction neighbor genes of the seed, calculate the increased coverage (denoted as *Pf*) when the gene is added to the sub-pathway.
Step 4: The direct interaction neighbor genes of the seed will be added to the sub-pathway one by one according to their *Pf* value from high to low. The direct neighbor genes of the seed will be divided into two sets based on whether it increases the coverage of the sub-pathway during the adding process. Then, the genes increasing the coverage are defined as set *GI*; otherwise, defined as set *GN*. The genes in set *GI* are remained in the sub-pathway as new seeds. For each of the genes from set *GN*, its direct neighbor genes will be added one by one according to their *Pf* value from high to low. If at least one neighbor gene could increase the coverage of the sub-pathway, the gene in set *GN* will be remained in the sub-pathway, and its neighbor genes increasing the coverage will be used as new seeds; otherwise, the gene in set *GN* will be excluded. So, we allowed a gene that cannot increase coverage in the process of adding genes.
Step 5: Based on the new seed genes identified in step 4, repeat step 3–4 until the seed genes don't have direct interaction neighbors in the connected component.

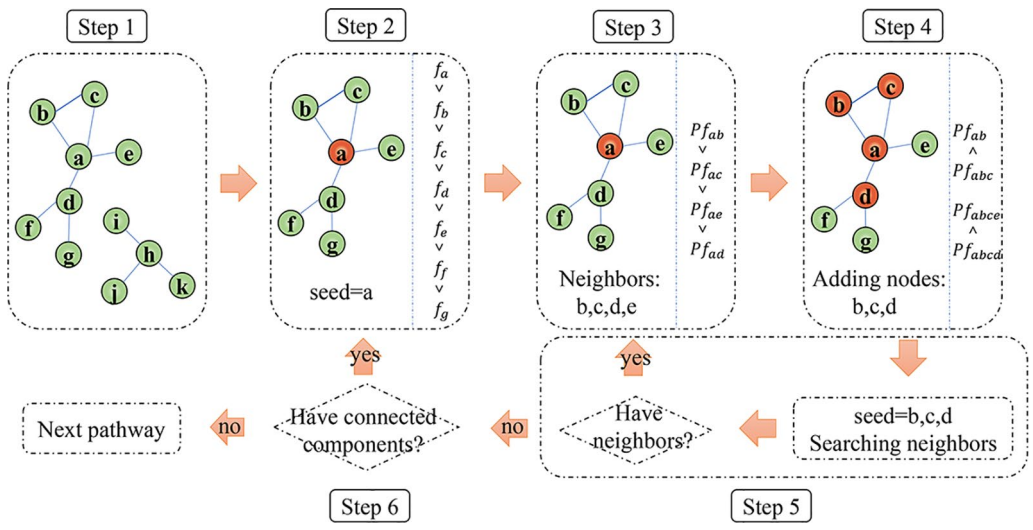


FIGURE 1 | The schematic diagram of the algorithm to identify common mutation sub-pathways. Orange nodes denote genes remained in the sub-pathway.

If the coverage of the sub-pathway is higher than a predefined parameter (default 90%) and the number of genes in the sub-pathway is at least five (default five), the sub-pathway will be output as a common sub-pathway. The two parameters, the frequency to define common sub-pathway and the minimum number of genes for a sub-pathway, could be adjusted by users.

Step 6: For all the connected components obtained from a significant pathway, repeat steps 2–5 to identify all the common sub-pathways.

Identification of Subtype-Specific Mutated Pathways and Sub-Pathways

First, the Fisher's exact test is used to identify the pathways that have significantly different mutation frequencies between two subtypes (subtype *A* and subtype *B*) of a cancer. The significant pathways are defined as subtype-specific mutated pathways. The *p* values are adjusted using the BH method to control the FDR.

After identifying the significantly subtype-specific mutated pathways for a cancer, the sub-pathways, that make the differences of the mutation frequencies between two subtypes as larger as possible, are further extracted in each of the significant pathways using the greedy search algorithm. Similar with the method to identify common sub-pathways, we also integrated the differences of mutation frequencies between subtype *A* and subtype *B* of one cancer and pathway information to identify subtype-specific sub-pathways (**Supplementary Figure 1**). The detailed algorithm to identify subtype-specific sub-pathways in a given pathway is described as follows.

Step 1: Convert a subtype-specific pathway to an undirected graph based on the interactions between genes annotated by KEGG database, and obtain all the connected components on the graph.

Step 2: For a certain connected component, calculate the mutation frequencies of each gene in subtype *A* and subtype *B* of one cancer, respectively, denoted as f_a and f_b. For a given gene, the mutation frequency difference between subtype *A* and subtype *B* is defined as $v = f_a - f_b$. According to the v value, the genes with $v>0$ are defined as subtype *A* specific genes (denoted as set G_a) and the genes with $v>0$ are defined as subtype *A* specific genes (denoted as set G_a) and the genes with $v<0$ are defined as subtype *B* specific genes (denoted as set G_b). Then, the two classes of genes (G_a and G_b) are used to identify subtype *A* specific and subtype *B* specific sub-pathways, respectively. Here, we example the process to identify subtype *A* specific sub-pathways to explain the algorithm.

Step 3: Firstly, select the gene with the largest $|v|$ value among G_a as a seed for an initial sub-pathway.

Step 4: For each of the direct interaction neighbor genes of the seed in G_a, calculate the increasing of coverage difference (denoted as *Pv*) between subtype *A* and subtype *B* when the gene is added to the sub-pathway.

Step 5: The direct neighbor genes of the seed in G_a will be added to the sub-pathway one by one according to their *Pv* value from high to low. When direct neighbor genes added to the sub-pathway increase the coverage differences, the genes will be remained in the sub-pathway as new seeds. Similarly, PathMG allows for at most one gene that doesn't increase the coverage difference. Therefore, for a gene that cannot increase coverage difference, its direct neighbor genes will be added one by one according to their *Pv* value from high to low. If at least one neighbor gene could increase the coverage difference of the sub-pathway, the gene will be remained in the sub-pathway and its neighbor genes increased coverage will be used as new seeds; otherwise, the gene will be excluded.

Step 6: Based on the new seeds identified in step 5, repeat step 4–5 until the seeds don't have direct neighbors among G_a in the connected component. Calculate the *p* values of Fisher's exact test ($p < 0.05$) for the sub-pathway, and output the subtype *A* specific sub-pathway if its mutation frequency in subtype *A* is higher than subtype *B* with a predefined parameter (default 0.25). Equally, we can identify subtype *B* specific sub-pathways for the connected component.

Step 7: For all the connected components, obtained from a subtype-specific pathway, repeat steps 2–6 to identify all the subtype *A* specific and subtype *B* specific sub-pathways, respectively.

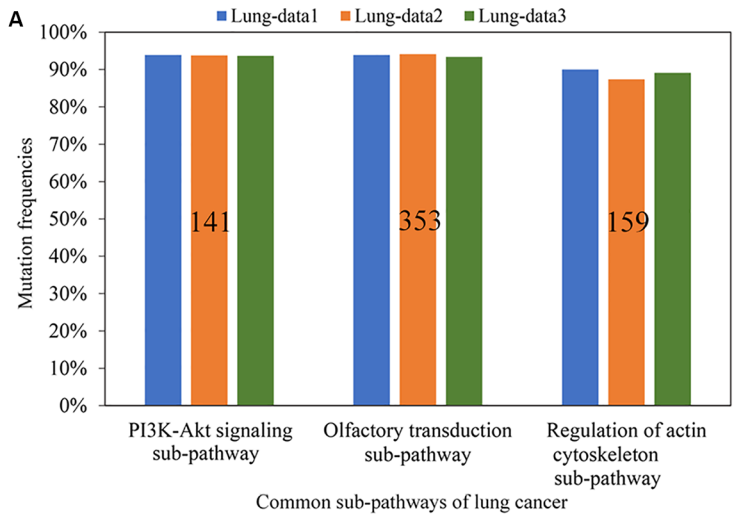
Sub-Pathways Enriched With Cancer Genes and Drug Target Genes

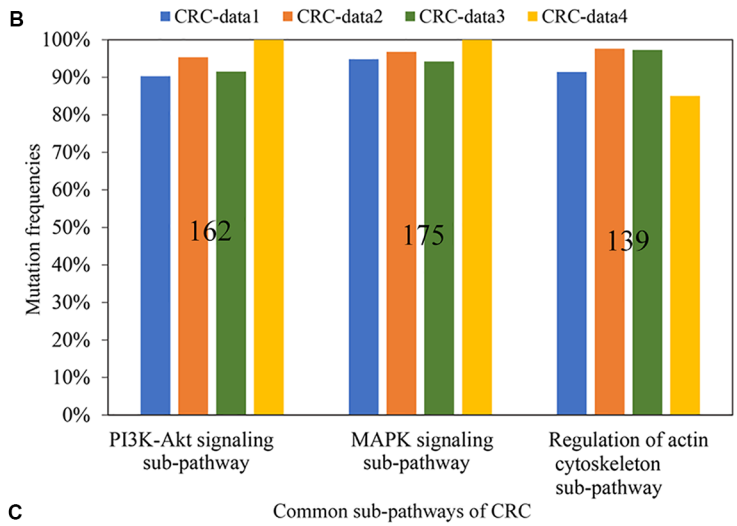
The cancer genes were downloaded from the Catalogue of Somatic Mutations in Cancer (COSMIC) database, which collected a total of 719 cancer genes (Bamford et al., 2004). We also collected 7,463 target genes for the commonly used drugs for lung cancer therapy, such as carboplatin, cisplatin, and docetaxel, from the Comparative Toxicogenomics Database (http://ctdbase.org/) (Davis et al., 2009). Among the 7,463 target genes, 2,661 genes were included in the KEGG pathways. Simultaneously, 935 target genes for the commonly used drugs for CRC therapy were also downloaded from the Comparative Toxicogenomics Database, among which 527 genes were included in the KEGG pathways.

RESULTS

Identify Common Sub-Pathways for Lung Cancer and Colorectal Cancer

Firstly, random experiments were done to identify the significantly mutated pathways. Then, the sub-pathways commonly mutated in at least 90% of patients were identified in each of the significantly mutated pathways. Here, considering the existence of large measurement variation and low quality of tissue samples, we defined the sub-pathways covering more than a predefined parameter (here, default 90%) of patients rather than 100% of patients in a cancer as commonly mutated sub-pathways. The details of the methods were described in *Materials* and *Methods*.





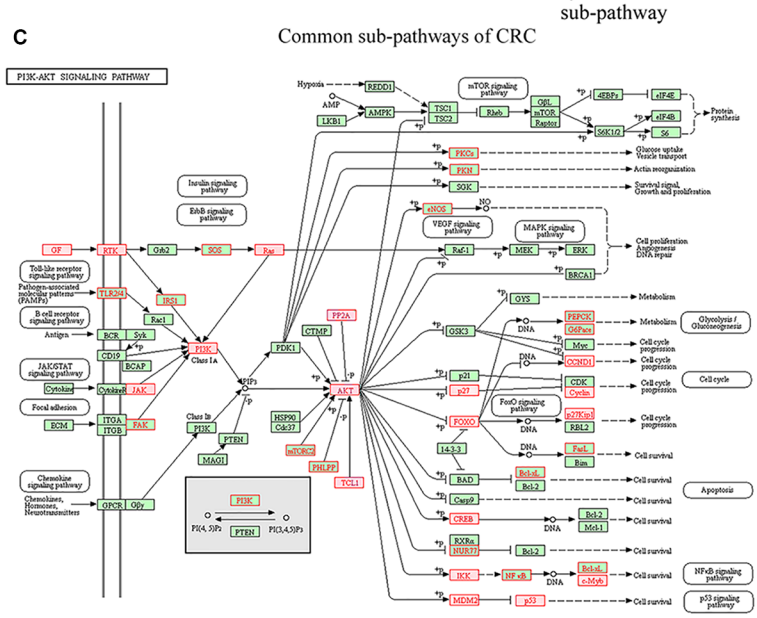


FIGURE 2 | The mutation frequencies of the common sub-pathways across different datasets for lung cancer **(A)** and CRC **(B)**, respectively. The figures on the bars represent the number of genes within the identified sub-pathways. **(C)** The sub-pathway identified from PI3K-Akt signaling pathway. The genes with red font were genes in the sub-pathway and the genes in squares filled with red color were cancer genes.

Using the 408 mutation profiles of Lung-data1, we firstly identified 116 significantly mutated pathways for lung cancer (FDR < 0.05). Then, three sub-pathways, PI3K-Akt signaling sub-pathway, olfactory transduction sub-pathway, and regulation of actin cytoskeleton sub-pathway, were identified as the common sub-pathways of lung cancer (**Figure 2A**). In the two independent validation datasets (Lung-data2 for 1,031 samples and Lung-data3 for 746 samples), two of the three common sub-pathways covered at least 93% samples, whereas the regulation of actin cytoskeleton sub-pathway covered 87 and 89% samples in Lung-data2 and Lung-data3, respectively. The result indicated that the common mutated sub-pathways were highly reproducible in different sets of lung cancer samples, which suggests that the mutation genes within the common sub-pathways could be candidate panels of mutation genes for lung cancer diagnosis. Moreover, except for the sub-pathway of olfactory transduction, two of the three common mutated sub-pathways were significantly enriched with cancer genes documented in the database COSMIC and target genes for the commonly used drugs for lung cancer therapy (**Supplementary Tables 2** and **3**, hypergeometric test, $p < 0.05$). For example, the PI3K-Akt signaling sub-pathway included 141 genes and 43 of them were cancer genes, which was unlikely to happen by chance (**Figure 2C**, hypergeometric test, $p = 4.28E{-}29$).

Similarly, 125 significantly mutated pathways were identified from the 619 CRC samples from CRC-data1. Among the sub-pathways identified from these significant pathways, three sub-pathways, each of which covered at least 90% of 619 CRC samples, were identified as the common sub-pathways of CRC (**Figure 2B**). Notably, the mutation frequencies of all the three common sub-pathways were higher than 91% in the two independent datasets (CRC-data2 for 536 samples and CRC-data3 for 224 samples, **Figure 2B**). Moreover, all the three reproducible common sub-pathways were significantly enriched with cancer genes documented in the database COSMIC and drug targets for commonly used drugs for colon cancer therapy (**Supplementary Tables 2** and **3**, hypergeometric test, $p < 0.05$).

With whole-exome sequencing, we further measured 13 CRC samples with different proportions of the tumor epithelial cell to validate the three common sub-pathways. For the 13 samples, nine samples were sampled from three patients each with three different tumor locations and the other four samples were sampled from two patients each with two different tumor locations. The results showed that two of the three common sub-pathways covered all the 13 CRC samples. For the remained common sub-pathway of regulation of actin cytoskeleton, it covered 11 of the 13 CRC samples (**Figure 2B**). Overall, these results suggest that these common sub-pathways may be reliable diagnosis marker for CRC even when the proportion of the tumor epithelial cell is as low as 40%.

Identify Subtype-Specific Sub-Pathways for Lung Cancer and Colorectal Cancer

Based on the mutation profiles of 230 LUAD and 178 LSCC samples from Lung-data1, 43 pathways with significantly different mutation frequencies between LUAD and LSCC were identified using Fisher's exact test (FDR < 0.05). Here, we developed an algorithm to identify subtype-specific sub-pathways with at least 0.25 mutation frequency difference between LUAD and LSCC ($p < 0.05$).

Based on the 43 subtype-specific pathways, a total of 13 subtype-specific sub-pathways with at least 0.25 mutation frequency difference between the 230 LUAD and 178 LSCC samples were identified in Lung-data1 ($p < 0.05$), including 6 LUAD-specific sub-pathways and 7 LSCC-specific sub-pathways. In the two independent Lung-data2 and Lung-data3 datasets, all the six LUAD-specific sub-pathways were validated as LUAD-specific sub-pathways. For the seven LSCC-specific sub-pathways, six were validated as LSCC-specific sub-pathways. Only one sub-pathway, inositol phosphate metabolism, had p value less than 0.05 in both two validation datasets, but its mutation frequency difference was 0.18 and 0.12 in Lung-data2 and Lung-data3, respectively (**Supplementary Table 4**). Notably, all the top five sub-pathways with the largest differences of mutation frequencies between LUAD and LSCC in Lung-data1 were reproducible in both the two independent datasets (**Figure 3**).

Based on the knowledge that LSCC patients suffered poorer prognoses than LUAD patients (Gyorffy et al., 2013), we performed

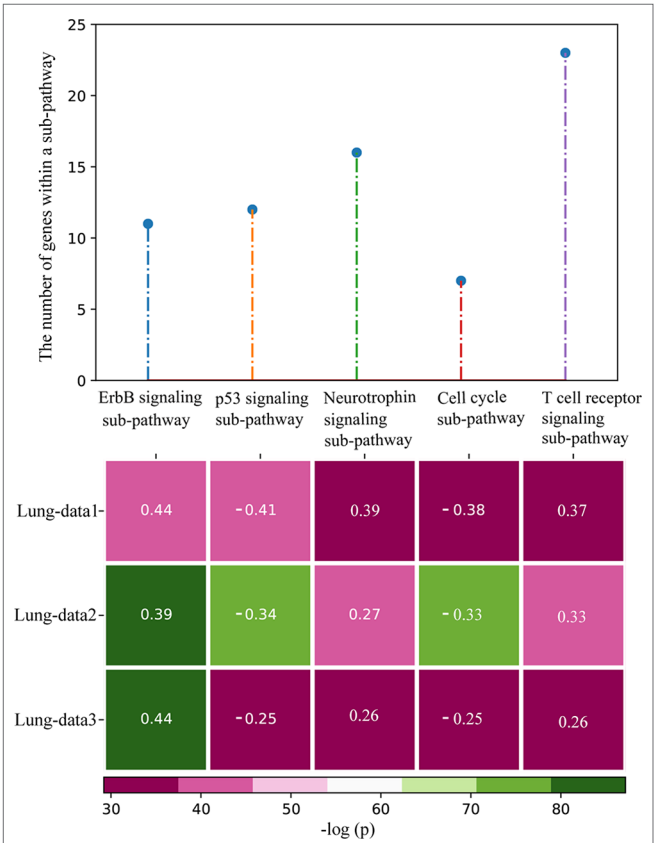


FIGURE 3 | The top five most significant subtype-specific sub-pathways with the largest differences of mutation frequencies. The heatmap shows the p values of the sub-pathways calculated by Fisher's exact test, and the figures on the heatmap represent the mutation frequency differences between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LSCC). The mutation frequency difference was calculated as the mutation frequency of the sub-pathway in LUAD minus the mutation frequency of the sub-pathway in LSCC. When the figure on the heatmap was positive (negative), the sub-pathway was LUAD-specific (LSCC-specific) sub-pathway.

survival analysis using the overall survival data of the 87 LUAD and 79 LSCC samples from Lung-data1. These patients were at the stage I and treated with complete surgical resection to exclude the effects of stage and chemotherapy on prognosis. We evaluated whether the patients with and without mutation of a sub-pathway were significantly different in overall survival (OS) time. Finally, five of the 12 reproducible subtype-specific sub-pathways were found to be associated with OS (the univariate Cox proportional-hazards regression model, $p < 0.05$) (**Supplementary Table 5**). Among the five top sub-pathways, three sub-pathways, p53 signaling pathway, T cell receptor signaling pathway, and cell cycle, were related to the OS of lung cancer. For example, a LSCC-specific sub-pathway of cell cycle, including seven genes, was mutated in 102 of 166 patients, which had significantly poorer overall survival than the other 64 patients without the mutation of this sub-pathway (log-rank $p = 0.02$, **Figure 4**).

Similarly, 221 subtype-specific pathways were identified for CRC using the mutation profiles of 166 LCC and 315 RCC samples from CRC-data1 (Fisher's exact test, FDR < 0.05). Based on the subtype-specific pathways, 42 subtype-specific sub-pathways were further identified with mutation frequency difference higher than 0.25 ($p < 0.05$). All the 42 sub-pathways were RCC-specific, which further validated the report that RCC was hyper-mutational (N. Cancer Genome Atlas, 2012). Because only CRC-data3 had the information of tumor location in the validation datasets, we then validated the 42 RCC-specific sub-pathways in CRC-data3. Among the 42 RCC-specific sub-pathways, 34 sub-pathways were validated in CRC-data3 (**Supplementary Table 4**). For the remained eight sub-pathways, only two sub-pathways had $p \geq 0.05$ and the other six sub-pathways had $p < 0.05$ with mutation frequency differences ranging from 0.17 to 0.24 in CRC-data3. Moreover, 39 of the 42 RCC-specific sub-pathways were enriched with cancer genes, and 32 of the 42 RCC-specific sub-pathways were enriched with target genes for the commonly used CRC therapy drugs (**Supplementary Table 6**, hypergeometric test, $p < 0.05$).

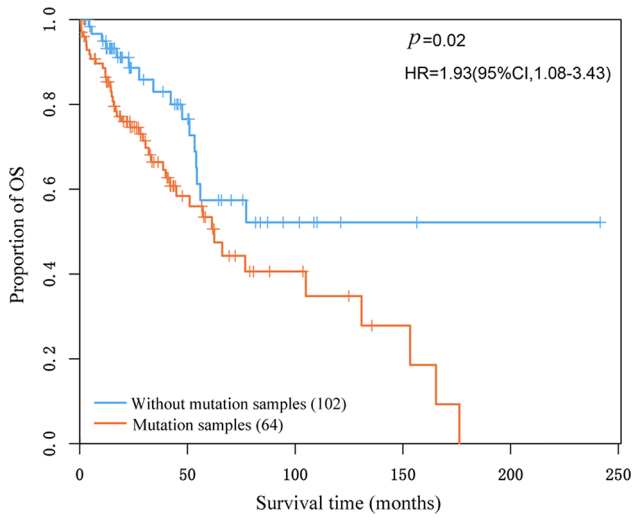


FIGURE 4 | Kaplan–Meier estimates of overall survival according to whether lung squamous cell carcinoma-specific sub-pathway of cell cycle mutated in the patients.

DISCUSSION

In this study, we developed an empirical algorithm, named PathMG, to identify commonly mutated sub-pathways for a specific cancer. For lung cancer, two of three common sub-pathways were identified from the PI3K-Akt signaling pathway and the regulation of actin cytoskeleton pathway, which were known cancer hallmarks (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2011). Both the two sub-pathways were enriched with cancer genes and cancer drug targets. Another common sub-pathway extracted from olfactory transduction pathway covered more than 90% samples of each dataset for lung cancer. It has been reported that olfactory transduction pathway can affect apoptosis of lung cancer cells (Lu et al., 2013), which may be new hallmark of lung cancer. Similarly, we also identified three common sub-pathways for CRC. Among them, two common sub-pathways were also identified from PI3K-Akt signaling pathway and Regulation of actin cytoskeleton pathway. It suggests that the two pathways may be hallmark for pan-cancer. Due to the high reproducibility of common sub-pathways, they may be reliable cancer diagnosis markers. Especially, the common sub-pathways identified in CRC were reproducible even when the proportion of the tumor epithelial cell was as low as 40%. Because the mutation profiles of circulating tumor DNA (ctDNA) were lacking, we only applied the algorithm to identify common sub-pathways using mutation profiles of tissues. However, the application of the algorithm is not restricted to tissues, it can also be used to analyze mutation profiles of ctDNA.

Simultaneously, PathMG can provide definite subtype-specific sub-pathways for a cancer with two known subtypes, which may give a novel way to identify subtype diagnosis signatures. Here, we identified six reproducible LUAD-specific sub-pathways and six reproducible LSCC-specific sub-pathways for lung cancer. Most of these sub-pathways were enriched with cancer genes and target genes for the commonly used lung cancer drugs (**Supplementary Table 6**). Similarly, we also identified 42 subtype-specific sub-pathways for CRC. All the sub-pathways were RCC-specific, which further validated that RCC had higher mutation rate than LCC (N. Cancer Genome Atlas, 2012).

Here, the default coverage to identify common sub-pathways was defined as 90% which can be adjusted by users. This parameter will affect the discovery of the number of common sub-pathways for a particular cancer. When the parameter was defined as 85%, six common sub-pathways were identified in Lung-data1. In the two independent validation datasets (Lung-data2 and Lung-data3), five of the six common sub-pathways covered at least 86% samples, whereas the remained sub-pathway of phospholipase D signaling pathway covered 84% samples in both the two validation datasets (**Supplementary Table 7**). Similarly, seven reproducible common sub-pathways were obtained for CRC (**Supplementary Table 7**). The results indicated that the common sub-pathways identified in different coverages (90 or 85%) were highly reproducible. For subtype-specific sub-pathway analysis, we considered the mutation frequency difference of a sub-pathway between two subtypes of a cancer as a parameter, and the default value was defined as 0.25. As expected, the larger the parameter, the more likely the discovered sub-pathways to be reproducible in independent validated datasets.

For example, the top 10 sub-pathways in Lung-data1 were ranked within the top 12 sub-pathways in the two independent validation datasets (**Supplementary Table 8**).

Besides, because all the mutation profiles analyzed in this study were detected by whole-exome sequencing, the total exon length of a gene will affect the mutation frequency of a gene theoretically. To analyze the effect, we studied the distribution of mutation counts among genes with different total exon lengths. The result showed that the mutated genes with total exon lengths shorter than 18,000 bp accounted for about 94% mutated genes (**Supplementary Figure 2**). The average mutation counts of these genes ranged from 4.6 to 10.3, indicating that the mutation counts didn't vary widely for most genes. Then, we didn't consider this factor. One possible way of addressing this limitation is to use the algorithm, called PathScan, to identify significant pathways which considered the effect of gene length (Wendl et al., 2011). Then, we can identify common sub-pathways based on the significant pathways identified by PathScan. The number of random experiments was also a limitation of PathMG. To assure the power of the algorithm, the experiment was repeated 1,000 times which could be adjusted by users. In this study, we only analyzed KEGG pathways to interpret the algorithm of PathMG, which will limit the number of identified common and subtype-specific sub-pathways. When using PathMG to identify diagnostic markers for a cancer, we had better integrate more pathways from different databases to obtain the optimal sub-pathway marker.

In summary, PathMG can be used to identify common and subtype-specific sub-pathways for a particular cancer, which may help users to prioritize panels of mutations at the sub-pathway level to aid cancer diagnosis and sub-pathway targeted treatment.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Fujian medical university biomedical research ethics committee. 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

ZG, HY, and XD conceived and designed the overall study. HY, XD, HC, JC, JH, QG, ML, JXie and JXia analyzed the data. HY, YG and ZG wrote the manuscript. HY, XD and HC revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China [grant numbers. 61801118]. The education research project for young and middle-aged teachers in Fujian province [grant number. JAT170214] and the Fujian natural science foundation [grant number. 2019J01678].

ACKNOWLEDGMENTS

We thank the TCGA database for sharing the mutation data of lung cancer and CRC.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01228/full#supplementary-material

REFERENCES

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., et al. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* 91 (2), 355–358. doi: 10.1038/sj.bjc.6601894

Bertrand, D., Chng, K. R., Sherbaf, F. G., Kiesel, A., Chia, B. K., Sia, Y. Y., et al. (2015). Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res.* 43 (7), e44. doi: 10.1093/nar/gku1393

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., et al. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* 48 (6), 607–616. doi: 10.1038/ng.3564

Cho, A., Shim, J. E., Kim, E., Supek, F., Lehner, B., and Lee, I. (2016). MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.* 17 (1), 129. doi: 10.1186/s13059-016-0989-x

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31 (3), 213–219. doi: 10.1038/nbt.2514

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22 (2), 398–406. doi: 10.1101/gr.125567.111

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45 (10), 1127–1133. doi: 10.1038/ng.2762

Clifford, H. W., Cassidy, A. P., Vaughn, C., Tsai, E. S., Seres, B., Patel, N., et al. (2016). Profiling lung adenocarcinoma by liquid biopsy: can one size fit all?. *Cancer Nanotechnol.* 7 (1), 10. doi: 10.1186/s12645-016-0023-8

Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wiegers, T. C., and Mattingly, C. J. (2009). Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.* 37 (Database issue), D786–D792. doi: 10.1093/nar/gkn580

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22 (8), 1589–1598. doi: 10.1101/gr.134635.111

Ece Solmaz, A., Onay, H., Atik, T., Aykut, A., Cerrah Gunes, M., Ozalp Yuregir, O., et al. (2015). Targeted multi-gene panel testing for the diagnosis of Bardet Biedl syndrome: Identification of nine novel mutations across BBS1, BBS2, BBS4,

BBS7, BBS9, BBS10 genes. *Eur. J. Med. Genet.* 58 (12), 689–694. doi: 10.1016/j.ejmg.2015.10.011

Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* 6 (3), 271–281 e7. doi: 10.1016/j.cels.2018.03.002

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 (4), 857–865. doi: 10.1016/j.celrep.2016.03.075

Gu, Y., Zhao, W., Xia, J., Zhang, Y., Wu, R., Wang, C., et al. (2011). Analysis of pathway mutation profiles highlights collaboration between cancer-associated superpathways. *Hum. Mutat.* 32 (9), 1028–1035. doi: 10.1002/humu.21541

Gyorffy, B., Surowiak, P., Budczies, J., and Lanczky, A. (2013). Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One* 8 (12), e82241. doi: 10.1371/journal.pone.0082241

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100 (1), 57–70. doi: 10.1016/S0092-8674(00)81683-9

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

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10 (11), 1108–1115. doi: 10.1038/nmeth.2651

Hristov, B. H., and Singh, M. (2017). Network-based coverage of mutational profiles reveals cancer genes. *Cell Syst.* 5 (3), 221–229 e4. doi: 10.1016/j.cels.2017.09.003

Hua, X., Xu, H., Yang, Y., Zhu, J., Liu, P., and Lu, Y. (2013). DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am. J. Hum. Genet.* 93 (3), 439–451. doi: 10.1016/j.ajhg.2013.07.003

Judeh, T., Johnson, C., Kumar, A., and Zhu, D. (2013). TEAK: topology enrichment analysis framework for detecting activated biological subpathways. *Nucleic Acids Res.* 41 (3), 1425–1437. doi: 10.1093/nar/gks1299

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502 (7471), 333–339. doi: 10.1038/nature12634

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38 (Database issue), D355–D360. doi: 10.1093/nar/gkp896

Leiserson, M. D., Wu, H. T., Vandin, F., and Raphael, B. J. (2015). CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* 16, 160. doi: 10.1186/s13059-015-0700-7

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47 (2), 106–114. doi: 10.1038/ng.3168

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, C., Han, J., Yao, Q., Zou, C., Xu, Y., Zhang, C., et al. (2013). Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic Acids Res.* 41 (9), e101. doi: 10.1093/nar/gkt161

Li, J., Meeks, H., Feng, B. J., Healey, S., Thorne, H., Makunin, I., et al. (2016). Targeted massively parallel sequencing of a panel of putative breast cancer susceptibility genes in a large cohort of multiple-case breast and ovarian cancer families. *J. Med. Genet.* 53 (1), 34–42. doi: 10.1136/jmedgenet-2015-103452

Lu, C., Xiong, M., Luo, Y., Li, J., Zhang, Y., Dong, Y., et al. (2013). Genome-wide transcriptional analysis of apoptosis-related genes and pathways regulated by H2AX in lung cancer A549 cells. *Apoptosis* 18 (9), 1039–1047. doi: 10.1007/s10495-013-0875-x

Merid, S. K., Goranskaya, D., and Alexeyenko, A. (2014). Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. *BMC Bioinformatics* 15, 308. doi: 10.1186/1471-2105-15-308

N. Cancer Genome Atlas Research (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455 (7216), 1061–1068. doi: 10.1038/nature07385

N. Cancer Genome Atlas Research (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474 (7353), 609–615. doi: 10.1038/nature10166

N. Cancer Genome Atlas Research (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489 (7417), 519–525. doi: 10.1038/nature11404

N. Cancer Genome Atlas Research (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511 (7511), 543–550. doi: 10.1038/nature13385

N. Cancer Genome Atlas (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487 (7407), 330–337. doi: 10.1038/nature11252

Nam, S., Chang, H. R., Kim, K. T., Kook, M. C., Hong, D., Kwon, C. H., et al. (2014). PATHOME: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene* 33 (41), 4941–4951. doi: 10.1038/onc.2014.80

Sato, K. A., Hachiya, T., Iwaya, T., Kume, K., Matsuo, T., Kawasaki, K., et al. (2016). Individualized mutation detection in circulating tumor DNA for monitoring colorectal tumor burden using a cancer-associated gene sequencing panel. *PLoS One* 11 (1), e0146275. doi: 10.1371/journal.pone.0146275

Shrestha, R., Hodzic, E., Sauerwald, T., Dao, P., Wang, K., Yeung, J., et al. (2017). HIT'nDRIVE: patient-specific multidriver gene prioritization for precision oncology. *Genome Res.* 27 (9), 1573–1588. doi: 10.1101/gr.221218.117

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* 3, 2650. doi: 10.1038/srep02650

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18 (3), 507–522. doi: 10.1089/cmb.2010.0265

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26 (12), i237–i245. doi: 10.1093/bioinformatics/btq182

Wendl, M. C., Wallis, J. W., Lin, L., Kandoth, C., Mardis, E. R., Wilson, R. K., et al. (2011). PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* 27 (12), 1595–1602. doi: 10.1093/bioinformatics/btr193

Youn, A., and Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* 27 (2), 175–181. doi: 10.1093/bioinformatics/btq630

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 © 2019 Yan, Deng, Chen, Cheng, He, Guan, Li, Xie, Xia, Gu and Guo. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 04 December 2019
doi: 10.3389/fbioe.2019.00388

CancerClock: A DNA Methylation Age Predictor to Identify and Characterize Aging Clock in Pan-Cancer

Tongtong Zhu[1†], Yue Gao[1†], Junwei Wang[2†], Xin Li[1], Shipeng Shang[1], Yanxia Wang[1], Shuang Guo[1], Hanxiao Zhou[1], Hongjia Liu[1], Dailin Sun[1], Hong Chen[2], Li Wang[1*] and Shangwei Ning[1*]*

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] Department of Respiratory Medicine, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Lixin Cheng,
Shenzhen People's Hospital, China
Jing Gong,
Huazhong Agricultural University, China

***Correspondence:**
Shangwei Ning
ningsw@ems.hrbmu.edu.cn
Li Wang
wangli@hrbmu.edu.cn
Hong Chen
chenhong744563@aliyun.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 06 September 2019
Accepted: 18 November 2019
Published: 04 December 2019

Citation:
Zhu T, Gao Y, Wang J, Li X, Shang S, Wang Y, Guo S, Zhou H, Liu H, Sun D, Chen H, Wang L and Ning S (2019) CancerClock: A DNA Methylation Age Predictor to Identify and Characterize Aging Clock in Pan-Cancer. Front. Bioeng. Biotechnol. 7:388. doi: 10.3389/fbioe.2019.00388

Many biological indicators related to chronological age have been proposed. Recent studies found that epigenetic clock or DNA methylation age is highly correlated with chronological age. In particular, a significant difference between DNA methylation age (m-age) and chronological age was observed in cancers. However, the prediction and characterization of m-age in pan-cancer remains an explored area. In this study, 1,631 age-related methylation sites in normal tissues were discovered and analyzed. A comprehensive computational model named CancerClock was constructed to predict the m-age for normal samples based on methylation levels of the extracted methylation sites. LASSO linear regression model was used to screen and train the CancerClock model in normal tissues. The accuracy of CancerClock has proved to be 81%, and the correlation value between chronological age and m-age was 0.939 ($P < 0.01$). Next, CancerClock was used to evaluate the difference between m-age and chronological age for 33 cancer types from TCGA. There were significant differences between predicted m-age and chronological age in large number of cancer samples. These cancer samples were defined as "age-related cancer samples" and they have some differential methylation sites. The differences between predicted m-age and chronological age may contribute to cancer development. Some of these differential methylation sites were associated with cancer survival. CancerClock provided assistance in estimating the m-age in normal and cancer samples. The changes between m-age and chronological age may improve the diagnosis and prognosis of cancers.

Keywords: chronological age, methylation age, pan-cancer, LASSO, survival

BACKGROUND

Cancer has long been a major threat to human health and age, and it is a complex disease, is often characterized by abnormal and uninhibited cell growth (Eeghen et al., 2015; Mcguire et al., 2015; Taber et al., 2016). Unhealthy living environment leads to delay in the repair of damaged body cells, which accumulate with increase in individual's chronological age, thus greatly increasing the risk of cancer (Palmer et al., 2018). Although it has been confirmed that there is a general correlation between individual age and the occurrence of cancer (Zinger et al., 2017), the degree of this correlation may be different due to the age distribution of different tumor sites.

More and more diagnostic, recurrence and prognostic biomarkers of cancer have been identified (Zhou et al., 2015a, 2016). Many age-related biomarkers such as DNA methylation, telomere length changes, cell metabolic factor regulation, transcription level, and protein expression level have been discovered and further studied (Aubert and Lansdorp, 2008; Vilchez et al., 2014; Zubakov et al., 2016). DNA methylation could modulate gene expression during development and cancer progression (Wang et al., 2013; Zhang et al., 2019). In comparison to normal somatic tissues, the cancer methylome is typically characterized by a pattern of global hypomethylation coupled with site-specific promoter hypermethylation (Urbano et al., 2019). Through comparative analysis of a variety of existing age-related biomarkers, DNA methylation has been found as one of the most promising biomarkers for age prediction. DNA methylation refers to the addition of a methyl group to C at the 5th position of cytosine to promote or inhibit gene expression (Dirk, 2015; Nwanaji-Enwerem et al., 2018). In recent years, DNA methylation age (m-age) biomarkers has been able to accurately estimate the age of any tissue throughout life (Horvath, 2013). In addition, m-age biomarkers are valuable tools for evaluating tumor process, which can be a predictor of human health (Perna et al., 2016; Dhingra et al., 2018). However, the differences between m-age and chronological age of different tissues in tumor and the biological processes involved remained to be studied.

There are many studies on the age of individuals using DNA methylation level (Horvath, 2013; Galamb et al., 2016; Dhingra et al., 2018). In 2013, it was proposed for the first time to construct human multi-tissue age predictor based on DNA methylation to measure the degree of aging in human (Horvath, 2013). Subsequently, Weidnei et al. constructed age predictors using three CpG sites as characteristics in 575 healthy samples (Weidner et al., 2014; Nwanaji-Enwerem et al., 2018). Zbiec-piekarska et al. used 5 CpG loci as characteristics in 420 healthy samples to predict age (Spólnicka et al., 2018). Due to tissue specificity of methylation level, Stubbs et al. proposed to construct a multi-tissue age predictor based on methylation level in biological mice model (Stubbs et al., 2017). In addition, DNA methylation is maintained throughout life course in multiple tissues, linking many known early life factor to cancer risk (Kristina et al., 2015). The changes in DNA methylation with age occur at regulatory regions and contribute to tumor development (Johnson et al., 2017). Methylation of SLFN11 is a biomarker for poor prognosis in colorectal cancer and methylations of SLIT1, SLIT2, and SLIT3 are abnormal in gastric cancer. F2RL3 methylation is recently identified as a biomarker closely reflecting both current and past smoking exposure, causing lung cancer (Yan et al., 2015; Kim et al., 2016; He et al., 2017). However, the difference between the m-age for cancer and healthy samples is still unknown. In addition, the effect of age-related methylation characteristics on the survival risk of diverse cancer needs to be further confirmed.

In the present study, we found that DNA methylation levels were correlated with chronological age, and age-associated methylation sites were identified in human cancers. A comprehensive age prediction model named CancerClock was constructed using these age-associated methylation sites. Further, the characteristics of these age-associated methylation sites were described from three main aspects, including GO functions, biological phenotype, and the genomic location of the feature. Moreover, the m-age of common types of human cancers was described. The differences between predicted m-age and chronological age were also identified using CancerClock. For each cancer type, we detected a series of methylation sites, which could influence m-age compared to healthy samples. Finally, these differences were analyzed using cox regression model, and their effect on tumor survival risk was analyzed. In summary, the findings of this study would provide assistance to depict the age clock of cancers, characterize the m-age of cancer samples, and clinically evaluate the cancer progression.

METHODS

Clinical and Methylation Profile of Cancers

The methylation profile of IlluminaMethylation450 chip for 33 cancer types and their matched normal tissues were obtained from the cancer genome atlas (TCGA) data portal (TCGA Release 14.0, https://portal.gdc.cancer.gov/) (Hutter and Zenklusen, 2018), and each sample contained 485,512 methylation sites. The values in the methylation profile represent methylation degree of the sites ranging from 0 to 1, which is calculated as the ratio of methylated signal. The sample age ranged from 14 to 89 years. Totally, 8,692 samples including 7,988 tumor samples and 704 normal tissue samples were extracted for the present study (**Supplementary Table S1**). The corresponding clinical information of the samples (including age, gender, survival prognosis, cancer stage, and type of tumor samples) was also described.

Methylation Data Processing

First, we calculated the mean, standard deviation, and maximum and minimum values of age for each cancer type sample without clinical information were excluded. Methylation sites with NA values >10% in samples were removed. Then, we imputed the remaining NA values using 10-nearest neighbor method with the function knnImputation in DMwR package by R.

Abbreviations: m-age, DNA methylation age; TCGA, The cancer genome atlas; MSE, Mean-squared error; GO, Gene Ontology; EWAS, epigenome-wide association study software; BLCA, Bladder urothelial carcinoma; BRCA, Breast invasive carcinoma; COAD, Colon adenocarcinoma; ESCA, Esophageal carcinoma; HNSC, Head and neck squamous cell carcinoma; KICH, Kidney chromophobe; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; PRAD, Prostate adenocarcinoma; READ, Rectum adenocarcinoma; STAD, Stomach adenocarcinoma; THCA, Thyroid carcinoma; UCEC, Uterine corpus endometrial carcinoma; CHOL, Cholangiocarcinoma; OV, Ovarian serous cystadenocarcinoma; GBM, Glioblastoma multiforme; PAAD, Pancreatic adenocarcinoma; CESC, Cervical squamous cell carcinoma; PCPG, Pheochromocytoma and paraganglioma; SARC, Sarcoma; THYM, Thymoma; SKCM, Skin cutaneous melanoma; ACC, Adrenocortical carcinoma; CCSK, Clear cell sarcoma of the kidney; DLBC, Lymphoid neoplasm diffuse large B-cell lymphoma; MESO, Mesothelioma; LGG, Brain lower grade glioma; TGCT, Testicular germ cell tumors; UCS, Uterine carcinosarcoma; UVM, Uveal melanoma.

DNA methylation profiles of normal and tumor samples were annotated using CpG probe annotation file from GENCODE (Release 29, https://www.gencodegenes.org) (Frankish et al., 2019). All processes were performed by R software (R 3.3.3).

The Segmentation and Statistics of Age for All Cancer and Normal Samples

Human age in this study was divided into four sections according to the novel age subsection proposed by the United Nations and World Health Organization (WHO) after 1994: (Eeghen et al., 2015) Young people: under 44 years old (Taber et al., 2016), middle-aged people: between 45 and 59 years old (Mcguire et al., 2015), older-young people: between 60 and 74 years old, and (Palmer et al., 2018) old people: over 75 years old (**Supplementary Table S2**). The chronological ages of samples were obtained from clinical information and were divided using above age segmentations rules (**Supplementary Table S3**).

Identification of Age-Associated Methylation Sites

To identify age-associated methylation sites, Spearman correlation coefficients (SCCs) were calculated based on age and methylation levels based on all the normal samples (Stubbs et al., 2017). Then, a series of methylation sites were obtained with $P < 0.05$ and $|SCC|>0.3$. These methylation sites were proved to be significantly correlated with age, and can be the candidate features for constructing the model.

Construction of Age Predictor Known as CancerClock Based on Methylation Level in Normal Samples

To predict human age, we performed LASSO regression model in the generalized linear mode, which through some regression coefficients was strictly set to 0 and we obtained a model with good performance and strong explanatory power (Hepp et al., 2016). Firstly, we obtained a set of M*N methylation matrix X, in which the value in the matrix is the methylation level of the methylation site M in N samples. Each row of X represented the 1,631 age-associated sites, and the column represented the normal samples. Then, we normalized the methylation values as follows:

$$\sum_{i=1}^{m} x_{ij} = 0, \sum_{i=1}^{m} x_{ij}^2 = 1$$

From the formulae, i = 1,2,3…m, j = 1,2,3…n, m represents the 1,631 age-associated sites, n represents the sample size, and x_{ij} represents the methylation level.

Next, we assumed a set of linear regression models:

$$mAge = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots + \beta_m x_m$$

mAge is the model prediction age, vector β is the model coefficient, and vector x is the methylation expression level.

RSSs (residual sum of squares) with the least square estimation:

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

I ($n = 1, 2, 3 \ldots n$) represents the sample size, y_i describes the actual age of sample i, and x_{ij} represents the methylation level. The smaller the RSSs, the better the fitting effect.

LASSO-minimality:

$$J(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \gamma \sum_{j=1}^{p} |\beta_j| = RSS + \gamma \sum_{j=1}^{p} |\beta_j|$$

N means the total number of samples. $\gamma \sum_{j=1}^{p} |\beta_j|$ represents compressed penalty term. γ means adjustment, which controls the influence of RSS and compressed penalty term on regression coefficient. β_0 is excluded from the compressed penalty term, which represents the mean of response variables when $x_j = 0$. Next, coordinate axis descent method was used to solve the model. The initial β values were all 0. When the βvalue was lower than the threshold, we set the two adjacent iterations; the vector β_k would be regarded as the final feature of our model.

The CancerClock model was constructed based on GLMNET package in R 3.3.3 (Friedman et al., 2010). We evaluated the stability and accuracy of the model by means of 10-fold cross-validation. All the healthy samples were randomly divided into 10 subsets, 9 of which were selected as the training set. The test set was used in the verification of the model.

Functional Analysis and Phenotypic Traits for Characteristics in CancerClock

With the Enrichr tool online web server (http://amp.pharm.mssm.edu/Enrichr) using default parameters, functional enrichment was performed for genes across core clusters (Kuleshov et al., 2016), and we obtained enriched GO terms ($P < 0.01$, FDR < 0.05). The degree of methylation at chromosomal sites affected post-translational modification of proteins, which in turn contributed to the expression of biological traits. In this study, EWAS atlas, which is online analysis software, was used in describing the selected 282 loci with biological traits. EWAS Atlas (http://bigd.big.ac.cn/ewas) (Li et al., 2019) is a database of epigenetic traits, which include the correlation of information between nearly 330,000 methylation sites and 305 different biological traits through the integration and analysis of existing literatures.

Characteristics of Genome Location Regions for Age-Related Methylation Sties in CancerClock

To explore the distribution of CancerClock features in genome, we got the annotation information of the genome region from

GENCODE database (Release 29, https://www.gencodegenes.org) and we mapped the methylation features on the genome using BEDTools (V2.29.0, https://bedtools.readthedocs.io/en/latest/content/overview.html). In addition, we defined the promoter region as the upstream and downstream 1.5 kb of TSS (transcription start site); thus, the regions included promoter, exon, 5′-UTR, 3′-UTR, and TSS.

Prediction of m-Age for Tumor Types

To predict human methylation age, we used 282 age-related methylation characteristics and 704 healthy samples to construct age prediction model. To investigate further whether tumor causes changes in methylation age, we used the CancerClock prediction model to predict the 33 cancer types in TCGA based on methylation level. For cancer samples, their methylation age and chronological age showed some difference. Among the age prediction of cancer samples, we defined the error value as the absolute age difference of the quarter-point after the age prediction rank, and we showed the difference between the biology age and methylation age for different cancer types. Thus, some age-influenced cancer samples, in which methylation age was significantly different from chronological age, were obtained.

Differential Analysis of Methylation Sites Between Normal and "Age-Related Cancer Samples"

To determine which age-related methylation sites were changed in tumor samples, we conducted *t*-test for the methylation sites in healthy samples and age-influenced cancer samples, and we obtained significantly differential methylation sites (FDR < 0.01, Fold Change value>2 or <0.5). These cancer samples were defined as "age-related cancer samples" and they have some differential methylation sites.

The Survival Analysis for Differential Methylation Sites

An integrated pipeline was constructed to explore the associations between these age-related differential methylation sites and cancer survival (Zhou et al., 2019). First, we divided the samples into two groups including hyper-methylated and hypo-methylated sample groups according to the median methylation level. Next, the age-related differential methylation sites of the 16 cancer types were selected to establish a cox risk regression model for each tumor type, which was selected according to $P < 0.05$ (Gellar et al., 2015). The covariance matrix was composed of the cancer stage, gender, age, and methylation level. Next, the risk score for each cancer patient was calculated according to the linear combination of the methylation values weighted by the coefficient from multivariate Cox regression analysis. The median risk score was used as the cut-off point to divide the cancer patients into high and low risk groups. Finally, Kaplan-Meier survival analysis was conducted for the two groups, and statistical significance was assessed using the log-rank test. The survival results were considered significant when $P < 0.05$. All analyses were performed within the R 3.3.3 framework.

RESULTS

Correlation Between the Levels of Some DNA Methylation Sites in Human Normal Tissues and Chronological Age

Chronological age is referred to the real age of an individual, which is recorded by researchers. To study age-associated methylation levels in human over a wide range of ages and tissues, we collected 704 normal tissue samples from 14 to 89 years (**Figure 1A**). We found that 145,523 DNA methylation sites across all tissues were correlated with age (a multiple testing corrected $p < 0.05$) (**Figure 1B**). Further, we screened the correlations and we obtained 1,631 methylation sites that were significantly correlated with age, which were considered to be significantly correlated with chronological age ($|cor| > 0.3$). Overall, we identified up to 1,609 positive correlations, accounting for 98.7% of the total number of methylation sites, which suggested that in most samples, the higher the level of age-related methylation, the older the organism in the sample (**Figure 1C**). We classified tissue samples follow tissues to verify tissue specificity of the 1,631 age-related sites stated above. We performed correlation analysis with chronological age of each tissue, and we compared them with all-tissue correlation scores. The correlations of the most specific tissues were higher than the overall correlation level. This indicated that the selected sites were not only significantly correlated with age in all samples, but also more highly correlated with chronological age of samples in a single tissue (**Figure 1D**).Therefore, we demonstrated that these 1,631 methylation sites were related to the chronological age of individuals not only in specific tissues but also in all samples.

Feature Selection and Model Construction of CancerClock

CancerClock model was constructed to predict the m-age in human normal samples based on adjacent normal tissue for each cancer (**Figure 2A**, see Methods). The LASSO linear regression model was used to screen the 1,631 age-related methylation sites in the training set using the GLMNET package of R. Finally, 282 sites were selected from 1,631 methylation site as model features (**Figure 2B**, **Supplementary Table S4**). According to the least Mean-Squared Error (MSE, reflect the degree of difference between the estimator and true value, the smaller the MSE, the better the model fit) of LASSO linear regression model, we found that the value β_0 of the model was 34.63 when the adjustment parameter was 0.1419941 (**Figure 2C**). Among the 282 methylation characteristics in the construction of model, the levels of cg08461576, cg05923914, cg27641628, cg13221458, cg05632420, and cg07103722 were significantly and negatively correlated with sample age, and the linear model coefficients of the six sites above were negative as well (**Figure 2D**). We described the methylation level for these 282 methylation characteristics and we promoted them to effectively describe the chronological age of all normal samples through the heatmap (**Figures 2A,D**). The results showed that it was difficult to characterize the methylation age of samples by single methylation sites, and the methylation clusters might produce a better result.

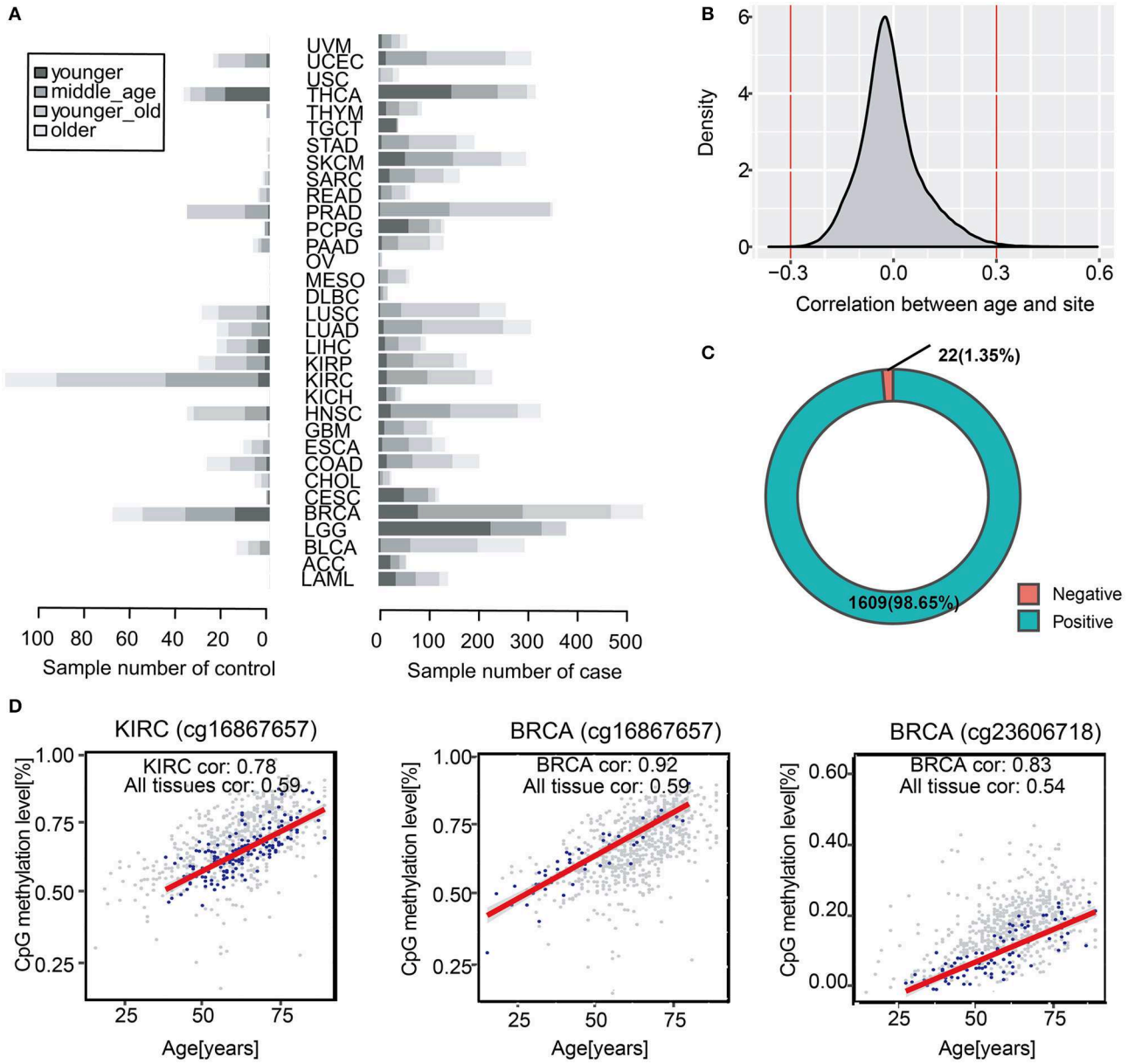


FIGURE 1 | Identification of chronological age-related DNA methylation sites in human normal tissues. **(A)** The barplot shows sample distribution of four age segments in diverse cancer types. The lighter color represents larger age of samples. **(B)** Distribution of correlation scores between all methylation sites and chronological age. **(C)** Spearman correlation between age and methylation ($P < 0.05$, $|cor| > 0.3$). The pie chart shows the proportion of 1,631 positive and negative methylation sites correlated with chronological age. **(D)** The scatter diagram shows the correlation between methylation sites (cg16867657, cg23606718) and chronological age in KIRC and BRCA. The overall background is the methylation sites of all samples (gray color), and the blue color represents the correlation between chronological age and methylation age in a certain type of samples such as BRCA and KIRC.

To evaluate the accuracy of the model, we used the training set and the test set to verify the model, separately. The results showed that the model responded well to the training set samples, and the correlation between chronological age and m-age was as high as 0.939 ($P < 2.38e{-}295$) (**Figure 2E**). Meanwhile, for test set samples, the correlation between chronological age and m-age predicted by the model was up to 0.843 ($P < 5.96e{-}20$) (**Figure 2E**). The average accuracy of the model was 81%. To determine the difference between chronological age and predicted m-age in normal samples, we ordered the absolute value of the difference between the chronological age and predicted m-age in descending order, and we selected the quartile as the error value. The results showed that the age error value was 3 years, which indicated that the prediction error of over 75% of the sample age was controlled within 3 years, which is far

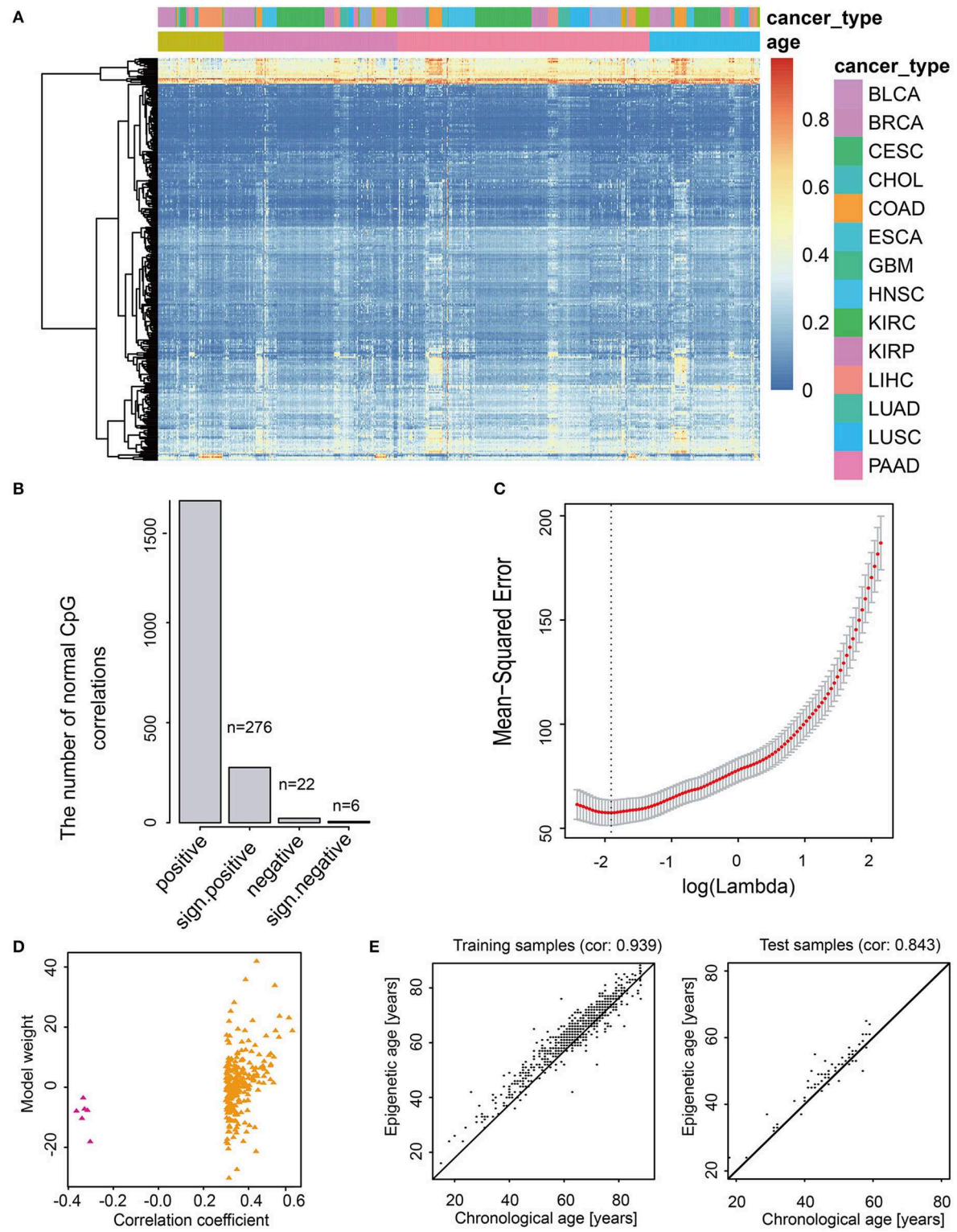


FIGURE 2 | Feature selection and model construction of CancerClock. **(A)** The heatmap shows the 282 age-related methylation sites extracted by the model in all cancer samples. The red and blue represent high and low levels of methylation sites. **(B)** The barplot shows the number of 282 positive and negative age-related methylation sites. **(C)** The selection of thresholds in LASSO regression model. The line represents the coefficient values, and the minimum mean-square error corresponding to log(Lambda) is−1.95197. **(D)** Scatter diagram shows the correlations between correlation coefficient and model weight of CancerClock. **(E)** The correlation between chronological age and predicted m-age in the training set and the correlation between chronological age and predicted m-age in the test set.

lower than the age error predicted by previous researchers through telomeres.

Biological Processes and Phenotypic Traits of Age-Related Methylation Sites for CancerClock

Gene Ontology (GO) analysis of the genes for the 282 model characteristics revealed that these genes were associated with some GO terms such as "lysine catabolic process" (GO:0006554) and "lysine metabolic process" (GO:0006553) (**Figure 3A**). A network consisting of GO terms and genes was constructed (**Figure 3B**). These results indicated that age-related methylation sites could alter many important biology processes (**Supplementary Table S5**). The online analysis software of EWAS Atlas was used to analyze the biological characteristics enrichment of 282 selected features, and it was found that 128 of them were significantly enriched in age traits [$-\log10\ (p) > 318$], which was manifested as significant methylation quantitative trait loci (meQTLs) related to age. Meanwhile, 128 sites were enriched in age ontology entry (GO:0007568). In addition, we found that most methylation sites were enriched in human acute leukemia [$-\log10\ (p) > 50$] (**Figure 3C**, **Supplementary Table S6**). When we mapped these characteristics on the corresponding gene positions, we found that most of the loci corresponded to a single gene, while few loci corresponded to multiple genes (**Figure 3D**). To describe the distribution of these features of CancerClock in genomic regions, we determined the position of these features in the genome. We found 37 of the 282 characteristics located in the promoter region of genes (**Figure 3E**).

Predicted m-Age and Chronological Age of Cancer Patients

To depict the difference between m-age and chronological age in cancer samples, we utilized the CancerClock based on the methylation levels of all the 33 tumor types. We found that the m-age was generally different from the chronological age, but the degree of the difference was closely related to tumor types. Here, we ranked the absolute age differences between chronological age and we predicted the m-age and selected the quartile values as the age difference scores of tumors, separately (**Figure 4A**, **Supplementary Table S7**). A number of tumors showed higher age difference score. Uterine carcinosarcoma (USC) had the highest age difference score, followed by ovarian serous cystadenoma carcinoma (OV) and uterine corpus endometrial carcinoma (UCEC). We found that the top three diseases with the largest age difference scores were all women tumors, suggesting that their chronological age is related to the pathological changes of the reproductive organ, and that female diseases affect chronological age to some extent. In addition, the distribution of m-age compared to the chronological age was described, separately (**Figure 4B**). We found that the m-age of different disease types was specific to their chronological age. For example, USC showed that all the 57 cancer samples were younger than the chronological age, and 96% of 54 testicular tumors showed age decline in men. According to our results, in the 121 and 184 samples with thymoma and ganglioma, 95% of the m-age was below the chronological age, respectively. In addition, in the 247 and 80 samples with renal cell carcinoma (RCC) and melanoma, 94% had younger m-age than chronological age, respectively.

Differential Levels of Some Methylation Sites Between m-Age and Chronological Age in Age-Related Cancer Samples

To explore further the difference between m-age and chronological age, we selected 25% age-related samples with the largest difference, and then performed differential analysis with normal samples in each cancer type (**Supplementary Table S8**). In the age-related samples of cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), there were 77 methylation sites levels that were, respectively, different (FDR < 0.01, $|\log\ (FC)|>1$) (**Supplementary Table S9**). Moreover, there were 38 methylation sites with significant changes in breast invasive carcinoma (BRCA) and 22 methylation sites with significant changes in kidney renal clear cell carcinoma (KIRC). The differential methylation sites of some cancers showed hyper-methylation in tumor samples as follows: cervical squamous cell carcinoma (CESC) and bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), sarcoma (SARC), and uterine corpus endometrial carcinoma (UCEC). It is suggested that hyper-methylation of these age-related differences may lead to changes in m-age and cancer development.

Association Between the Differential Methylation Sites in Age-Related Samples and Survival

Predicting survival state of cancer patients was critical and challenging (Zhou et al., 2017; Bao et al., 2019). To evaluate the influence of the methylation level on patient survival, cox risk regression model was performed for the differential methylation sites in age-related samples of each tumor type (**Supplementary Table S10**). The patients were divided into two groups according to median risk score. In most cancer types, these differential methylation sites in age-related samples were associated with survival (**Figure 5**). The higher the risk scores of BRCA, COAD, esophageal carcinoma (ESCA), liver hepatocellular carcinoma (LIHC), and stomach adenocarcinoma (STAD), the lower the survival of the patients. The results indicated that differential methylation sites of age-related samples maybe could be an effective prognostic biomarker for cancers.

DISCUSSION

DNA methylation levels were closely associated with age and related to the occurrence of tumors (Wang et al., 2016). In this study, we explored the correlations between DNA methylation level and age in normal tissues. Importantly, we established CancerClock, a predictive model of age based on

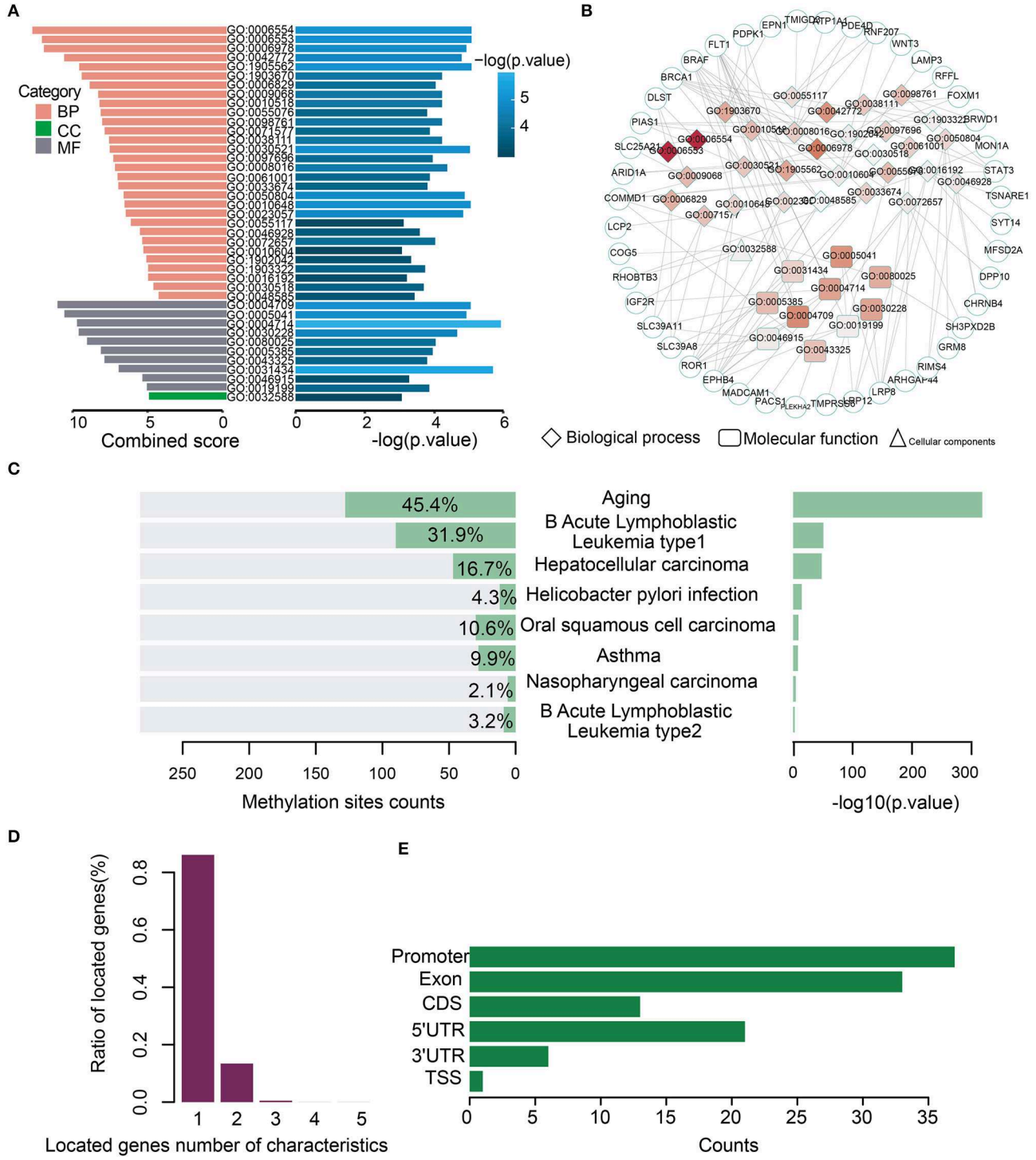


FIGURE 3 | Biological processes and phenotypic traits of age-related methylation sites for CancerClock. **(A)** The barplot shows the combined score and –log (*p*-values) for enrichment GO terms of 282 extracted methylation sites. Pink, green, and gray colors represent biological process (BP), molecular function (MF), and cellular component (CC), respectively. **(B)** Network shows the interactions between GO terms and genes. The color of node in network indicates enrichment strength, and the three different shapes represent different biological types of GO terms. The circle represents the genes. **(C)** The bar chart shows the enrichment counts and significance *P*-values of each trait from EWAS atlas analysis. **(D)** The corresponding relationship between methylation site and the genes in which it is located. The relationship is usually one-to-one or one-to-many. **(E)** The distribution of 282 features of CancerClock model in genome position.

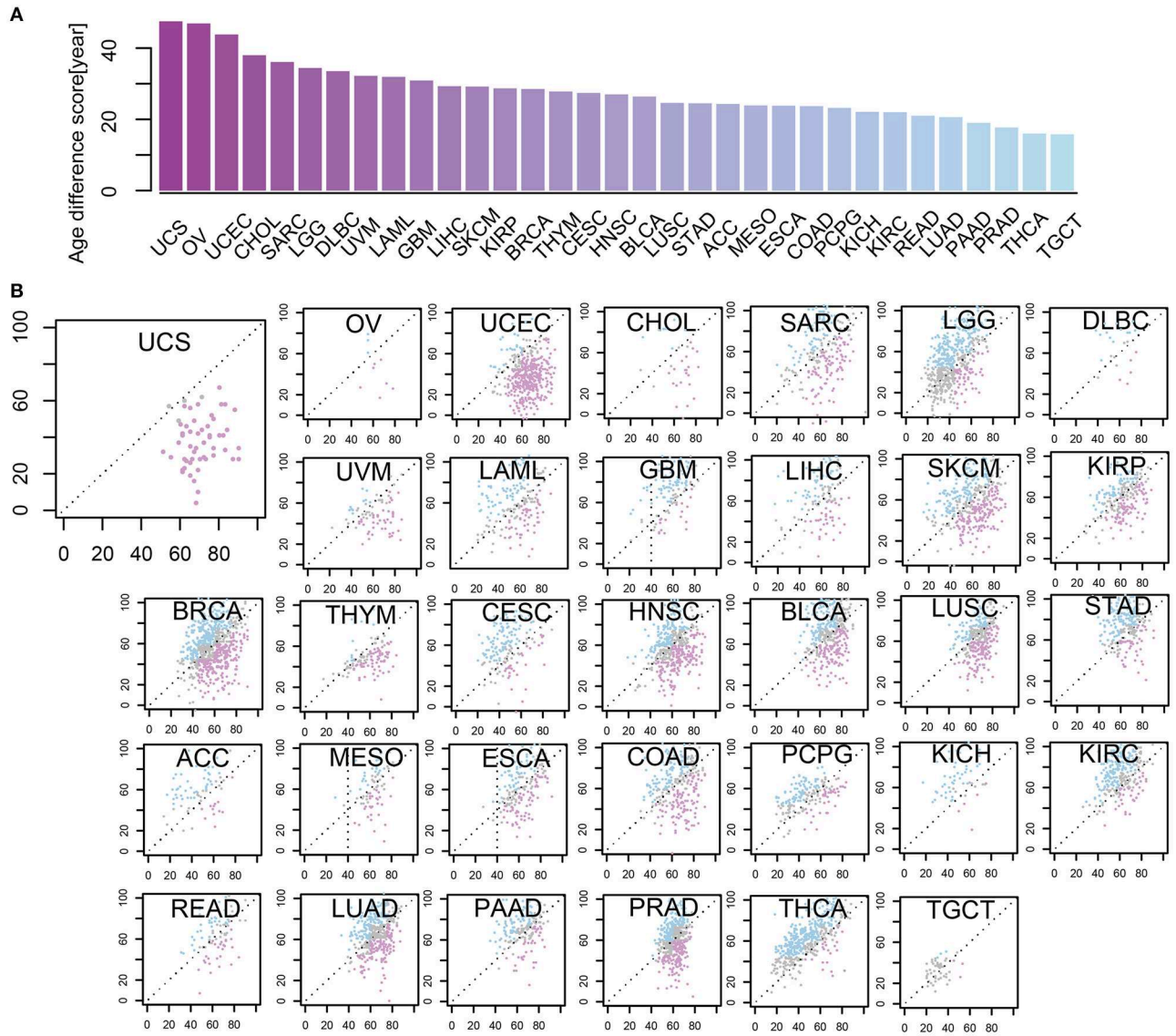


FIGURE 4 | The levels of some methylation sites were differential in age-related cancer samples between m-age and chronological age. **(A)** Age different scores between chronological age and predicted m-age among 33 cancer types. **(B)** The relationship between chronological age and predicted m-age for all the 33 disease types. Compared to the chronological age group, the pink color indicates that the predicted m-age group was down-regulated, the blue color indicates that the predicted m-age group was up-regulated, and the gray color indicates that the predicted m-age group remained unchanged.

DNA methylation level to depict the age clock in tumors. This predicted the age based on methylations of 282 sites from different tumor samples and it allowed us to access the m-age among methylation datasets. In addition, we found that the model was affected by a variety of biological processes, which may indicate the molecular influence of methylation on age. Age difference score in tumor samples showed the extent to which age affected different tumors. Meanwhile, methylation sites affecting the m-age of these tumors were identified. Through weighted survival analysis of cancer samples, we finally determined the impact of these age-causing sites on tumor survival.

CancerClock model is a multi-tissue age prediction model based on the methylation level of diverse cancer types in TCGA. It adopted a similar way as Horvath clock in building the age prediction model of normal samples (Horvath, 2013). More importantly, we applied this model to cancer samples and we depicted the difference between m-age and chronological age. These differences between m-age and chronological age may provide assistance to understand cancer development. Previous study also showed that epigenetic age acceleration is associated with colorectal cancer molecular characteristics and can be a significant predictor of overall survival, as well as age and tumor stage (Zheng et al., 2019). DNA methylation-based measures of biological age may be an important predictors of breast cancer risk (Kresovich et al., 2019). In the present study, we depicted this

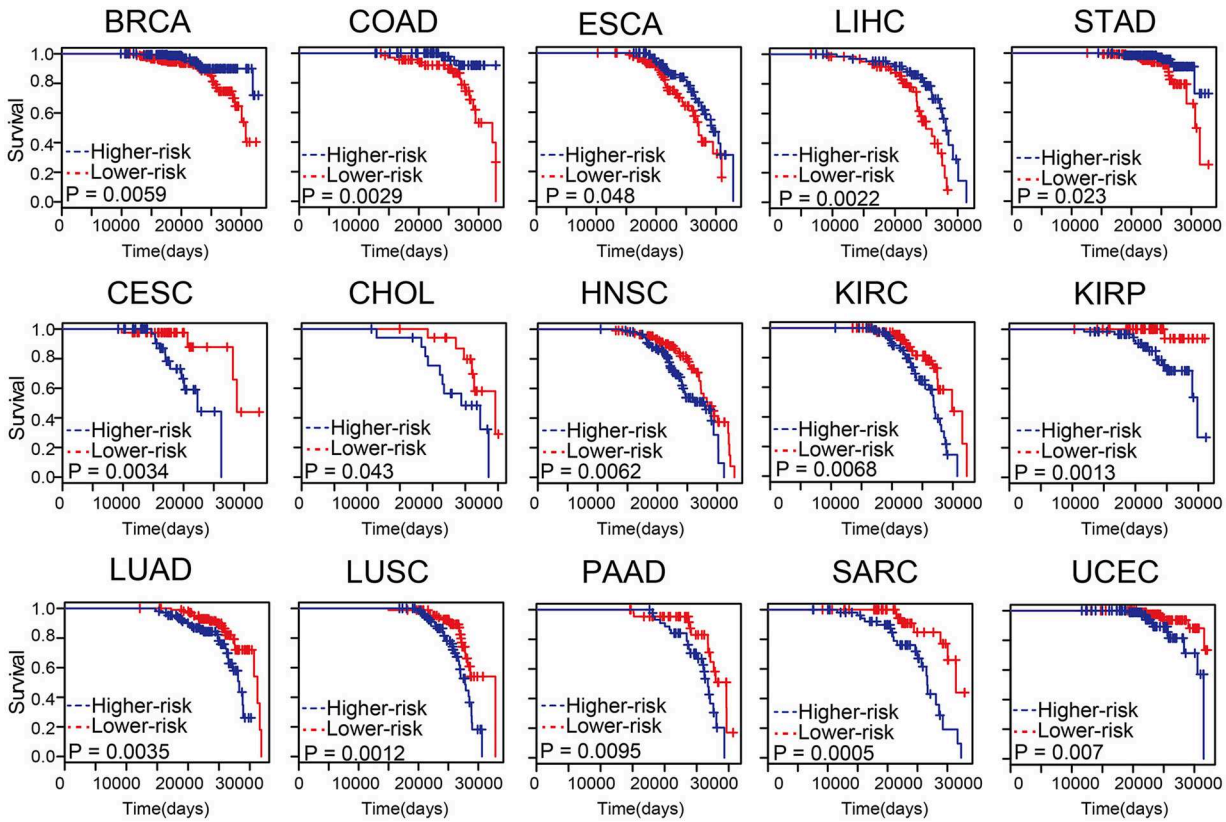


FIGURE 5 | The differential methylation sites in age-related samples were associated with survival. Kaplan-Meier survival analysis of two groups of patients with high- (blue line) and low- score (red line) groups. Survival days are shown along the X-axis. Overall survival rates are shown along the Y-axis.

phenomenon in multiple cancer types and we tried to explain the mechanism by which methylation levels contribute to m-age and cause cancers.

Many studies used expression of gene and lncRNA to predict cancer development and prognosis (Zhou et al., 2015b, 2018; Sun et al., 2019). In present work, m-age was considered as cancer biomarker for development and prognosis. CancerClock could be applied in predicting biological age for normal samples. The differences between biological and m-age were also could be identified. These differences could be used to explore the roles of methylation in cancer development and prognosis. In future work, more samples and experiments should be used to validate the present work.

In summary, the present study suggested that some methylation sites were associated with chronological age. Comprehensive age predicator CancerClock could predict m-age for normal samples and could find the differences between m-age and chronological age in age-related cancer samples. We further discovered the differential methylation sites between age-related cancer samples and normal samples. These differential methylation sites were associated with survival in cancers. In addition, the present study suggested that DNA methylation-based measures of chronological age might be important predictors of cancer risk.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

SN, LW, and HC conceived the project. TZ, YG, JW, and XL acquired the data. SS, YW, and SG analyzed the data. HZ, HL, and DS constructed the data resource. TZ, SN, and YG wrote the manuscript. All authors read and accepted the final version of the manuscript.

FUNDING

This work was supported by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province [UNPYSCT-2017071], Postdoctoral Scientific Research Developmental Fund, and Yu Weihan Outstanding Youth Training Fund of Harbin Medical University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00388/full#supplementary-material

REFERENCES

Aubert, G., and Lansdorp, P. M. (2008). Telomeres and aging. *Physiol. Rev.* 88, 557–579. doi: 10.1152/physrev.00026.2007

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief. Bioinformatics*. bbz118. doi: 10.1093/bib/bbz118

Dhingra, R., Nwanaji-Enwerem, J. C., Samet, M., and Ward-Caviness, C. K. (2018). DNA methylation age—environmental influences, health impacts, and its role in environmental epidemiology. *Curr. Environ. Health Rep*. 5, 317–327. doi: 10.1007/s40572-018-0203-2

Dirk, S. (2015). Function and information content of DNA methylation. *Nature* 517, 321–326. doi: 10.1038/nature14192

Eeghen, E. E., Van Bakker, S. D., and Aart, V. B., Loffeld, R. J. L. F. (2015). Impact of age and comorbidity on survival in colorectal cancer. *J. Gastrointest. Oncol.* 6, 605–612. doi: 10.3978/j.issn.2078-6891.2015.070

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D73. doi: 10.1093/nar/gky955

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

Galamb, O., Kalmár, A., Barták, B. K., Patai, Á. V., Leiszter, K., Péterfia, B., et al. (2016). Aging related methylation influences the gene expression of key control genes in colorectal cancer and adenoma. *World J. Gastroenterol.* 22, 10325–10340. doi: 10.3748/wjg.v22.i47.10325

Gellar, J. E., Colantuoni, E., Needham, D. M., and Crainiceanu, C. M. (2015). Cox regression models with functional covariates for survival data. *Stat. Model.* 15, 256–278. doi: 10.1177/1471082X14565526

He, T., Zhang, M., Zheng, R., Zheng, S., Linghu, E., Herman, J. G., et al. (2017). Methylation of SLFN11 is a marker of poor prognosis and cisplatin resistance in colorectal cancer. *Epigenomics* 9, 849–862. doi: 10.2217/epi-2017-0019

Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., and Mayr, A. (2016). Approaches to regularized regression - a comparison between gradient boosting and the lasso. *Methods Inf. Med.* 55, 422–430. doi: 10.3414/ME16-01-0033

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* 14:R115. doi: 10.1186/gb-2013-14-10-r115

Hutter, C., and Zenklusen, J. C. (2018). The cancer genome atlas: creating lasting value beyond its data. *Cell* 173:283. doi: 10.1016/j.cell.2018.03.042

Johnson, K. C., Houseman, E. A., King, J. E., and Christensen, B. C. (2017). Normal breast tissue DNA methylation differences at regulatory elements are associated with the cancer risk factor age. *Breast Cancer Res.* 19:81. doi: 10.1186/s13058-017-0873-y

Kim, M., Kim, J. H., Baek, S. J., Kim, S. Y., and Kim, Y. S. (2016). Specific expression and methylation of SLIT1, SLIT2, SLIT3, and miR-218 in gastric cancer subtypes. *Int. J. Oncol.* 48, 2497–2507. doi: 10.3892/ijo.2016.3473

Kresovich, J. K., Xu, Z., O'Brien, K. M., Weinberg, C. R., Sandler, D. P., and Taylor, J. A. (2019). Methylation-based biological age and breast cancer risk. *J. Natl. Cancer Inst.* 111, 1051–1058. doi: 10.1093/jnci/djz020

Kristina, H., Gwen, H., Paula, S., Louise, S., Horgan, G. W., Elizabeth, S., et al. (2015). Breast cancer risk and imprinting methylation in blood. *Clin. Epigenetics* 7, 1–9. doi: 10.1186/s13148-015-0125-x

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–W7. doi: 10.1093/nar/gkw377

Li, M., Zou, D., Li, Z., Gao, R., Sang, J., Zhang, Y., et al. (2019). EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.* 47, D983–D8. doi: 10.1093/nar/gky1027

Mcguire, A., Brown, J. A. L., Malone, C., Mclaughlin, R., and Kerin, M. J. (2015). Effects of age on the detection and management of breast cancer. *Cancers* 7, 908–929. doi: 10.3390/cancers7020815

Nwanaji-Enwerem, J. C., Weisskopf, M. G., and Baccarelli, A. A. (2018). Multi-tissue DNA methylation age: molecular relationships and perspectives for advancing biomarker utility. *Ageing Res. Rev.* 45, 15–23. doi: 10.1016/j.arr.2018.04.005

Palmer, S., Albergante, L., Blackburn, C. C., and Newman, T. J. (2018). Thymic involution and rising disease incidence with age. *Proc. Natl. Acad. Sci. U.S.A.* 115, 1883–1888. doi: 10.1073/pnas.1714478115

Perna, L., Zhang, Y., Mons, U., Holleczek, B., Saum, K. U., and Brenner, H. (2016). Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin. Epigenetics* 8, 1–7. doi: 10.1186/s13148-016-0228-z

Spólnicka, M., Pośpiech, E., Pepłonska, B., Zbieć-Piekarska, R., Makowska, Z, Pieta, A., et al. (2018). DNA methylation in ELOVL2 and C1orf132 correctly predicted chronological age of individuals from three disease groups. *Int. J. Legal Med.* 132, 1–11. doi: 10.1007/s00414-017-1636-0

Stubbs, T. M., Bonder, M. J., Stark, A. K., Krueger, F., Bi, A. C. T., Meyenn, F. V., et al. (2017). Multi-tissue DNA methylation age predictor in mouse. *Genome Biol.* 18:68. doi: 10.1186/s13059-017-1203-5

Sun, J., Zhao, H., Lin, S., Bao, S., Zhang, Y., Su, J., et al. (2019). Integrative analysis from multi-centre studies identifies a function-derived personalized multi-gene signature of outcome in colorectal cancer. *J. Cell. Mol. Med.* 23, 5270–5281. doi: 10.1111/jcmm.14403

Taber, J. M., Klein, W. M. P., Suls, J. M., and Ferrer, R. A. (2016). Lay awareness of the relationship between age and cancer risk. *Ann. Behav. Med.* 51, 1–12. doi: 10.1007/s12160-016-9845-1

Urbano, A., Smith, J., Weeks, R. J., and Chatterjee, A. (2019). Gene-specific targeting of DNA methylation in the mammalian genome. *Cancers* 11:E1515. doi: 10.3390/cancers11101515

Vilchez, D., Saez, I., and Dillin, A. (2014). The role of protein clearance mechanisms in organismal ageing and age-related diseases. *Nat. Commun.* 5:5659. PubMed PMID: 25482515. doi: 10.1038/ncomms6659

Wang, T., Liu, Q., Li, X., Wang, X., Li, J., Zhu, X., et al. (2013). RRBS-analyser: a comprehensive web server for reduced representation bisulfite sequencing data analysis. *Hum. Mutat.* 34, 1606–1610. doi: 10.1002/humu.22444

Wang, Y., Zhang, J., Xiao, X., Liu, H., Wang, F., Li, S., et al. (2016). The identification of age-associated cancer markers by an integrative analysis of dynamic DNA methylation changes. *Sci. Rep.* 6:22722. doi: 10.1038/srep22722

Weidner, C. I., Lin, Q., Koch, C. M., Eisele, L., Beier, F., Ziegler, P., et al. (2014). Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* 15:R24. doi: 10.1186/gb-2014-15-2-r24

Yan, Z., Ben, S. T., José, O. E. M., Bernd, H., Rongxi, Y., Barbara, B., et al. (2015). F2RL3 methylation, lung cancer incidence and mortality. *Int. J. Cancer* 137, 1739–1748. doi: 10.1002/ijc.29537

Zhang, Y., Yang, L., Kucherlapati, M., Hadjipanayis, A., Pantazi, A., Bristow, C. A., et al. (2019). Global impact of somatic structural variation on the DNA methylome of human cancers. *Genome Biol.* 20:209. doi: 10.1186/s13059-019-1818-9

Zheng, C., Li, L., and Xu, R. (2019). Association of epigenetic clock with consensus molecular subtypes and overall survival of colorectal cancer. *Cancer Epidemiol. Biomarkers Prevent*. doi: 10.1158/1055-9965.EPI-19-0208

Zhou, M., Diao, Z., Yue, X., Chen, Y., Zhao, H., Cheng, L., et al. (2016). Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer. *Oncotarget* 7, 56383–56394. doi: 10.18632/oncotarget.10891

Zhou, M., Guo, M., He, D., Wang, X., Cui, Y., Yang, H., et al. (2015b). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J. Transl. Med.* 13:231. doi: 10.1186/s12967-015-0556-3

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2019). Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Brief. Bioinformatics* 20, 598–608. doi: 10.1093/bib/bby021

Zhou, M., Zhao, H., Wang, Z., Cheng, L., Yang, L., Shi, H., et al. (2015a). Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. *J. Exp. Clin. Cancer Res.* 34:102. doi: 10.1186/s13046-015-0219-5

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA

biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

Zinger, A., Cho, W. C., and Ben-Yehuda, A. (2017). Cancer and aging - the inflammatory connection. *Aging Dis.* 8, 611–27. doi: 10.14336/AD.2016.1230

Zubakov, D., Liu, F., Kokmeijer, I., Choi, Y., van Meurs, J. B., van IJcken, W. F., et al. (2016). Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length. *Forensic Sci. Int. Genet.* 24, 33–43. doi: 10.1016/j.fsigen.2016.05.014

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 © 2019 Zhu, Gao, Wang, Li, Shang, Wang, Guo, Zhou, Liu, Sun, Chen, Wang and Ning. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 10 December 2019
doi: 10.3389/fbioe.2019.00364

Whole-Genome Methylation Analysis of Phenotype Discordant Monozygotic Twins Reveals Novel Epigenetic Perturbation Contributing to the Pathogenesis of Adolescent Idiopathic Scoliosis

Gang Liu[1,2†], Lianlei Wang[1,2†], Xinyu Wang[3†], Zihui Yan[1,2], Xinzhuang Yang[2,4], Mao Lin[1,2], Sen Liu[1,2], Yuzhi Zuo[1,2], Yuchen Niu[2,4], Sen Zhao[1,2], Yanxue Zhao[1,2], Jianguo Zhang[1,2,5], Jianxiong Shen[1,2,5], Yipeng Wang[1,2,5], Guixing Qiu[1,2,5], Zhihong Wu[2,4,5] and Nan Wu[1,2,5*] on behalf of the Deciphering Disorders Involving Scoliosis and Comorbidities (DISCO) Study*

[1] Department of Orthopedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China, [2] Beijing Key Laboratory for Genetic Research of Skeletal Deformity, Beijing, China, [3] Institute of Biomedical Big Data, Wenzhou Medical University, Wenzhou, China, [4] Department of Central Laboratory, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China, [5] Medical Research Center of Orthopedics, Chinese Academy of Medical Sciences, Beijing, China

Background: Adolescent idiopathic scoliosis (AIS) is a complex disease affecting a large number of teenagers, especially in female. This study reveals novel epigenetic perturbation to the pathogenesis of AIS.

Methods: A female monozygotic (MZ) twin pair discordant for AIS were examined for whole-exome sequencing and epigenome difference. Sets of differentially methylated regions (DMRs) were validated using MethylTarget™ method in 20 AIS female patients and 20 healthy female controls.

Results: Few exome difference but several potential DMRs were found between the MZ twins. We identified 313 hypermethylated DMRs and 397 hypomethylated DMRs, respectively. Most of them were enriched in the MAPK and PI3K-Akt signaling pathway, which may contribute to the discordance of AIS. Several DMRs related to scoliosis genes were tested, and the *NDN*: TSS-DMR (chr15:23932133-23932304, hg19) was confirmed in additional samples. The methylation level of this DMR was significantly higher in the AIS group than in the control group ($p = 0.04$).

Conclusions: We described the epigenome difference in an AIS female discordant MZ twin pair using Whole Genome Bisulfite Sequencing (WGBS). The *NDN*: TSS-DMR had higher methylation level in female AIS, which can help elucidate the potential etiology of AIS.

Keywords: adolescent idiopathic scoliosis (AIS), monozygotic twins, whole exome sequencing (WES), DNA methylation, whole genome bisulfite sequencing (WGBS)

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Enrique Medina-Acosta,
Universidade Estadual do Norte Fluminense Darcy Ribeiro, Brazil
Jie Zhang,
Tongji University, China

***Correspondence:**
Nan Wu
dr.wunan@pumch.cn
Zhihong Wu
orthoscience@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *19 August 2019*
Accepted: *12 November 2019*
Published: *10 December 2019*

Citation:
Liu G, Wang L, Wang X, Yan Z, Yang X, Lin M, Liu S, Zuo Y, Niu Y, Zhao S, Zhao Y, Zhang J, Shen J, Wang Y, Qiu G, Wu Z and Wu N (2019) Whole-Genome Methylation Analysis of Phenotype Discordant Monozygotic Twins Reveals Novel Epigenetic Perturbation Contributing to the Pathogenesis of Adolescent Idiopathic Scoliosis. Front. Bioeng. Biotechnol. 7:364. doi: 10.3389/fbioe.2019.00364

INTRODUCTION

Adolescent idiopathic scoliosis (AIS) is a 3-dimensional spinal deformity affecting 1–3% of children in the world (Yawn et al., 1999). Although studies to discover the underlying mechanisms of AIS had been conducted for many years, the etiology is still unclear. Previous studies showed that genetic factors from single nucleotide variant (SNV) to copy number variants (CNV), played a pivotal role in the development of AIS (Kesling and Reinker, 1997; Li et al., 2016; Gao et al., 2017). Based on genome-wide association study (GWAS) and other similar association studies, single nucleotide polymorphisms (SNPs) of several genes such as *GPR126* (Kou et al., 2013; Liu et al., 2018), *PAX1* (Sharma et al., 2015; Liu et al., 2019), *LBX1* (Takahashi et al., 2011; Liu et al., 2017), and *BNC2* (Ogura et al., 2015), were known to increase the risk of AIS. There was also evidence that rare mutations of some genes may be responsible for AIS, such as *FBN1* (Buchan et al., 2014a), *AKAP2* (Li et al., 2016) and *MAPK7* (Gao et al., 2017). Other studies illustrated the association between CNV and AIS (Buchan et al., 2014b). All of those genetic factors can only explain the etiology of 2–7.6% of AIS (Buchan et al., 2014a; Ikegawa, 2016).

The abnormal, or aberrant DNA methylation pattern is universally recognized as an important factor in diseases, especially complex diseases. Recent studies indicated that differential methylation of key genes or CpG site were related to AIS, such as site cg01374129 (Meng et al., 2018), *COMP* (Mao et al., 2018), and *PITX1* (Shi et al., 2018). But there were few types of research focusing on the methylation status of AIS on the whole-genome scale and no research of monozygotic twins discordant for AIS.

Monozygotic (MZ) twins are outstanding subjects to analyze epigenetic mechanisms. Theoretically, they share almost the same genotype, but the epigenome could be different, especially for discordant MZ twins who display phenotypically discordant on disease traits. Therefore discordant MZ twins are analyzed for deciphering complex diseases, such as congenital heart disease (Lyu et al., 2018), type 1 diabetes (Elboudwarej et al., 2016), and congenital renal agenesis (Jin et al., 2014).

In this study, we enrolled a MZ female twin pair discordant for AIS from Chinese Han population. In this twin pair, one was diagnosed with AIS, but her twin sister was healthy and had no spinal deformity. Their parents had no abnormality of the skeletal system or related family history. We compared their exome variants and genome methylation difference aiming to find the contributions to the pathogenesis of AIS.

MATERIALS AND METHODS

Patients and Materials

Monozygotic twin discordant for AIS was recruited from Peking Union Medical College Hospital (PUMCH). There was no consanguinity between the parents. The parents and monozygotic twin of the patient were healthy without abnormalities of the skeletal system. Twenty sporadic AIS female patients and 20 healthy female controls were enrolled as a replication cohort. Genomic DNA was extracted from their peripheral blood using DNeasy Blood & Tissue Kits (QIAGEN, Eastwin Scientific, Inc., Beijing, China) according to the manufacturer's instructions. This study was approved by the Ethical Review Board of Peking Union Medical College Hospital. Written informed consent was obtained from all the participants or their parents.

Whole Exome Sequencing (WES) and Analysis

WES was performed on DNA extracted from the peripheral blood. In brief, libraries were prepared from DNA samples and subjected to whole-exome capture using the VCRome SeqCap EZ Chice HGSC 96 Reactions capture reagent (Roche), followed by sequencing on an Illumina HiSeq 4000 platform with 150-bp pair-end reads mode. Whole-exome sequencing for all individuals in this family was performed and each generated fully reads with a mean depth of 107 (**Table S1**).

The variant-calling and annotation were performed by the in-house developed Peking Union Medical college hospital Pipeline (PUMP), same as in our previous studies (Wang et al., 2018a,b). In brief, single-nucleotide variants and internal duplications and/or deletions (aka indels) were called using the HaplotypeCaller of the Genome Analysis Toolkit (v3.4.0). Annotation of the *de novo*, compound heterozygotes, and recessively inherited variants was calculated with Gemini (v0.19.1) for *in silico* subtraction of parental variants from the proband's variants, with accounting for read number information extracted from BAM files. Computational prediction tools Gerp++ (Davydov et al., 2010), CADD (Kircher et al., 2014), SIFT (Vaser et al., 2016), Polyphen-2 (Adzhubei et al., 2010), and MutationTaster (Schwarz et al., 2014) were used to predict the conservation and pathogenicity of candidate variants. All variants were compared against Deciphering Disorders Involving Scoliosis & Comorbidities (DISCO) study in-house database and publicly available databases including the 1000 Genomes Project (http://www.internationalgenome.org/), the Exome variant server, NHLBI GO Exome Sequencing Project (ESP) (http://evs.gs.washington.edu/EVS/), and the Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/).

After data annotation, we first identified *de novo* mutations in the AIS patient taking her parents as a reference and then comparison was made between the MZ twins. For the mutations between the MZ twins including SNVs and short insertions/deletions (InDels), we classified them into four types (**Figure S1**). In the Equal type, both of the twins had the same mutations comparing to the Reference Genome. In the different type1, only the case had the mutation. In the different type2, the mutations only presented in the Control. In the different type3, both of the twins had different mutation compared to the Reference Genome. We hypothesized that the different type1 and type3 were potentially pathogenic.

Whole Genome Bisulfite Sequencing (WGBS)

Library Preparation and Quantification

A total amount of 5.2 micrograms genomic DNA spiked with 26 ng lambda DNA were fragmented by sonication to 200–300

bp with Covaris S220, followed by end repair and adenylation. Cytosine-methylated barcodes were ligated to sonicated DNA as per manufacturer's instructions. Then these DNA fragments were treated twice with bisulfite using EZ DNA Methylation-GoldTM Kit (Zymo Research). And the resulting single-strand DNA fragments were PCR amplificated using KAPA HiFi HotStart Uracil + ReadyMix (2X).

Library concentration was quantified by Qubit® 2.0 Fluorometer (Life Technologies, CA, USA) and quantitative PCR, and the insert size was checked on Agilent Bioanalyzer 2100 system.

Clustering and Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumia) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on Illumina Hiseq 2500 platform and 100 bp single-end reads were generated. Image analysis and base calling were performed with the standard Illumina pipeline, and finally, 100 bp paired-end reads were generated.

WGBS Data Analysis

Raw WGBS reads were mapped to the human reference genome (hg19, NCBI build 37.1) using BSMAP v2.90 (default parameters). Adaptor, low-quality and duplicated reads were automatically trimmed by BSMAP v2.90 with default thresholds. Single base methylation ratio was measured as the proportion of methylated Cs in all mapped reads from both strands. Only CpG sites with covered reads ≥4 reads were considered for further analysis.

Differentially methylated regions (DMRs) between the twin pair were identified using MethPipe 3.4.3 with significant differentially CpGs ≥ 5 and a minimal number of 10 CpGs that the DMR spans. DMRs were classified to "hypermethylation group" and "hypomethylation group" based on methylation level difference and these regions were ranked by absolute methylation difference between scoliosis and normal sample in the MZ twin (**Figure S2**). Microarray probes were designed for DMRs from the twin pair based on physical location, the methylation state of each interval was quantified by the mean of methylation beta values. Pearson correlation score was used to estimate the reliability of WGBS data.

Annotation of DMRs Associated Genes

Genomic position distribution of DMRs was performed using ChIPseeker (v1.14.0) Bioconductor package. Every DMR starting from 3 kb downstream to 3 kb upstream of the TSS was assigned to the corresponding RefSeq gene. Functional enrichment analysis of DMRs associated genes at KEGG pathways was performed using clusterProfiler (v3.6.0) Bioconductor package.

Enrichment With Scoliosis Related Genes in DMRs

We searched public database including HPO, OMIM, Clinvar, and Pubmed using scoliosis as key term, 940 associated genes were collected as a scoliosis related gene set (**Table S2**). For these genes, an enrichment *P*-value was calculated based on hypergeometric testing:F

$$P = \frac{C_M^x C_{N-M}^{k-x}}{C_N^k},\ k \in \{0,\ 1,\ 2, \ldots, m\}$$

N represents the total gene counts of RefSeq database, *M* means the total number of scoliosis-related genes, *k* is the number of genes that have mutations identified from WES data, and *x* is the number of intersected genes between these two gene sets.

For the 940 scoliosis related genes, an enrichment *P*-value was also calculated based on hypergeometric testing above. Here *N* is total RefSeq genes (27967), *M* is the number of scoliosis-related genes, *k* is the number of DMRs associated genes, and *x* is the number of intersected genes between the two gene sets.

Validation in the Replication Cohort

Bisulfite Conversion and Multiplex Amplification

DNA methylation level was analysis by MethylTarget™ (Genesky Biotechnologies Inc., Shanghai, China), an NGS-based multiple Targeted CpG methylation analysis method. Specifically, the genomic regions of interest were analyzed and transformed into bisulfite-converted sequences by geneCpG software. PCR primer sets were designed with the Methylation Primer software from bisulfate converted DNA.

Genomic DNA (400 ng) was subjected to sodium bisulfite treatment using EZ DNA Methylation™-GOLD Kit (Zymo Research) according to the manufacturer's protocols. Multiplex PCR was performed with optimized primer sets combination. A 20 μl PCR reaction mixture was prepared for each reaction and included 1x reaction buffer (Takara), 3 mM Mg^{2+}, 0.2 mM dNTP, 0.1 μM of each primer, 1U HotStarTaq polymerase (Takara) and 2 μl template DNA. The cycling program was 95°C for 2 min; 11 cycles of 94°C for 20 s, 63°C for 40s with a decreasing temperature step of 0.5°C per cycle, 72°C for 1 min; then followed by 24 cycles of 94°C for 20 s, 65°C for 30 s, 72°C for 1 min; 72°C for 2 min.

Index PCR

PCR amplicons were diluted and amplified using indexed primers. Specifically, a 20 μl mixture was prepared for each reaction and included 1x reaction buffer (NEB Q5™), 0.3 mM dNTP, 0.3 μM of F primer, 0.3 μM of index primer, 1 U Q5™ DNA polymerase (NEB) and 1 μL diluted template. The cycling program was 98°C for 30 s; 11 cycles of 98°C for 10 s, 65°C for 30 s, 72°C for 30 s; 72°C for 5 min. PCR amplicons (170bp-270bp) were separated by agarose electrophoresis and purified using QIAquick Gel Extraction kit (QIAGEN).

Sequencing

Libraries from different samples were quantified and pooled together, followed by sequencing on the Illumina MiSeq platform according to manufacturer's protocols. Sequencing was performed with a 2 × 150 bp paired-end mode.

Data Analysis

Quality control of sequencing reads was performed by FastQC. Filtered reads were mapped to genome by Blast. After reads

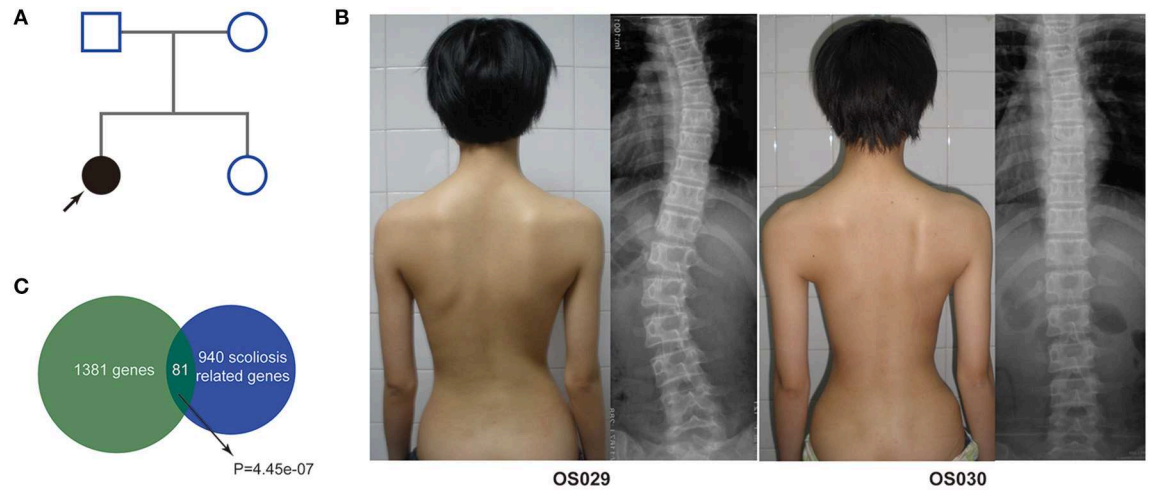


FIGURE 1 | Clinical and WES findings of the AIS discordant monozygotic twin. **(A)** Pedigree of the patient. **(B)** Clinical characteristics and imaging of the proband and her healthy twin sister*. **(C)** Enrichment of WES mutation genes with scoliosis related genes. *Written informed consent was obtained from their parents for the publication of this image.

recalibration with USEARCH, methylation and haplotype were analyzed using Perl script. Statistics were performed by *U*-test and ANOVA.

RESULTS

Clinical Information

The monozygotic twins were 15-year-old girls, one of them (OS029) was diagnosed with AIS at 12. The patient had two apexes of scoliosis, T3-T11 with Cobb angle of about 27 degrees and T12-L3 with Cobb angle of about 26 degrees. The spine images of the healthy sister (OS030) and parents were normal (**Figures 1A,B**). In the replication cohort, the mean age of the AIS patients was 15 years old, and the mean age of the controls was 28 years old. All of the participants were Chinese females.

Whole-Exome Sequence Results

After strict filtering and genomic annotation, we failed to identify pathogenic variants responsible for the discordant phenotype in the MZ twin. Finally, only three *de novo* variants were identified, one missense SNV in the exon of *PASD1* (c.425C>T), one missense SNV in the exon of *SLC44A4* (c.1281G>C) and one synonymous SNV in the exon of *AGTR2* (c.1080C>A), but all the *de novo* exome variants occur in both MZ twins (**Table S3**). Then we compared all the exome variants between the MZ twins. There were 30,700 same mutations, 32 Type 1 mutations and 1,830 Type 3 mutations. There were no Type 2 mutations (**Table 1**). Altogether there were 1,862 potential pathogenic mutations located in 1,381 genes (**Table S4**), and 81 were overlapped with scoliosis related genes (**Figure 1C**). By comparing with randomized gene list, these potential pathogenic genes undergoing mutations were significantly enriched in scoliosis genes ($P = 4.45 \times 10^{-07}$). Since both of the parents had no spinal deformity, these potential pathogenic mutations may only increase the susceptibility of scoliosis without causing scoliosis directly.

TABLE 1 | Different mutations in the MZ twin discordant for AIS.

Type	Count
Equal type (both mutation)	30,700
Different type 1 (both mutation)	32
Different type 2 (only mutation in Control)	0
Different type 3 (only mutation in Case)	1,830

DNA Methylation Differences in the MZ Twins

The global methylation status between the MZ twins were very similar, only a small percentage of promoter existed difference (348/27967, 1.24%, **Table S5**). We compared whole-genome DNA methylation level between the MZ twins, and found that both the hyper- and hypo- DMRs were differently distributed (**Figure 2A**). Finally, we identified 313 hypermethylated DMRs and 397 hypomethylated DMRs (**Table S5**). The average hypermethylation difference was 38% and the average hypomethylation difference was 31% between the two samples. We found that more DMRs locate in the promoter regions (**Figure 2B**) and most of them located in CGI (CpG island) (**Figure 2C**). Significant methylation level difference between the twins may influence the function of DMR-related genes and contribute to the etiology of this disease. Interestingly, these DMRs overlapped with 25 allele-specific methylated regions, suggesting that they were related to abnormal genomic imprinting (**Figure S3**) (de Sa Machado Araujo et al., 2018).

We enriched the genes associated with these DMRs and then subjected these genes to KEGG (Kyoto Encyclopedia

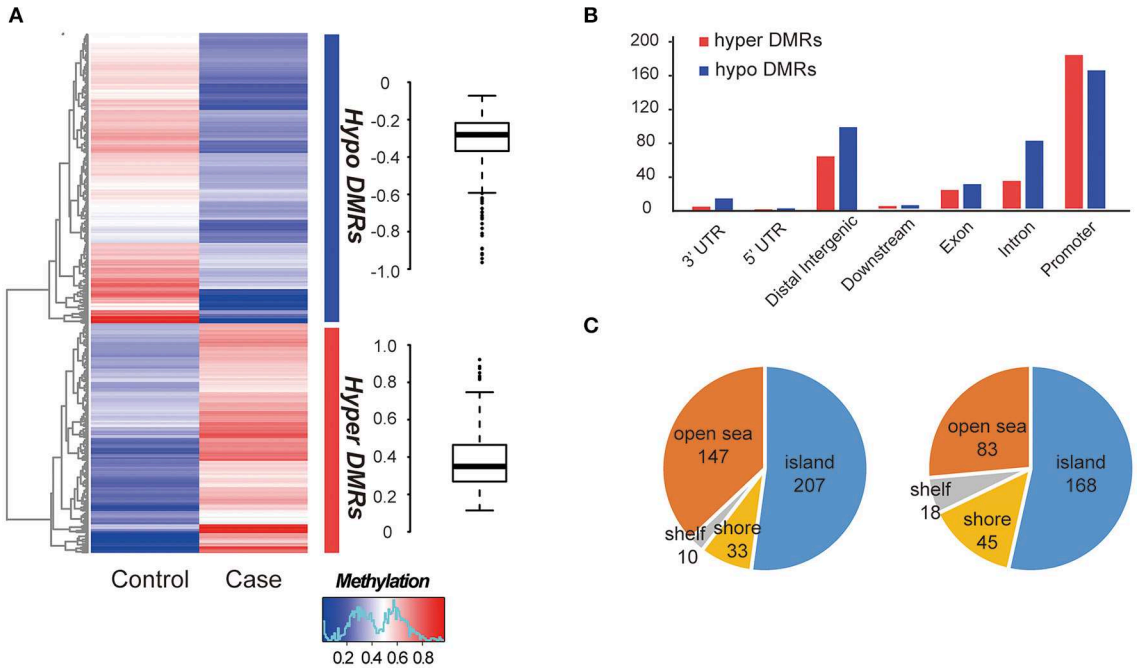


FIGURE 2 | Comparison of DMRs between two samples. **(A)** Hierarchical clustering of DNA methylation profiles of AIS twins. The distribution of the methylation difference was shown in the box plot at the right side. **(B)** Characteristics of different functional element DNA methylation profiles of AIS twins. **(C)** CpG island distributions of hypomethylated DMRs (left) and hypermethylated DMRs (right).

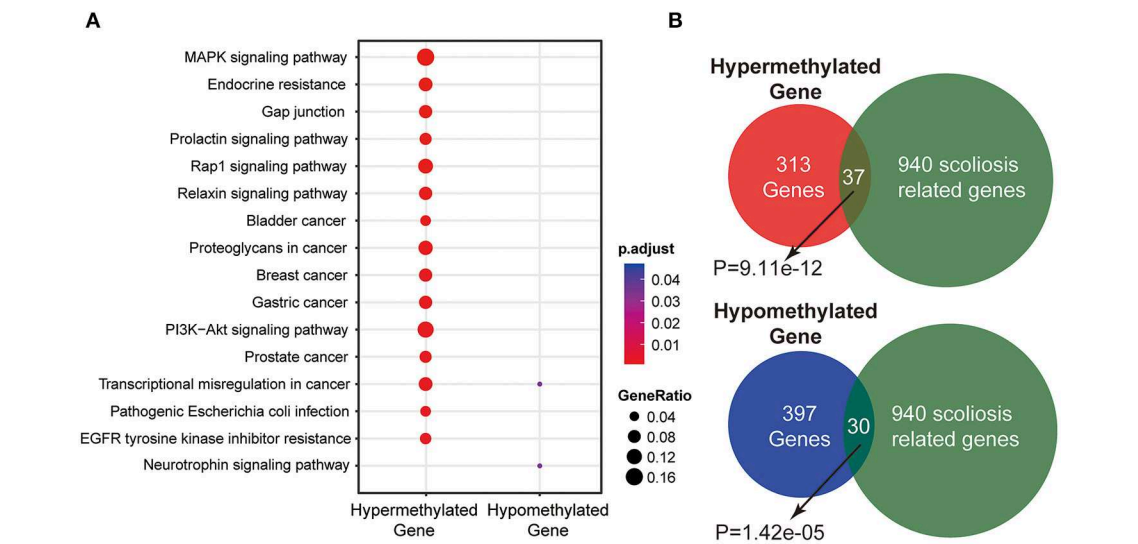


FIGURE 3 | Annotations of DMR related genes between two samples. **(A)** Scatterplot of DMR related genes enrichment with KEGG pathway. **(B)** Enrichment of DMR related genes with scoliosis related genes.

of Genes and Genomes) pathway analysis. It revealed that DMR associated genes were enriched in several pathways. Three pathways with the most significant association were the MAPK signaling pathway, PI3K-Akt signaling pathway and Rap1 signaling pathway (**Figure 3A**), indicating that these pathways could be associated with AIS.

Enrichment of Scoliosis Related Genes in DMRs

To explore the relationship between these DMRs and scoliosis, we enriched the DMR related genes and found they were significantly overlapped with scoliosis related genes. In hypermethylated DMR associated genes group, 37 genes were

TABLE 2 | Selected epigenome DNAm difference region in the replicated cohort.

Name of DMRs	Selected DMR position	Mean methylation level in AIS group	Mean methylation level in control group	*P*-value	*Z*-value	MethylDiff
CASK:TSS-DMR	chrX:41783051-41783257	4.31×10^{-1}	4.26×10^{-1}	0.68	0.46	5.39×10^{-3}
EMD:TSS-DMR	chrX:153606503-153607014	3.92×10^{-1}	4.01×10^{-1}	0.20	−0.84	-9.31×10^{-3}
ESR1:TSS-DMR	chr6:152128858-152128958	2.98×10^{-2}	3.14×10^{-2}	0.74	0.64	-1.64×10^{-3}
FMR1:TSS-DMR	chrX:146993416-146993736	4.28×10^{-1}	4.17×10^{-1}	0.83	0.95	1.06×10^{-2}
GDI1:TSS-DMR	chrX:153665399-153665481	4.93×10^{-1}	4.86×10^{-1}	0.95	1.61	6.81×10^{-3}
GPC3:TSS-DMR	chrX:133119090-133119453	4.01×10^{-1}	3.98×10^{-1}	0.43	−0.18	3.41×10^{-3}
KCNQ1OT1:TSS-DMR	chr11:2720966-2721111	4.05×10^{-1}	4.09×10^{-1}	0.55	0.12	-4.21×10^{-3}
MAP2K2:TSS-DMR	chr19:4123972-4124036	1.47×10^{-2}	1.60×10^{-2}	0.28	−0.59	-1.32×10^{-3}
NDN:TSS-DMR	chr15:23932133-23932304	3.78×10^{-1}	3.64×10^{-1}	**0.04**	−1.75	1.45×10^{-2}
PAX1:TSS-DMR	chr20:21687204-21687404	4.90×10^{-2}	5.45×10^{-2}	0.11	−1.24	-5.51×10^{-3}
PHF8:TSS-DMR	chrX:54069169-54069423	4.85×10^{-1}	4.67×10^{-1}	0.54	0.09	1.77×10^{-2}
RPS6KA3:TSS-DMR	chrX:20285105-20285182	3.17×10^{-1}	3.23×10^{-1}	0.33	−0.45	-6.05×10^{-3}
TUBB2B:TSS-DMR	chr6:3227646-3227824	2.25×10^{-2}	2.40×10^{-2}	0.28	−0.59	-1.45×10^{-3}
TUBB3:TSS-DMR	chr16:89989983-89990156	3.71×10^{-2}	4.07×10^{-2}	0.11	−1.20	-3.59×10^{-3}

The bold values mean statistically significant.

overlapped with these scoliosis genes. In hypomethylated DMR associated genes group, 30 genes were overlapped with scoliosis related genes (**Figure 3B**). After comparing to randomized gene list, the DMR associated genes were significantly overlapped with scoliosis genes ($p = 9.11 \times 10^{-12}$ and 1.42×10^{-05}, respectively), indicating that the methylation of scoliosis-related genes may be associated with the risk of AIS development.

Validation of Candidate DMRs in the Replication Cohort

For further identification of specific pathogenic epigenetic variants, we selected 14 DMRs from the literature for validation based on the related gene's function in the replication cohort (**Table 2**). All of the 14 genes were related to scoliosis and manifested with high methylation difference in the MZ twins. We found the methylation level of *NDN*: TSS-DMR (chr15:23932133-23932304, hg19) was significantly differentiated in the replication cohort. The mean methylation level in the AIS group was significantly higher than the mean methylation level in the control group (3.78×10^{-1} and 3.64×10^{-1}, $p = 0.04$). This DMR locates at the promoter region of gene *NDN*. The different methylation of this gene may influence gene expression and be associated with the phenotype of AIS.

DISCUSSION

This study represented the first combined analysis of whole exome and whole epigenome of MZ twins discordant for AIS. According to previous studies, genetic variations can predispose to AIS (Buchan et al., 2014a; Gao et al., 2017). Therefore, we first performed WES to detect genetic variations. We found three *de novo* SNVs, but the healthy twin sister also had these SNVs, indicating that pathogenesis of AIS may not be on account of these variants. Epigenetic variation is widely acknowledged to be involved in the pathogenesis of various diseases (Jin et al., 2014; Lyu et al., 2018). In this study, we compared whole-genome DNA methylation between the discordant twins using WGBS. Potential pathological DMRs were identified. The DMR related genes were enriched predominantly in the MAPK signaling pathway, and other 12 signaling pathways (i.e., PI3K-Akt and Rap1 signaling pathways etc.). The MAPK and PI3K-Akt signaling pathways have been reported to play an important role in osteoblast differentiation and skeletogenesis (Ge et al., 2007; Zhang et al., 2014; Iezaki et al., 2018). Previous studies also indicated that Rap1 signaling pathway is critical for myogenic differentiation and osteoclast function (Pizon et al., 1996; Zou et al., 2013). Abnormalities of these pathways could give rise to disorders manifested with scoliosis (Tiffin et al., 2013; Martinez-Lopez et al., 2017; Tsai et al., 2018). The previous study also showed that MAPK signaling pathway and PI3K-Akt signaling pathway were downregulated in bone marrow mesenchymal stem cells (BM-MSCs) of AIS, which may contribute to the AIS initiation and development (Zhuang et al., 2016). We hypothesized that hypermethylation of these DMR related genes could regulate MAPK, PI3K-Akt, and Rap1 signaling pathways, which may influence the initiation of AIS. This regulation mechanism may contribute to the discordance of AIS in this twin pair.

We also found the DMR associated genes were significantly overlapped with scoliosis related genes, which indicated that the methylation of scoliosis-related genes may play a pivotal role in the development of AIS. To explore specific epigenetic variants, we compared the specific methylation difference of several DMRs in the replication cohort. We found the *NDN*: TSS-DMR (chr15:23932133-23932304, hg19) was significantly differentiated in the replication cohort. The AIS group had higher methylation level of this DMR than the control group. This DMR locates in the promoter area of *NDN* and has 146 base pairs distance to TSS. *NDN* locates at 15q11.2 and is an intronless gene located in the Prader-Willi syndrome (PWS) deletion region. Previous studies hypothesized that lack of its coding protein Necdin during development may contribute to PWS (Jay et al.,

1997; Miller et al., 2009). PWS is a rare disease associated with a variety of musculoskeletal abnormalities, about 43.4% of patients were afflicted with scoliosis (Odent et al., 2008). Besides, *NDN* is a well-known maternally imprinted gene and is expressed exclusively from the paternal allele, and its methylation are persistent markers of gene regulation (Lau et al., 2004). Previous studies had proved that imprinted DMRs (iDMRs) could be perturbed in kinds of diseases (de Sa Machado Araujo et al., 2018). In our study, we hypothesized that higher methylation could decrease the expression of *NDN*, which predisposed the patient to AIS.

There are several limitations in our study. First, since WES only covered most exons of the human reference genome, we did not assess the rest of the genome and other types of variants, such as karyotype or structural variants. Therefore, we cannot exclude the possibility that genome variants contribute to the etiology of this MZ twin pair discordant for AIS. Second, since all of the AIS patients were teenagers, the age of the control group is not matched with the case group in the replication cohort. The unmatched age may cause bias of the implications. Third, the sample size of the MZ twins and the replication cohort was relatively small. Larger samples are needed for the replication.

CONCLUSION

In conclusion, we described the genome methylation difference in an AIS discordant MZ twin pair using WGBS and selected DMR methylation difference in a replication cohort. We found that the *NDN*: TSS-DMR had higher methylation level in AIS, which may elucidate the potential etiology of AIS.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the http://discostudy.org/data/DISCO2019-1.html.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethical Review Board of Peking Union Medical College Hospital. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

GL, LW, and NW performed the research, analyzed, and interpreted the data. GL drafted the manuscript. XW, XY, ML, ZY, SL, YZu, YN, SZ, and YZh helped sample collection. JZ and JS performed phenotyping of patients. XW, ZY, and XY helped analysis and interpretation of NGS data. YN provided technique support. GQ, YW, ZW, JZ, and JS offered professional discussions and instructions. ZW and NW conceived and designed the study, revised the manuscript, and provided the final approval of the manuscript.

FUNDING

This research was funded in part by the National Natural Science Foundation of China (81822030 to NW, 81902271 to GL, 81772299 and 81930068 to ZW, 81772301 and 81972132 to GQ, 81672123 to JZ, 81572097 and 81871746 to YW), Beijing Natural Science Foundation (7172175 to NW, 7191007 to ZW, and 7184232 to SL), CAMS Initiative Fund for Medical Sciences (2016-I2M-3-003 to GQ and NW, 2016-I2M-2-006 and 2017-I2M-2-001 to ZW), the Central Level Public Interest Program for Scientific Research Institute (2018RC31003 to NW), the National Key Research and Development Program of China (No. 2018YFC0910506 to NW and ZW), and the International Program Associate grant from RIKEN of Japan (190038 to ML).

ACKNOWLEDGMENTS

We give our great thanks to Dr. Shugang Li, Hong Zhao, and Yu Zhao for the patient enrollment and clinical evaluation; and to the patients for their active participation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00364/full#supplementary-material

REFERENCES

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, 248–249. doi: 10.1038/nmeth0410-248

Buchan, J. G., Alvarado, D. M., Haller, G., Aferol, H., Miller, N. H., Dobbs, M. B., et al. (2014b). Are copy number variants associated with adolescent idiopathic scoliosis? *Clin. Orthop. Relat. Res.* 472, 3216–3225. doi: 10.1007/s11999-014-3766-8

Buchan, J. G., Alvarado, D. M., Haller, G. E., Cruchaga, C., Harms, M. B., Zhang, T., et al. (2014a). Rare variants in FBN1 and FBN2 are associated with severe adolescent idiopathic scoliosis. *Hum. Mol. Genet.* 23, 5271–5282. doi: 10.1093/hmg/ddu224

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6:e1001025. doi: 10.1371/journal.pcbi.1001025

de Sa Machado Araujo, G., da Silva Francisco Junior, R., Dos Santos Ferreira, C., Mozer Rodrigues, P. T., Terra Machado, D., Louvain de Souza, T., et al. (2018). Maternal 5(m)CpG imprints at the PARD6G-AS1 and GCSAML differentially methylated regions are decoupled from parent-of-origin expression effects in multiple human tissues. *Front. Genet.* 9:36. doi: 10.3389/fgene.2018.00036

Elboudwarej, E., Cole, M., Briggs, F. B., Fouts, A., Fain, P. R., Quach, H., et al. (2016). Hypomethylation within gene promoter regions and type 1 diabetes in discordant monozygotic twins. *J. Autoimmun.* 68, 23–29. doi: 10.1016/j.jaut.2015.12.003

Gao, W., Chen, C., Zhou, T., Yang, S., Gao, B., Zhou, H., et al. (2017). Rare coding variants in MAPK7 predispose to adolescent idiopathic scoliosis. *Hum. Mutat.* 38, 1500–1510. doi: 10.1002/humu.23296

Ge, C., Xiao, G., Jiang, D., and Franceschi, R. T. (2007). Critical role of the extracellular signal-regulated kinase-MAPK pathway in osteoblast differentiation and skeletal development. *J. Cell Biol.* 176, 709–718. doi: 10.1083/jcb.200610046

Iezaki, T., Fukasawa, K., Horie, T., Park, G., Robinson, S., Nakaya, M., et al. (2018). The MAPK Erk5 is necessary for proper skeletogenesis through a molecular axis that involves Smurfs-Smads-Sox9. *Development* 145:dev164004. doi: 10.1242/dev.164004

Ikegawa, S. (2016). Genomic study of adolescent idiopathic scoliosis in Japan. *Scoliosis Spinal Disord.* 11:5. doi: 10.1186/s13013-016-0067-x

Jay, P., Rougeulle, C., Massacrier, A., Moncla, A., Mattei, M. G., Malzac, P., et al. (1997). The human necdin gene, NDN, is maternally imprinted and located in the Prader-Willi syndrome chromosomal region. *Nat. Genet.* 17, 357–361. doi: 10.1038/ng1197-357

Jin, M., Zhu, S., Hu, P., Liu, D., Li, Q., Li, Z., et al. (2014). Genomic and epigenomic analyses of monozygotic twins discordant for congenital renal agenesis. *Am. J. Kidney Dis.* 64, 119–122. doi: 10.1053/j.ajkd.2014.01.423

Kesling, K. L., and Reinker, K. A. (1997). Scoliosis in twins. A meta-analysis of the literature and report of six cases. *Spine* 22, 2009–2014; discussion 2015. doi: 10.1097/00007632-199709010-00014

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, 310–315. doi: 10.1038/ng.2892

Kou, I., Takahashi, Y., Johnson, T. A., Takahashi, A., Guo, L., Dai, J., et al. (2013). Genetic variants in GPR126 are associated with adolescent idiopathic scoliosis. *Nat. Genet.* 45, 676–679. doi: 10.1038/ng.2639

Lau, J. C., Hanel, M. L., and Wevrick, R. (2004). Tissue-specific and imprinted epigenetic modifications of the human NDN gene. *Nucleic Acids Res.* 32, 3376–3382. doi: 10.1093/nar/gkh671

Li, W., Li, Y., Zhang, L., Guo, H., Tian, D., Li, Y., et al. (2016). AKAP2 identified as a novel gene mutated in a Chinese family with adolescent idiopathic scoliosis. *J. Med. Genet.* 53, 488–493. doi: 10.1136/jmedgenet-2015-103684

Liu, G., Liu, S., Li, X., Chen, J., Chen, W., Zuo, Y., et al. (2019). Genetic polymorphisms of PAX1 are functionally associated with different PUMC types of adolescent idiopathic scoliosis in a northern Chinese Han population. *Gene* 688, 215–220. doi: 10.1016/j.gene.2018.12.013

Liu, G., Liu, S., Lin, M., Li, X., Chen, W., Zuo, Y., et al. (2018). Genetic polymorphisms of GPR126 are functionally associated with PUMC classifications of adolescent idiopathic scoliosis in a Northern Han population. *J. Cell. Mol. Med.* 22, 1964–1971. doi: 10.1111/jcmm.13486

Liu, S., Wu, N., Zuo, Y., Zhou, Y., Liu, J., Liu, Z., et al. (2017). Genetic polymorphism of LBX1 is associated with adolescent idiopathic scoliosis in northern chinese han population. *Spine* 42, 1125–1129. doi: 10.1097/BRS.0000000000002111

Lyu, G., Zhang, C., Ling, T., Liu, R., Zong, L., Guan, Y., et al. (2018). Genome and epigenome analysis of monozygotic twins discordant for congenital heart disease. *BMC Genomics* 19:428. doi: 10.1186/s12864-018-4814-7

Mao, S. H., Qian, B. P., Shi, B., Zhu, Z. Z., and Qiu, Y. (2018). Quantitative evaluation of the relationship between COMP promoter methylation and the susceptibility and curve progression of adolescent idiopathic scoliosis. *Eur. Spine J.* 27, 272–277. doi: 10.1007/s00586-017-5309-y

Martinez-Lopez, A., Blasco-Morente, G., Perez-Lopez, I., Herrera-Garcia, J. D., Luque-Valenzuela, M., Sanchez-Cano, D., et al. (2017). CLOVES syndrome: review of a PIK3CA-related overgrowth spectrum (PROS). *Clin. Genet.* 91, 14–21. doi: 10.1111/cge.12832

Meng, Y., Lin, T., Liang, S., Gao, R., Jiang, H., Shao, W., et al. (2018). Value of DNA methylation in predicting curve progression in patients with adolescent idiopathic scoliosis. *EBioMedicine* 36, 489–496. doi: 10.1016/j.ebiom.2018.09.014

Miller, N. L., Wevrick, R., and Mellon, P. L. (2009). Necdin, a Prader-Willi syndrome candidate gene, regulates gonadotropin-releasing hormone neurons during development. *Hum. Mol. Genet.* 18, 248–260. doi: 10.1093/hmg/ddn344

Odent, T., Accadbled, F., Koureas, G., Cournot, M., Moine, A., Diene, G., et al. (2008). Scoliosis in patients with Prader-Willi Syndrome. *Pediatrics* 122, e499–e503. doi: 10.1542/peds.2007-3487

Ogura, Y., Kou, I., Miura, S., Takahashi, A., Xu, L., Takeda, K., et al. (2015). A functional SNP in BNC2 is associated with adolescent idiopathic scoliosis. *Am. J. Hum. Genet.* 97, 337–342. doi: 10.1016/j.ajhg.2015.06.012

Pizon, V., Cifuentes-Diaz, C., Mege, R. M., Baldacci, G., and Rieger, F. (1996). Expression and localization of RAP1 proteins during myogenic differentiation. *Eur. J. Cell Biol.* 69, 224–235.

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* 11, 361–362. doi: 10.1038/nmeth.2890

Sharma, S., Londono, D., Eckalbar, W. L., Gao, X., Zhang, D., Mauldin, K., et al. (2015). A PAX1 enhancer locus is associated with susceptibility to idiopathic scoliosis in females. *Nat. Commun.* 6:6452. doi: 10.1038/ncomms7452

Shi, B., Xu, L., Mao, S., Xu, L., Liu, Z., Sun, X., et al. (2018). Abnormal PITX1 gene methylation in adolescent idiopathic scoliosis: a pilot study. *BMC Musculoskelet. Disord.* 19:138. doi: 10.1186/s12891-018-2054-2

Takahashi, Y., Kou, I., Takahashi, A., Johnson, T. A., Kono, K., Kawakami, N., et al. (2011). A genome-wide association study identifies common variants near LBX1 associated with adolescent idiopathic scoliosis. *Nat. Genet.* 43, 1237–1240. doi: 10.1038/ng.974

Tiffin, H. R., Jenkins, Z. A., Gray, M. J., Cameron-Christie, S. R., Eaton, J., Aftimos, S., et al. (2013). Dysregulation of FHL1 spliceforms due to an indel mutation produces an emery-dreifuss muscular dystrophy plus phenotype. *Neurogenetics* 14, 113–121. doi: 10.1007/s10048-013-0359-8

Tsai, I. C., McKnight, K., McKinstry, S. U., Maynard, A. T., Tan, P. L., Golzio, C., et al. (2018). Small molecule inhibition of RAS/MAPK signaling ameliorates developmental pathologies of Kabuki Syndrome. *Sci. Rep.* 8:10779. doi: 10.1038/s41598-018-28709-y

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9. doi: 10.1038/nprot.2015.123

Wang, K., Zhao, S., Liu, B., Zhang, Q., Li, Y., Liu, J., et al. (2018a). Perturbations of BMP/TGF-beta and VEGF/VEGFR signalling pathways in non-syndromic sporadic brain arteriovenous malformations (BAVM). *J. Med. Genet.* 55, 675–684. doi: 10.1136/jmedgenet-2017-105224

Wang, K., Zhao, S., Zhang, Q., Yuan, J., Liu, J., Ding, X., et al. (2018b). Whole-exome sequencing reveals known and novel variants in a cohort of intracranial vertebral-basilar artery dissection (IVAD). *J. Hum. Genet.* 63, 1119–1128. doi: 10.1038/s10038-018-0496-x

Yawn, B. P., Yawn, R. A., Hodge, D., Kurland, M., Shaughnessy, W. J., Ilstrup, D., et al. (1999). A population-based study of school scoliosis screening. *JAMA* 282, 1427–1432. doi: 10.1001/jama.282.15.1427

Zhang, Y., Pizzute, T., and Pei, M. (2014). A review of crosstalk between MAPK and Wnt signals and its impact on cartilage regeneration. *Cell Tissue Res.* 358, 633–649. doi: 10.1007/s00441-014-2010-x

Zhuang, Q., Mao, W., Xu, P., Li, H., Sun, Z., Li, S., et al. (2016). Identification of differential genes expression profiles and pathways of bone marrow mesenchymal stem cells of adolescent idiopathic scoliosis patients by microarray and integrated gene network analysis. *Spine* 41, 840–855. doi: 10.1097/BRS.0000000000001394

Zou, W., Izawa, T., Zhu, T., Chappel, J., Otero, K., Monkley, S. J., et al. (2013). Talin1 and Rap1 are critical for osteoclast function. *Mol. Cell. Biol.* 33, 830–844. doi: 10.1128/MCB.00790-12

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 © 2019 Liu, Wang, Wang, Yan, Yang, Lin, Liu, Zuo, Niu, Zhao, Zhao, Zhang, Shen, Wang, Qiu, Wu and Wu. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 17 December 2019
doi: 10.3389/fbioe.2019.00427

A Novel Approach to Identify Enhancer lincRNAs by Integrating Genome, Epigenome, and Regulatome

Hui Liu[1†], Tiantongfei Jiang[1†], Shuyuan Wang[1†], Xiang Chen[1], Xiaoyan Jin[1], Qi Wang[1], Xinhui Li[1], Jiaqi Yin[1], Tingting Shao[1], Yongsheng Li[1,2*], Juan Xu[1,2*] and Qiong Wu[3*]*

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] Key Laboratory of Tropical Translational Medicine of Ministry of Education, College of Biomedical Informatics and Engineering, Hainan Medical University, Haikou, China, [3] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Hongwei Wang,
Sun Yat-sen University, China
Jianfeng Xu,
Baylor College of Medicine,
United States

***Correspondence:**
Qiong Wu
kigo@hit.edu.cn
Juan Xu
xujuanbiocc@ems.hrbmu.edu.cn
Yongsheng Li
liyongsheng@ems.hrbmu.edu.cn
Tingting Shao
shaotingting_05@163.com

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

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 14 October 2019
Accepted: 02 December 2019
Published: 17 December 2019

Citation:
Liu H, Jiang T, Wang S, Chen X, Jin X, Wang Q, Li X, Yin J, Shao T, Li Y, Xu J and Wu Q (2019) A Novel Approach to Identify Enhancer lincRNAs by Integrating Genome, Epigenome, and Regulatome. Front. Bioeng. Biotechnol. 7:427. doi: 10.3389/fbioe.2019.00427

LincRNAs enriched with high H3K4me1 and low H3K4me3 signals often have the enhancer-like features which are named as enhancer-associated lincRNAs (elincRNAs). ElincRNAs are considered to be indispensable for target gene transcription, which play important roles in development, signaling events, and even diseases. In this study, we developed a regularized regression model to identify elincRNAs by integrating the genomic, epigenomic, and regulatory data. Application of the proposed method to mouse ESCs reveals that besides the basic well-known epigenetic features H3K4me1 and H3K4me3, more specific epigenetic features, such as high DNA methylation, high H3K122ac, and H3K36me3 were contributed to mark elincRNAs with the best accuracy and precision. Finally, 3729 elincRNAs were identified in mouse ESCs. Furthermore, the elincRNAs and canonical lincRNAs exhibit distinct genomic features, and elincRNAs have the higher CGI enrichment and lower sequence conservation. Through the analysis of transcription regulation, we found that elincRNAs were significantly regulated by NANOG, POU5F1, SOX2 and ESRRB, and were involved in the core transcriptional regulatory circuitry controlling ES cell state Function enrichment analysis further discovered that elincRNAs tended to regulate specific embryonic development biological processes. These results indicated that these two types of lincRNAs had both specific epigenetic and transcriptional regulation mechanism and display distinct functional characters. In conclusion, we presented a credible computational model to prioritize novel elincRNAs, and depicted the atlas of elincRNAs in mouse ESCs, which would help dissect the function roles of lncRNAs during the mammalian development and diseases.

Keywords: enhancer lincRNA, mouse embryonic stem cell, regularized regression model, epigenome, regulatome

INTRODUCTION

Long non-coding RNAs (lncRNAs) are a class of RNA transcripts longer than 200 nucleotides which are not transcribed into proteins (Derrien et al., 2012). Most lncRNAs are PolII-dependent transcribed, whose transcripts have the exon-intron structure with the 5′ capping and poly-A trail (Dinger et al., 2008). In comparison with mRNAs, lncRNAs generally have lower expression

abundance and sequence conservation. Recently, crucial roles of lncRNAs in cell differentiation, embryonic development and complex diseases have been confirmed in multiple studies (Lee and Bartolomei, 2013; Flynn and Chang, 2014; Hamazaki et al., 2015; Zhou et al., 2017, 2018, 2019). Guttmanet al. knocked down 147 lncRNAs expressed in mouse embryonic stem cells (ESCs), which resulted in the significant changes of the global gene expression in mouse ESCs, indicating that lncRNAs had key roles in the circuitry controlling ES cell state (Guttman et al., 2011). LncRNAs can regulate gene expression through a number of mechanisms, across epigenetic, transcriptional, and alternative splicing regulation (Gong and Maquat, 2011; Sun et al., 2014; Tay et al., 2014; Rutenberg-Schoenberg et al., 2016; Zhou et al., 2018). LncRNAs could regulate gene expressions in both cis and trans effects, influencing or interacting with nearby or distant genes (Signal et al., 2016). Furthermore, lncRNAs could regulate the gene expressions in mammalian development.

Intergenic lncRNAs (lincRNAs) are one type of the most widely studied lncRNAs, besides other lncRNA types, such as sense lncRNAs, antisense lncRNAs and intronic lncRNAs (St Laurent et al., 2015). Studies have shown that lincRNA tends to be located on functional elements, such as enhancers and promoters. Enhancers are the short DNA regions which could be bound by activators to enhance the transcription of the target genes in developmental patterns, cell differentiation, even in human diseases (Hnisz et al., 2013; Emera et al., 2016; Karnuta and Scacheri, 2018). Many studies have discovered and described the functional regulations of enhancers on gene expression (Cinghu et al., 2017; Catarino and Stark, 2018). LincRNAs, whose genomic location are overlapped with enhancers, generally transcribe from enhancer regions, and are often named as enhancer-associated lincRNAs (elincRNAs). Thus, they are also considered as enhancer RNAs (eRNAs), and could be involved in the enhancer-promoter looping inside topological associated domains (TAD) (Vance and Ponting, 2014). ElincRNAs could perform enhancer-like function that spatially and temporally regulate the target gene expression in cis or in trans format, during the mammalian development and diseases (Sakabe et al., 2012; Lam et al., 2014; Long et al., 2016). Hon et al. identified the e-lncRNAs from 27919 human lncRNA genes with high-confidence 5′ ends (Hon et al., 2017). More and more studies have shown that elincRNAs play key regulatory roles in cell differentiation and embryonic development. MyoD is a major regulator for muscle differentiation, and Mousavi et al. found that eRNA from the core enhancer could promote the MyoD expression (Mousavi et al., 2013). Recent researches indicated that elincRNAs were indispensable for the gene regulatory network by establishing and stabilizing the chromatin loops of enhancer-promoter interactions (Plank and Dean, 2014; Kim T. K. et al., 2015; Bose et al., 2017). Generally, there is another important type of RNA, named as canonical lincRNAs [also known as promoter-associated lincRNAs (plincRNAs)], presenting the canonical promoter-specific features with H3K4me3 enriched in their TSS intervals (Vance and Ponting, 2014; Kim T. K. et al., 2015).

Using epigenetic features, elincRNAs could be distinguished with the canonical lincRNAs (Signal et al., 2016). ElincRNAs are marked with high H3K4me1 and low H3K4me3 in the TSS regions, which are the enhancer-specific signatures, while canonical lincRNAs are marked with the high H3K4me3 and low or lacked H3K4me1 in the TSS regions which are canonical promoter features (Vance and Ponting, 2014). ElncRNAs and canonical lncRNAs were classified by H3K4me1/H3K4me3 ratio in TSS intervals in human monocytes by Ilott et al. (Ilott et al., 2014). Brain region-specific intergenic or intronic enhancer RNAs were marked with H3K4me1 and H3K27ac enrichment transcribed from enhancers in human genome. Bogu et al. also identified the elncRNAs and plncRNAs (marked by promoter features) across various tissues of mouse, by using the ChromHMM method to interrogate the chromatin status of enhancer and promoter regions (Bogu et al., 2015). Previous studies have shown that elincRNAs could be characterized by high H3K4me1/H3K4me3 ratio in their TSS intervals. As the key regulators for the establishment of the chromatin looping and activation of gene expression, the knowledge about genomic and epigenomic characteristics of elincRNAs is far from completeness. Moreover, H3K36me3 could enrich in the body regions of canonical lincRNAs, while lack in the elincRNAs body regions, just like the enhancers (Natoli and Andrau, 2012; Li et al., 2016). ElincRNA TSS intervals are CGI (CpG Island)-poor regions, while canonical lincRNA TSS intervals are the CGI-rich regions (Li et al., 2016). Thus, elincRNAs could be distinguished from canonical lincRNAs by integrating multi-omic features. However, there are few other known discriminatory chromatin modification or genomic features for the elincRNAs and canonical lincRNAs. DNA methylation could regulate H3K27ac at enhancer regions in mouse ESCs (King et al., 2016), and DNA methylation at enhancers could also identify distinct breast cancer lineages (Fleischer et al., 2017). However, whether this is a specific feature for elincRNAs is still unknown. Furthermore, which features could be contributing factors for elincRNAs have not been systematically interrogated from the perspective of genome, epigenome and regulatome. ElincRNAs could be better characterized by the integration of the significant genomic and epigenomic features. Furthermore, the annotation atlas of elincRNAs could be more comprehensive and complete.

As the acknowledged model organism, mouse is the well-known model for the researches of mammalian development and human diseases. ESCs are the primitive cells derived from preimplantation embryos that have the potential to differentiate into numerous specialized cell types. In this study, chromatin modification data and genomic features of mouse ESCs were collected derived from public online source to identify elincRNAs and canonical lincRNAs based on regularized regression model. Subsequently, elincRNAs and canonical lincRNAs were identified in the genome wide, using the predictive model with the specific features. Further, elincRNAs and canonical lincRNAs were characterized and compared from transcript structure, sequence features, epigenetic modifications and so on. Moreover, it is found that the TF binding patterns of the elincRNAs were different from canonical lincRNAs, which were enriched with specific development associated TFs. The identification of characteristic features of elincRNAs and canonical lincRNAs and the prediction of the two kind lincRNAs might be helpful for

the target gene expression regulation of enhancers and their transcriptions in development and human diseases.

MATERIALS AND METHODS

Epigenome, Regulatome, and Transcriptome Datasets

We collected epigenome, regulatome and transcriptome data of mouse ESCs. For epigenome data, the publicly available data contain 12 histone modifications, including 9 active histone modifications and 3 repressive histone modifications (Details see **Supplementary Table 1**). BS-Seq and DNase-Seq data of mouse ESCs were derived from GEO (**Supplementary Table 1**). For regulatome data, the 25 TF ChIP-seq data of mouse ESCs were obtained from public repository GEO (Details haven shown in **Supplementary Table 1**). For transcriptome data, the RNA-Seq data was derived from GEO with the ID GSE39619, and was used to quantify the expression levels of lincRNAs in mouse ESCs. The CAGE data was obtained from FANTOM5 project (http://fantom.gsc.riken.jp/5/).

Publicly Available Genomic and Functional Annotations

The genome annotation of known lincRNAs were derived from GENCODE Release M6 (GRCm38.p4), and was converted into mm9 assembly version, using NCBI remap_api.pl (https://www.ncbi.nlm.nih.gov/genome/tools/remap/docs/api). CGI and repeat elements annotations of mouse (mm9 version) were obtained from UCSC database. Mouse mm9 reference genome was sourced from UCSC. Conservation scoring by phyloP (phylogenetic *p*-values) for 20 placental mammal genomes, which contained mouse rat, pig, guinea, rabbit, human, chimp, and so on (Details see **Supplementary Table 2**). Enhancer annotation was derived from vista enhancer database, which contained 568 mouse enhancers (mm9 version) and 1747 human enhancers (hg19 version) (Visel et al., 2007). Promoter annotation was sourced from EPD (The Eukaryotic Promoter Database), in which 21239 mouse promoters (mm9 version) and 23248 human promoters (hg19 version) (**Supplementary Table 2**) (Dreos et al., 2017).

Sequencing Reads Alignment

Next Generation Sequencing data in SRA format, including RNA-Seq, ChIP-Seq and BS-Seq, were converted into fastq format by the NCBI-provided program sartoolkit. Trim_galore was used to remove the adapter sequence and low-quality reads of NGS data, and the quality control was performed by FastQC. The data passed the examination of quality control were mapped to the mouse reference genome for the following analysis. RNA-Seq data of mouse ESCs were processed and mapped into mouse reference genome (mm9 version) by HISAT2 (Kim D. et al., 2015), and the expression levels of the annotated and putative lincRNA sets were estimated by StringTie (Pertea et al., 2016). ChIP-Seq data of Histone modifications and TFs were aligned into mouse reference genome (mm9 version) using bowtie2 (Langmead and Salzberg, 2012). Peak callings which identified the enriched signal regions for ChIP-Seq were performed by MACS2 (Zhang Y. et al., 2008). BS-Seq for mouse ESCs was processed by Bismark, and the DNA methylation levels (βvalues) were extracted by bismark_methyaltion_extrasctor at single base resolution (Krueger and Andrews, 2011). The conversion of SAM to BAM file formats was conducted by samtools. The visualization of NGS data was performed by ngs.plot (Loh and Shen, 2016) and deepTools2 (Ramirez et al., 2016), including the meta profiles and the heatmaps of TSS, transcripts and the particular regions. In addition, the K-means clustering based on histone modifications was performed by DeepTools2 (Ramirez et al., 2016).

Predictive Model for elincRNAs and Canonical lincRNAs

In order to identify elincRNAs, we propose a multi-step predictive method. Firstly, by using H3K4me1 and H3K4me3 signals of TSS intervals, High confidence elincRNAs and canonical lincRNAs were identified as training set for the predictive model. Then, genomic, epigenomic, and regulatory features were calculated for high confidence set. Using the mutil-omic features, the predictive model was developed, based on regularized regression model. Further, the predictive model was evaluated by ten-fold cross-validation and independent test set with high assessment measures. Finally, elincRNAs were identified from the canonical lincRNAs by the predictive model with the specific features.

Identifying High Confidence Sets of Elincrnas and Canonical Lincrnas

For identifying high confidence sets of elincRNAs and canonical lincRNAs, TSS intervals of elincRNAs and canonical lincRNAs are defined as the regions from TSS up-stream 500 bp to TSS down-stream 500 bp. Gene body regions of elincRNAs and canonical lincRNAs are defined as the regions from TSS down-stream 500 bp to TTS (Transcription Termination Site). Then, 4,157 lincRNA transcripts were collected from the annotation of UCSC mm10 version, whose genomic coordinates were converted to the mm9 version. As significant makers for elincRNAs and canonical lincRNAs, H3K4me1 and H3K4me3 status in the expressed lincRNA TSS intervals were interrogated. Here, the expressed lincRNAs were defined as those with FPKM $\geq$ 0.5. For a certain expressed lincRNA with at least one histone modification signal $\geq$10, the H3K4me1/H3K4me3 count ratios in the TSS intervals were calculated. Furthermore, the read count for these two histone modifications were permutated, and random H3K4me1/H3K4me3 ratios were calculated. This process was repeated 10000 times. LincRNAs with the observed H3K4me1/H3K4me3 ratio more than 95 or <5% rank in the random ratio distribution were considered as high confident lincRNAs and canonical lincRNA, respectively.

Construction of Feature Sets

Genomic Features

CGI (CpG Island) and 7 repeat elements of mouse annotations in mm9 version were collected from UCSC (Details are shown in **Table 1** and **Supplementary Table 1**). The coverage ratio of CGI and repeat elements for TSS intervals and gene body regions of

TABLE 1 | Genomic, epigenomic and regulatomic data used in the prediction model.

Features	Cell type	Scource	Data type
Histone modifications (Active modifications: H3k4me1, H3k4me2, H3k4me3, H3k9ac, H3k27ac, H3k36me3, H3k79me2, H3k64ac, H3k122ac, Repressive modifications: H3k9me3, H3k20me3, H3k27me3)	ESCs	GEO	ChIP-Seq
DNA methylation	ESCs	GEO	ChIP-Seq
Repeat elements (DNA, LINE, SINE, LTR, Low complexity, Satellite, Simple repeat)	–	UCSC mm9	Bed format
CpG Island	–	UCSC mm9	Bed format
TFs (Pol2, NelfA, Spt5, Ctr9, Smad1, E2f1, Tcfcp2l1, CTCF, Zfx, STAT3, Klf4, Esrrb, c-Myc, n-Myc, GFP, p300, Suz12, nanog, Oct4, Sox2, Smc1, Smc3, Med12, Med1, Nipbl)	ESCs	GEO	ChIP-Seq

lincRNAs in training set were calculated as the genomic features for model training.

Epigenomic Features

DNA methylation and 12 histone modifications for mouse ESCs were collected for predictive model training (Details are shown in **Table 1** and **Supplementary Table 1**). For DNA methylation, the average β value for the probes in TSS intervals and gene body regions were calculated, respectively. The coverage ratio of each histone modification was also be calculated by using the read mapped in TSS intervals and gene body regions of training lincRNAs. DNA methylation and histone modifications were used as the epigenomic features with the value range from 0 to 1.

Regulatory Features

We also collected 25 TF ChIP-Seq data of mouse ESCs (Details are shown in **Table 1** and **Supplementary Table 1**). And for the regulatory features, TF binding coverage ratio for TSS intervals and gene body regions were calculated by the reads mapped in the relevant regions, respectively.

Building the Predictive Model Based on Regularized Regression

The predictive model for elincRNAs was built based on regularization regression model. Regularization regression model could avoid the over-fitting by penalizing high-valued regression coefficients, and which is also known as the shrinkage method. It is important for suppressing over-fitting problems in machine learning. Moreover, there are many characteristic variables in the regression model that do not contribute to the response variables. Therefore, feature selection is required for the predictive model. The regularization regression function was as follows (Equation 1):

$$\sum_{i=1}^{n} y_i = \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} \tag{1}$$

Where y_i represents the i th inputted lincRNA. When $y_i > 0$, the lincRNA was considered as elincRNA; if $y_i < 0$, the linRNA was considered as canonical lincRNA.x_{ij} represents the j th feature of the i th inputted lincRNA (y_i), and β_j (j=1,…,p) represents the contribution degree of the j th feature for the inputted lincRNA. If $\beta_j > 0$, it means that the j th feature is contributing factor for elincRNAs. Otherwise, it means that the j th feature is a contributing factor for canonical lincRNAs.

The regularization algorithm uses a cyclical coordinate descent method to obtain an optimized objective function $L(\beta)$ by continuously optimizing each parameter and iterating until convergence. The objective function $L(\beta)$ (Equation 2) is defined as combination of loss function [Residual Sum of Squares (*RSS*), Equation 3] and penalty term $P_\alpha(\beta)$ (Equation 4), which both named as penalized residual sum of squares (PRSS).

$$L(\beta) = RRS + P_\alpha(\beta) \tag{2}$$

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{P} \beta_j x_{ij})^2 \tag{3}$$

$$P_\alpha(\beta) = \sum_{j=1}^{P} [\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|] \tag{4}$$

The objective function $L(\beta)$ is minimized to estimate the vector of regression coefficients β_j, j=1,…,P.

A norm, L1–norm (L1 norm regularization), L2–norm (L2 norm regularization) or the combination of L1-norm and L2-norm could be added to the loss function of the regularization method as the penalty term. When α =1, the penalty term L1-norm is added into loss function, and the regulated regression model is lasso regression with the objective function $L(\beta)$ as follows (Equation 5):

$$L^{lasso}(\beta_1, \ldots, \beta_P) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{P} \beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j| \tag{5}$$

If α=0, L2-norm penalty is added into loss function, and the regulated regression model is Ridge regression whose objective function $L(\beta)$ as follows (Equation 6):

$$L^{ridge}(\beta_1, \ldots, \beta_P) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{P} \beta_j x_{ij})^2 + \lambda_2 \sum_{j=1}^{P} \beta_j^2 \tag{6}$$

If $0 < \alpha < 1$, combination of L1- and L2- norm is added into RSS, and the regulated regression model is elastic net regression (Equation 7).

$$L^{elastic-net}(\beta_1, \ldots, \beta_P) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{P} \beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j| + \lambda_2 \sum_{j=1}^{P} \beta_j^2 \tag{7}$$

Through the addition of a norm penalty, the model could obtain the optimal process of solving the function, thus prevent the over-fitting phenomenon and perform the feature selection.

The best combination of contributing factors for elincRNA identification could be prioritized with the lowest mean squared error. In this study, the predictive model for elincRNAs based on regularization regression was performed by R package Glmnet (https://cran.r-project.org/). The selection of α value for regularized regression model was determined by the minimized the error value which generated by cross-validation.

Model Evaluation

The 10-fold cross-validation and independent testing set were used to estimate the robustness of the predictive model. Training set was divided into 10 equal sized subsets, and 9 subsets were used as training set for model building, while the remaining one subset was used as the validation data for testing the model. This process was repeated for 10 times, in which each single subset was used as the validation data. Moreover, the ROC (Receiver Operating Characteristic) curve and PR (Precision-Recall) curve were drawn, respectively. ROC curve was drawn by plotting the sensitivity (also named as true positive rate, or recall) against the 1-specificity (also known as false positive rate) at various threshold settings. The PR (Precision-Recall) curve was performed by measuring the precision (positive predictive value) against sensitivity at various threshold settings. The AUC (area under the curve) values of ROC curve and PR curve were calculated, which were used to estimate the classification effect of the model. The closer the AUC value is approached to 1, the better performance of the prediction model gets.

The Independent Testing Set

To collect a comprehensive enhancer list, the vista enhancers of human and mouse were obtained. The genomic sequences of human enhancers were aligned to mouse genome (mm9) using blat program with the threshold 0.85. All the annotations of mouse genomic and the sequence conversed enhancers were overlapped with mouse lincRNAs, expected for the lincRNAs which identified as the high confidence set. And the lincRNAs covering more than half of an enhancer were considered as the elincRNAs. In the same method, the EPD promoter annotations of human and mouse were collected, and the human promoter sequences were aligned into mouse genome (mm9). The collected promoter set was compared with the mouse lincRNA set, and the lincRNAs which were overlapped with more than half of the collected promoters were considered as the canonical lincRNAs. In total, 37 elincRNAs and 69 canonical lincRNAs were obtained and used as the testing set for assessing performance of predictive models.

Identify Enriched TF Regulations for elincRNAs and Canonical lincRNAs

TF is thought to regulate elincRNAs or canonical lincRNAs, if a certain TF motif is enriched in the corresponding TSS intervals. TF motif data was used to analyze specific transcriptional regulation of elincRNAs and canonical linRNAs. In total, 358 mouse TF binding motif PWMs (Position Weighted Matrix) were collected form HOCOMOCO (HOmo sapiens COmprehensive MOdel COllection) Mouse v11 CORE (Kulakovskiy et al., 2018). AME (Analysis of Motif Enrichment) was used to detect enriched motifs in the TSS regions of lincRNAs with the statistically significance by Fisher's exact test (Bailey et al., 2015). FIMO (Find Individual Motif Occurrences) was used to screen the given TF motifs occurred in the TSS regions of elincRNAs and canonical lincRNAs (Bailey et al., 2015). One elincRNA/canonical linRNA was considered to be regulated by a TF, if this TF motif occurred in the TSS interval of elincRNA/canonical lincRNA, and the ChIP-Seq peak of TF was also observed within elincRNA/canonical lincRNA.

RESULTS

Identifying High Confidence Sets of elincRNAs and Canonical lincRNAs

It is widely acknowledged that, H3K4me1 and H3K4me3 are well-known active chromatin markers for enhancers and promoters, respectively. Thus, the two histone modification markers H3K4me1 and H3K4me3 within TSS intervals of the lincRNAs were interrogated. The average profiles of H3K4me3 and H3K4me1 in TSS intervals of 4,157 annotated lincRNA transcripts in mouse ESCs were shown in **Figure 1A**, revealing that the lincRNA TSS intervals were enriched by H3K4me3 and H3K4me1 with the pattern of bimodal and unimodal distribution, respectively (**Figure 1A**). Further, H3K4me1 and H3K4me3 intensities for lincRNA TSS intervals were investigated and shown in **Supplementary Figures 1A,B**. Unlike those lincRNAs displayed the mRNA-like promoter histone signatures, there was another lincRNA subset marked high H3K4me1 and low H3K4me3, which were enhancer-like histone signatures (**Supplementary Figures 1A,B**). H3K4me1 and H3K4me3 modification tags were counted in TSS intervals of these lincRNAs. The values of H3K4me1/H3K4me3 ratio were calculated, and the results showed that, more than 50% lincRNA TSS intervals were modified with low H3K4me1/H3K4me3 ratio (**Figure 1B**), which was consistent with the mRNA-like promoter feature. In addition, 27.41% lincRNAs were modified with high H3K4me1 (H3K4me1/H3K4me3 $\geq$ 2), which was enhancer signature. More interesting, through analyzing 1284 known lincRNAs with FPKM $\geq$ 0.5 in mouse ESCs by the LOESS (local polynomial regression) method, we draw the conclusion that the lincRNA expression levels were related with both H3K4me1 and H3K4me3 (**Figure 1C**). Particularly, canonical lincRNA expression levels might be higher than elincRNAs (**Figure 1C**). The results above were revealed that, the expressed lincRNAs were associated with the Histone modification H3K4me1 or H3K4me3 enriched in their TSS intervals.

LincRNAs could be divided into enhancer-associated and canonical lincRNAs by chromatin signatures H3K4me1 and H3K4me3 of TSS intervals (Signal et al., 2016). The reads of the two histone modifications in the 1,284 expressed lincRNA TSS intervals were counted for the subsequence analysis (Details were shown in Methods). As the result shown that, a bimodal distribution was presented for H3K4me1/H3K4me3 ratio for the lincRNAs, which was consistent with the results above (**Figure 1D**). It was revealed that expressed lincRNAs were associated with both H3K4me3 and H3K4me1. For

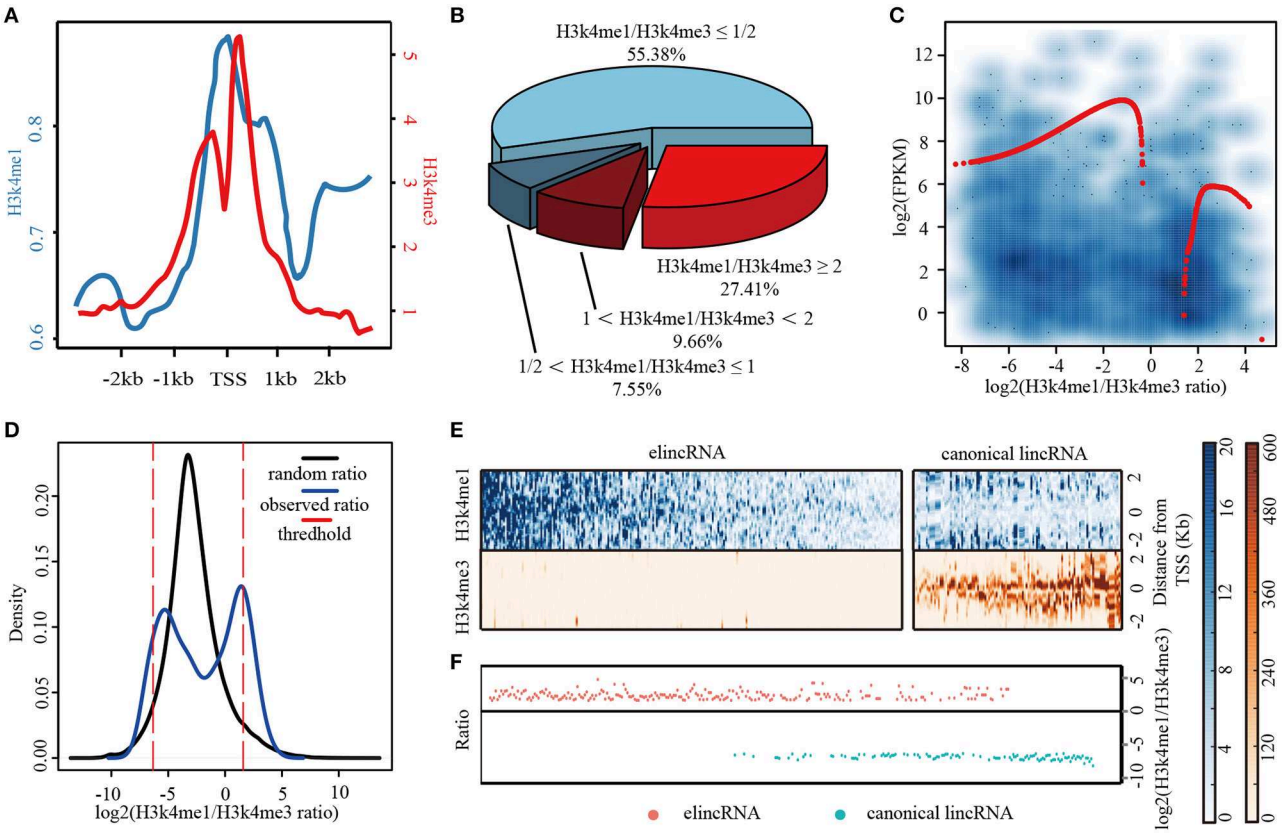


FIGURE 1 | Identification of elincRNAs and canonical lincRNAs with high confidence. **(A)** Density plots showing the distributions of H3K4me1 and H3K4me3 in known lincRNA TSS intervals. **(B)** Analysis of the H3K4me1/H3K4me3 ratio and the corresponding lincRNAs **(C)** Correlation analysis of the H3K4me1/H3K4me3 ratio and expression levels of the corresponding lincRNAs. **(D)** The identification of elincRNAs and canonical lincRNAs. **(E)** Heatmap showing the distributions of H3K4me1 and H3K4me3 in TSS intervals of identified elincRNAs and canonical lincRNAs. **(F)** Dots plots showing the log2 (H3K4me1/H3K4me3) ratio in TSS intervals of identified elincRNAs and canonical lincRNAs.

identifying the high confidence sets of elincRNAs and canonical lincRNAs, the signal intensities of H3K4me1 and H3K4me3 for these expressed lincRNAs were permutated, and the random H3K4me1/H3K4me3 ratios were calculated. This process was repeated for 10,000 times, thus, the distribution curve of random H3K4me1/H3K4me3 ratios could be performed with the normal distribution (black line in **Figure 1D**). The ratio values with the rank of 95 and 5% in the random distribution were used as the thresholds for elincRNAs and canonical lincRNAs, respectively (red dotted lines in **Figure 1D**) (Details were shown in Methods). By this method, 224 elincRNAs and 112 canonical lincRNAs were identified as the high confidence sets. The chromatin signatures of elincRNAs and canonical lincRNAs were investigated, as expected that, elinRNAs enriched H3K4me1 and depleted of H3K4me3 in TSS intervals (**Figure 1E**), as well as, H3K4me1/H3K4me3 ratios were >2 (**Figure 1F**). On the contrary, the canonical lincRNA TSS intervals were marked with low H3K4me1 and high H3K4me3 whose ratio values were <0.5 (**Figures 1E,F**). The intensity profiles of H3K4me1 and H3K4me3 for elincRNA and canonical lincRNA, was consistent with the previous studies (**Supplementary Figures 1C,D**).

To estimating the high confident sets of elincRNAs and canonical lincRNAs, two data of chromatin states identified by chromHMM in mouse ESCs were obtained (Yue et al., 2014; Bogu et al., 2015). The high confident elincRNAs or canonical lincRNAs with the coverage more than 0.3 by the relevant chromatin states were considered as the overlapped elincRNAs or canonical lincRNAs, respectively. For the chromatin states of Bogu's research, the number of elincRNAs overlapped by enhancer-like chromatin states were 103 (45.98%), while the of canonical lincRNAs overlapped by promoter-like chromatin states were 109 (97.32%) (**Supplementary Figure 2A**). Further, the random overlap distributions of elincRNAs and canonical lincRNAs were also acquired, through overlapping the random genomic regions equally with the observed lincRNA transcripts, using the corresponding chromatin state regions in the same criterion. And this process was repeated 10,000 times for Bogu's chromatin state data. It was revealed that the observed overlapped numbers were far from the random overlapped distributions (**Supplementary Figure 2A**). For Yue's chromatin states data (Yue et al., 2014), elincRNAs and canonical lincRNAs were estimated by the same method. ElincRNAs and canonical

lincRNAs overlapped by enhancer- and promoter-like chromatin states were 121 (54.02%) and 112 (100.00%), respectively, which were also far from the random distributions in the same method (**Supplementary Figure 2B**). Thus, the results indicated that the sets of elincRNAs and canonical lincRNAs identified would be used as the high confidence sets for the subsequence analysis.

Constructing a Novel Approach to Identify elincRNAs by Integrating Multi-Omic Features

For comprehensively characterizing and identifying elincRNAs, we integrated multi-dimensional features to build predictive model for elincRNAs. Genomic, epigenetic and regulatory features were collected, including CGI (CpG Island), 7 types of repeat elements, DNA methylation, 12 histone modifications and 25 TFs derived from the public sources (Details were shown in **Table 1** and **Supplementary Table 1**). Regularization regression model was performed to acquire the best combination of features for predictive model (**Figure 2A**). The high confident sets of elincRNAs and canonical lincRNAs were used as the positive and negative training set for identifying elincRNAs.

To examine the ability of different omic features to identify elincRNAs, we assessed the performance of predictive models combining distinct features, including sequence features (CGI and repeat elements), epigenomic features and regulatory features. Based on regularized regression model, the predictive models constructed by genomic, epigenomic or regulatory features alone were performed well. The results of 10-fold cross validation for each predictive model were shown in **Supplementary Table 3**, separately. When 25 TFs were used as the model features, the accuracy was 0.897, which was the lowest value in all of these models. The AUC values of ROC and PR curves were 0.931 and 0.959, respectively (**Supplementary Table 3** and **Figures 2B,C**). The accuracy of the predictive model based on CGI and repeats features was 0.923, which was slightly higher than TFs' model (**Supplementary Table 3** and **Figures 2B,C**). We observed that the model constructed by histone modifications alone and the combination with DNA methylation achieved the accuracies of 0.989 and 0.997, respectively. And the AUC values of ROC and PR curves were all more than 0.99 (**Supplementary Table 3**). This indicated that the three different features could contribute to identifying elincRNAs. However, it was shown that the TFs might not be the crucial factors for identifying elincRNAs. It was a remarkable fact that, among the TF features, P300 (EP300) which was a co-activator binding to enhancers was not identified as the contributing factor for elincRNAs, possibly because P300 also could be enriched in active promoter regions (Heintzman et al., 2007). Furthermore, genomic characteristics performed better, and the predictive model with epigenomic features performed best (**Supplementary Table 3** and **Figures 2B,C**). Further, to build the predictive model, the different combination of omic features were considered. The model constructed by the combination of genomic features, histone modifications and TFs was performed almost as well as the histone modification alone (accuracy was 0.989) (**Supplementary Table 3**). Moreover, when removed the TF feature, the predictive model with the combination of genomic and histone modifications performed similarly, with the equal accuracy 0.989. By combining all the features, the performance of the predictive model was a little better. Particularly, the predictive model with histone modification and DNA methylation performed best with the highest accuracy and AUCs values (**Supplementary Table 3** and **Figures 2B,C**). This indicated that epigenomic features including DNA methylation and histone modifications might be crucial for identifying the elincRNAs.

An independent data was also used to estimate the performance of the models (**Figure 2A**). The vista enhancers of human and mouse were collected to obtain a comprehensive enhancer list (Details were shown in Methods). In total, 37 elincRNAs and 69 canonical lincRNAs were obtained and used as the testing set for assessing the performance of predictive models. For the testing set, the predictive models with the combined features performed better than those with different feature alone (**Supplementary Table 4**). Predictive models with the combinations of genomic, histone modifications or with addition of the regulatory features got the better performances, with the accuracies of 0.755 and 0.774 respectively (**Supplementary Table 4**). And their AUCs values of ROC and PR curves were also better than the models with the different omic feature alone (0.668 and 0.444 for model with TF features, 0.815 and 0.608 for model with sequence feature) (**Figures 2D,E**, and **Supplementary Table 4**). These results showed the advantages of the integration of multi-omic features. It is noteworthy that the predictive model with the combination of histone modifications and DNA methylation acquired the best performance. And the accuracy achieved 0.859, which was a little higher than the model with all features (the accuracy was 0.858) (**Supplementary Table 4**). The results revealed that epigenetic features were the crucial signatures for elincRNAs. In brief, by combining DNA methylation and histone modifications, the regularized regression model performed effectively, which represented a novel approach to identify the elincRNAs.

According to the above analysis, the regularized regression model combining the histone modifications and DNA methylation performed best. The appropriate parameter α was interrogated for the optimal predictive model with the combination of epigenetic features. And the results were shown in **Supplementary Figure 3**, using the model contributing above, when the parameter α of regularized regression model was equal to 1, it was lasso regression model whose parameter λ was equal to 0.0016 with lowest mean squared error 0.0070 (**Supplementary Figure 3A**). However, the mean squared errors of ElasticNet and Ridge models were all higher than lasso model (**Supplementary Figures 3B–D**). Thus, when $\alpha = 1$, lasso model with the DNA methylation and histone modifications was used as the most effective predictive model to identify elincRNAs.

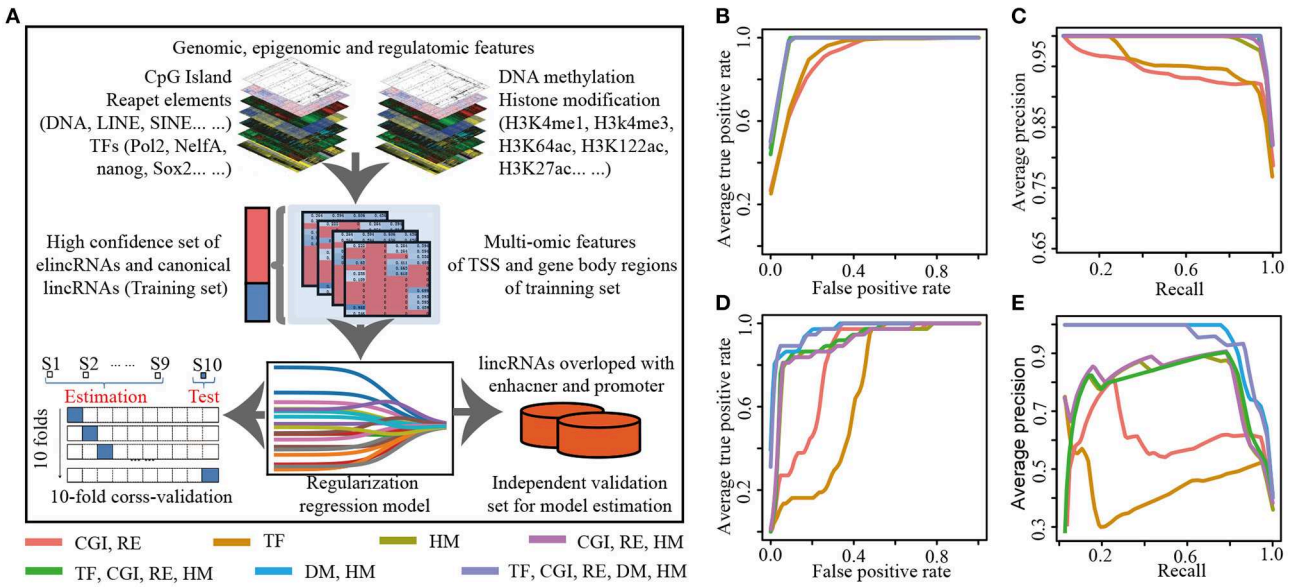


FIGURE 2 | The construction of regularization predictive model **(A)** Workflow of predictive model for identifying elincRNAs. **(B,C)** The average ROC **(B)** and PR **(C)** curves for 10-fold cross validation of the predictive models with different features. **(D,E)** The ROC **(D)** and PR **(E)** curves for the testing set of the predictive models with different features.

Both Histone Modifications and DNA Methylation Are Important Features

Biased on lasso model with the optimized parameter α, seven specific features among the 26 epigenetic features in TSS and gene body regions of lincRNAs, were identified for elincRNAs and canonical lincRNAs, respectively (**Figure 3A**). The regression equation, comprised of positive and negative regression coefficients, which represented the different contribution of features for identifying elincRNAs and canonical lincRNAs, respectively (**Figures 3A,B**) (Details were shown in Methods). The detailed regression equation was as follows:

$$y = 2.0860 + 5.79 \times x_1 + 3.16 \times x_2 + 0.53 \times x_3 + 0.46 \times x_4 - 5.17 \times x_5 - 4.13 \times x_6 - 0.87 \times x_7$$

In the equation, x_1, x_2, $x_3$and $x_4$represented TSS_DNA methylation, TSS_H3K4me1, Body_DNA methylation and Body_H3K122ac, respectively, which were contributing factors for elincRNAs. In addition, x_5, $x_6$and $x_7$represented Body_H3K36me3, TSS_H3K9ac and TSS_H3K4me3, which were contributing factors for canonical lincRNAs.

Further, based on the regression coefficients, we interrogated the contribution of the identified features for the predictive model. ElincRNAs were positively relative with TSS_methylation, TSS_H3K4me1, Body_methylation and Body_H3K122ac features (**Figures 3A,B**). As expected, H3K4me1 in TSS regions was the significant signature to predict elincRNAs whose contribution rate was 3.16. However, the greatest contribution feature for identifying elincRNAs was TSS_methylation with contribution rate 5.79 (**Figure 3B**). The results showed that DNA methylation of the TSS intervals contributed more to elincRNAs identification than H3K4me1 of the TSS intervals. The body_methylation feature could also be the predictor for elincRNAs, with the contribution rate of 0.53 (**Figure 3B**). Moreover, the DNA methylation signal intensities of TSS and body regions of elincRNAs and canonical lincRNAs were further compared. Indeed, the average DNA methylation levels of elincRNAs were significantly higher than canonical lincRNAs (**Figures 3C,D**), which was corresponding with previous results. In the study of Kundaje et al., they showed that the average DNA methylation of active enhancers was significantly higher than that of active TSSs (Roadmap Epigenomics et al., 2015). DNA methylation was enriched in both TSS and body regions of elincRNAs. In TSS and body regions of training and testing sets, the expression levels of elincRNAs were significantly lower than that of canonical lincRNAs, in both training and testing sets (**Figure 3E**), which was consistent with the expectations. The above results indicated that DNA methylation might be a crucial signature for elincRNAs identification. In addition, histone modification body_H3K122ac, was also identified as the lincRNA-related feature with value of the contribution rate of 0.46 (**Figure 3B**), which was consistent with recent researches that H3K122ac could mark active enhancers (Pradeepa et al., 2016). However, H3K27ac was not identified as the predictor of elincRNAs in the feature selection process of the predictive model. In the predictive model with histone modifications alone, TSS_H3K27ac was recognized as the significant marker for elincRNAs with the contribution rate of 0.20, which was higher than TSS_H3K122ac (contribution rate of 0.10), but lower than body_H3K122ac (contribution rate of 2.34) and TSS_H3K4me1 (contribution rate of 3.39) (**Supplementary Figure 4**). This was in accordance with Pradeepa's study that a set of active enhancers was uncovered which was marked by H3K122ac

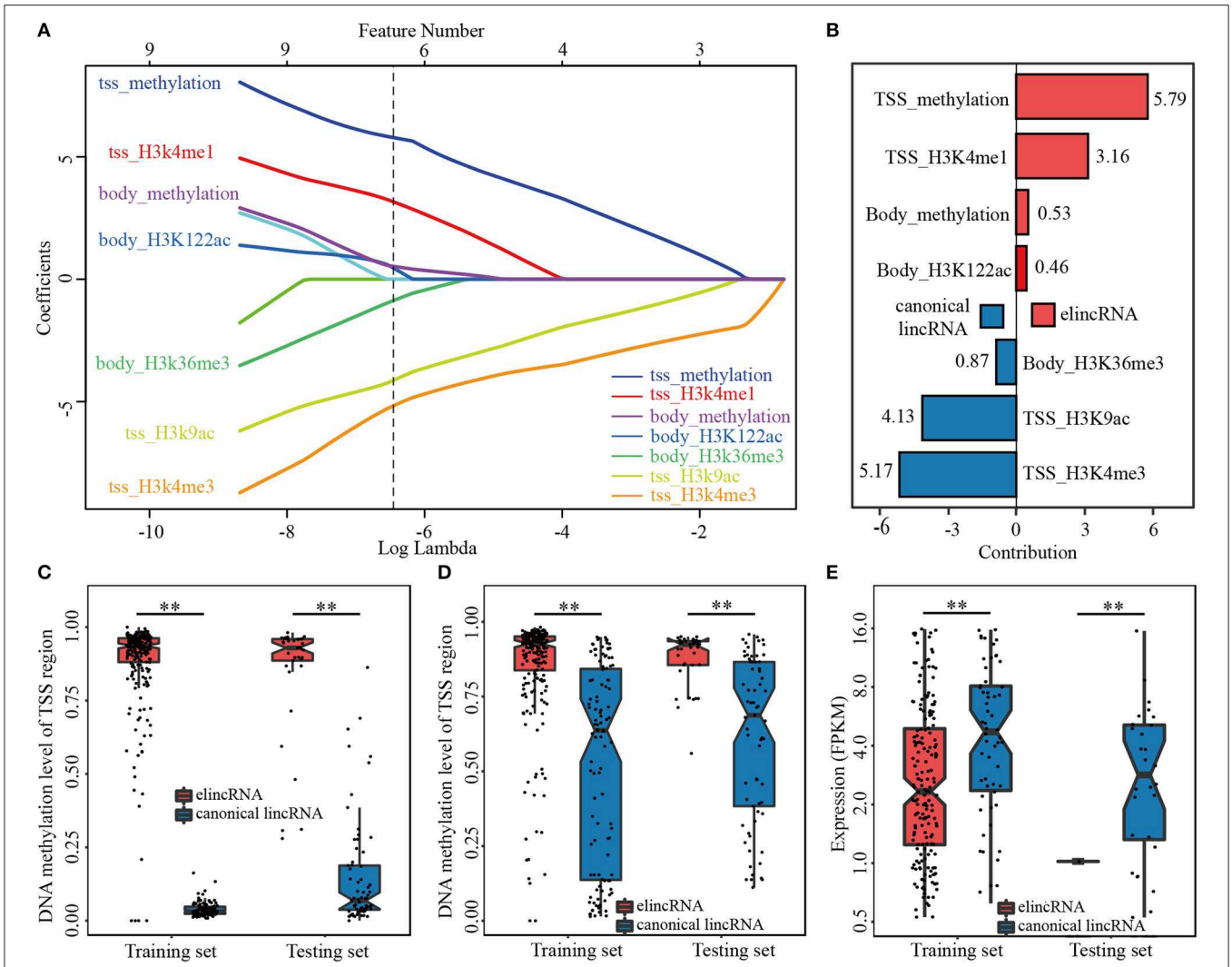


FIGURE 3 | The predictive model for elincRNA identification with the specific features. **(A)** Feature selection of the predictive model for elincRNA identification based on epigenetic features. **(B)** The contribution rates of the identified epigenetic features of predictive model based on the regression coefficients. **(C,D)** DNA methylation levels in the TSS **(C)** and body **(D)** regions of elincRNAs and canonical lincRNAs of training set and testing set in predictive model. **(E)** Expression levels of elincRNAs and canonical lincRNAs of training set and testing set in predictive model. $^{**}P < 0.01$.

but lack H3K27ac (Pradeepa et al., 2016). Further, the active enhancers marked by H3K27ac were also enriched with H3K122ac (Pradeepa et al., 2016). It was suggested that the predict efficiency of H3K27ac for elincRNAs might be less than TSS-enriched DNA methylation and body-enriched H3K122ac. Thus, the combination of TSS_methylation, TSS_H3K4me1, Body_methylation and Body_H3K122ac could be the significant signatures for elincRNAs.

On the other hand, canonical lincRNAs were relevant with TSS_H3K4me3, TSS_H3K9ac and Body_H3K36me3. As the canonical marker for promoters, TSS_H3K4me3 still had the great significant contribution for the identification of canonical lincRNAs with the contribution rate of 5.17 (**Figure 3B**). TSS_H3K9ac was also identified as the characteristic feature for canonical lincRNAs with the contribution value 4.13 (**Figure 3B**), which was an active regulatory marker of with preference for promoters (Consortium, 2012; Roadmap Epigenomics et al., 2015). Moreover, body_H3K36me3, which was an Pol II elongation marker associated with transcribed portions of active genes (Consortium, 2012), was also remarkable for canonical lincRNAs with contribution rate of 0.87 (**Figure 3B**). In addition to that, H3K36me3 was also found to be associated with active enhancers and be likely to correlate with enhancer RNA transcription (Consortium, 2012), however, H3K36me3 didn't be identified as the effective feature to mark elincRNAs. Therefore, through the regularization regression model, new features were identified to mark elinRNAs (TSS_methylation, Body_methylation and Body_H3K122ac), besides H3K4me1. It could characterize elinRNAs better, and could be helpful for identifying elincRNAs.

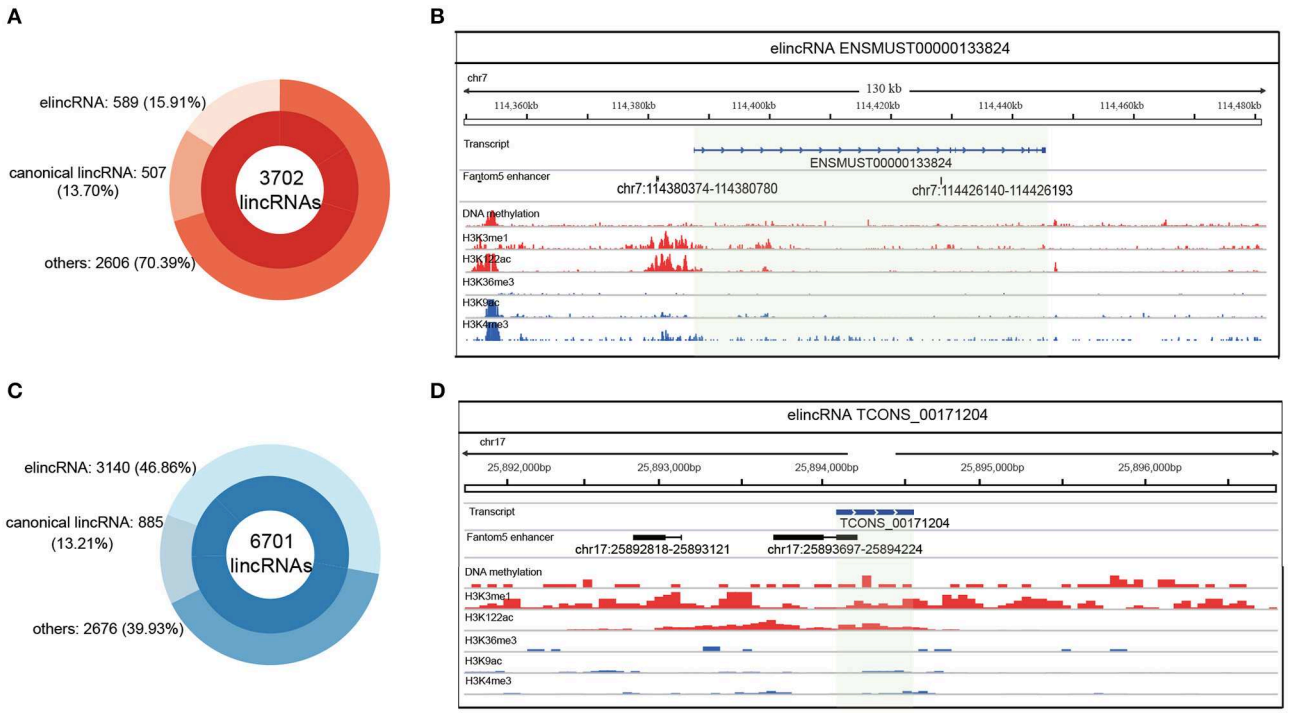


FIGURE 4 | Identification of elincRNAs based on the predictive model with specific epigenetic features. **(A)** Identification of elincRNAs from known lincRNA set by the predictive model. **(B)** Identification of elincRNAs from predicted lincRNA set expressed in mouse ESCs by the predictive model. **(C,D)** Examples for elincRNAs ENSMUST00000133824 **(C)** from known lincRNA set used as training set, and TCONS_00171204 **(D)** from predicted lincRNA set expressed in mouse ESCs identified by predictive model. **(C,D)** Were produced using IGV (Integrative Genomics Viewer), and the green transparent shadows represent the transcript regions of the examples.

Prediction of elincRNAs Based on Regularization Regression Model

For distinguishing more annotations of elincRNAs from canonical lincRNAs, elincRNAs and canonical lincRNAs were identified, using lasso model with the identified epigenetic features. Firstly, among the 3702 known lincRNAs which were not used as training and testing sets for predictive model, 589 elincRNAs and 507 canonical lincRNAs were identified by using the identified epigenetic features (**Figure 4A**). Moreover, the predictive model was applied to another lincRNA set, which contained 6701 putative lincRNAs in mouse ESCs that we previously identified (Liu et al., 2016). As result, 3140 elincRNAs and 885 canonical lincRNAs were identified (**Figure 4B**).

An identified elincRNA ENSMUST00000192129 sourced from known lincRNA set, which was located in reverse strand of Chr 1. And the specific epigenetic features were shown in **Supplementary Figure 5A**. Although H3K4me1 in TSS regions was poor, H3K122ac and DNA methtylation were enriched in the body region with no canonical lincRNA' features. The overlapped transcript ENSMUST00000192129 was an elincRNA identified in training set with the representative features of elincRNAs (**Supplementary Figure 5A**). Another elincRNA ENSMUST00000133824 used for model training was shown in **Figure 4C**, which was marked by the elincRNA specific features, and covered with a FANTOM5 annotated enhancer. A putative lincRNA TCONS_00171204 expressed in mouse ESC was identified as an elincRNA with only one exon (**Figure 4D**). This elinRNA was overlapped with an annotated enhancer with significant epigenetic features (**Figure 4D**). To mark a contrast with the elincRNAs, a canonical lincRNA ENSMUST00000180932 used for model training was shown in **Supplementary Figure 5B**, and this lincRNA was marked with high H2K9ac and H3K4me3 in TSS region with de H3K36me3 covered in body region. Thus, we obtained a relative comprehensive elincRNA set for mouse ESCs.

Characterization of elincRNAs and Canonical lincRNAs

The results above revealed that elincRNAs could be distinguished from canonical lincRNAs by specific epigenomic features. We detected that if there are any other different features between elincRNAs and canonical lincRNAs. By combining the high confident and predict elincRNAs and canonical lincRNAs, 3990 elincRNAs and 1573 canonical lincRNAs were collected (**Figures 5A,B**). And then, the elincRNAs and canonical lincRNAs were compared and characterized from various aspects. Firstly, the comparison of transcript length for elincRNAs and canonical lincRNAs were performed and the median values for elincRNAs and canonical lincRNAs were 1490.433 and 1550.179, respectively (**Supplementary Figure 6A**). It was indicated that

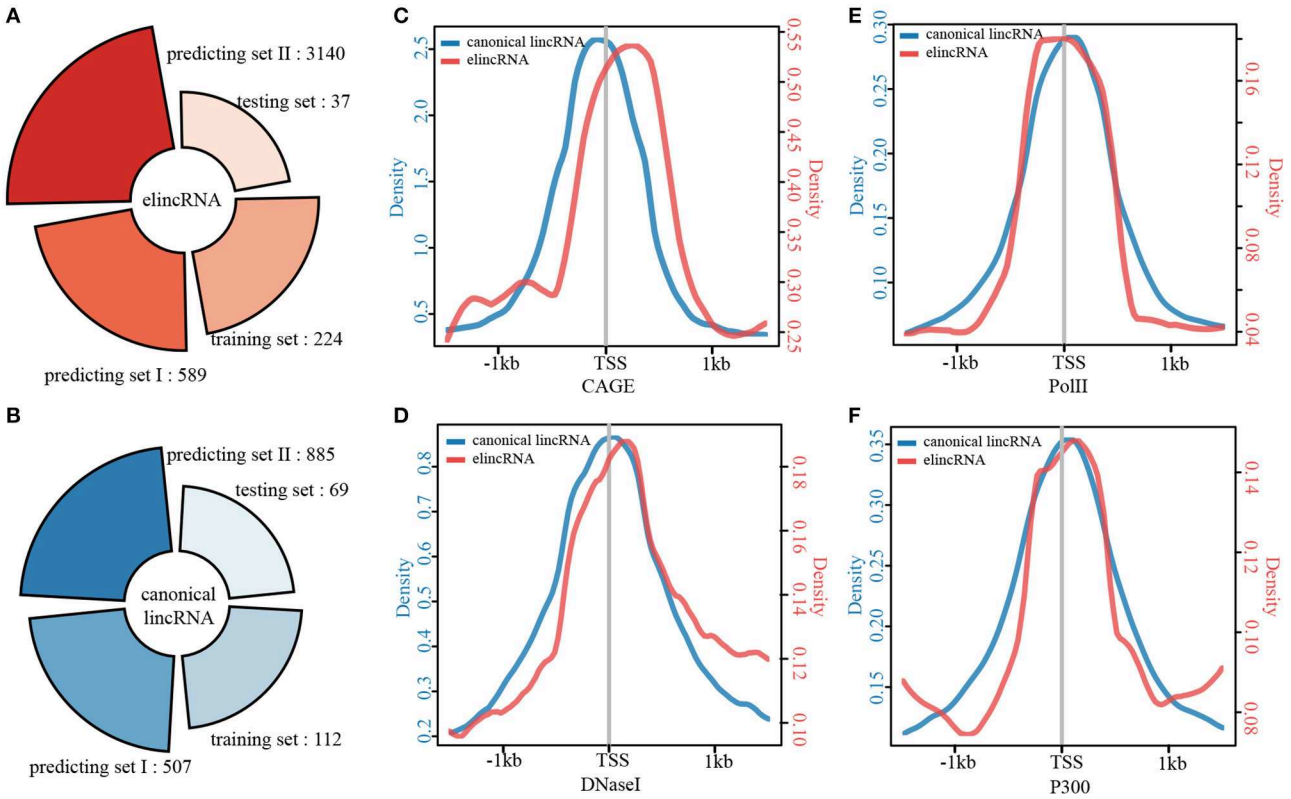


FIGURE 5 | Characterization of elincRNAs and canonical linRNAs. **(A)** ElincRNA set, and **(B)** Canonical lincRNA set, including training set, testing set, and two predicting sets. **(C–F)** The average profiles of CAGE reads **(C)**, DNaseI signals **(D)**, Pol II signals **(E)**, and P300 signals **(F)** around the TSS regions of elincRNA and canonical lincRNA sets.

the length of elincRNA transcripts was shorter than that of canonical lincRNAs with statistical significance (KS test p-value $< 2.2E\text{-}16$). Although with no statistical significance, the expression level of elincRNAs was a little lower than that of canonical lincRNAs (**Supplementary Figure 6B**), whose median values of expression levels were 6.544 and 7.134, respectively.

The genomic features were investigated for elincRNAs and canonical lincRNAs in the following analysis. It has been confirmed that enhancers are generally less conserved (Li et al., 2016), and the conservation of lincRNAs is also significantly lower than protein-encoding transcripts (Derrien et al., 2012). Therefore, it is necessary to interrogate the sequence conservation of exon regions for elincRNAs and canonical lincRNAs. The cumulative probability density distributions of the average conservation for the exon sequences of elincRNA and canonical lincRNA sets were shown in **Supplementary Figure 6C**. The conservation score of elincRNAs was significantly lower than that of canonical lincRNAs, whose median values were 0.036 and 0.064, respectively. Further, the coverage of CGIs (CpG Islands) in the TSS regions of these two lincRNA sets were compared, and the results were shown in **Supplementary Figure 6D**. It was found that, canonical lincRNA TSS intervals enriched CGIs with the median coverage 0.237, and elincRNA TSS intervals lacked the coverage of CGIs whose median coverage was only 0.002. Moreover, the CGI coverage of elincRNA TSS regions was much lower than canonical lincRNAs.

The average profiles of CAGE and chromatin modifications in TSS intervals for elincRNAs and canonical lincRNAs were compared. Although the CAGE intensity was significantly lower than that of the canonical lincRNA TSS regions, it still existed a peak in elincRNA TSS intervals (**Figure 5C**). DNase I hypersensitive sites are acknowledged to be characterized by open accessible chromatin. Thus, the DNaseI average profiles of the TSS regions for the two transcript sets were compared (**Figure 5D**). Similar results were obtained that elincRNAs and canonical lincRNAs showed a unimodal distribution in the TSS regions, but the enrichment of elincRNAs was significantly lower than that of canonical lincRNAs (**Figure 5D**). The phenomenon that the low abundances of the active signals in TSS intervals might be related to the universal low expression abundances of elincRNAs. Moreover, we compared PolII and P300 for TSS intervals of elincRNAs and canonical lincRNAs, and the results indicated that elincRNAs and canonical lincRNAs had distinct modification characteristics, respectively (**Figures 5E,F**).

Moreover, the coverage of repeat elements in TSS intervals were analyzed, including three retrotransposons (LINE, SINE and LTR) (**Supplementary Figures 7A–C**), DNA transposons (**Supplementary Figure 7D**), and three

tandom repeats (Satellite, Micro-satellite and Mini-Satellite) (**Supplementary Figures 7E–G**). The coverages of these seven repeat elements in TSS regions were significantly different between elincRNAs and canonical lincRNAs, indicating that the elincRNA TSS intervals might be enriched with repeat elements expect for mini-Satellite (**Supplementary Figures 7A–G**). Although CGI and repeat elements were not well predictors for identifying elincRNAs, both CGI and repeat elements were significantly different between the two lincRNA sets. We compared the elincRNAs with 680 dbSUPER annotated super-enhancers including 4343 constituents. The result showed that 273 elincRNAs were overlapped with 164 super-enhancers containing 426 constituents (**Supplementary Figure 7H**). Among the 3990 elincRNAs, 308 were overlapped with CAGE peaks or TSSs predicted by CAGE peaks of FANTOM in both forward and reverse strands, which might be bidirectional transcripts. Summarizing the results above, the elincRNAs and canonical lincRNAs exhibited distinctly specific transcript characteristics, sequence features and chromatin modification features.

ElincRNAs and Canonical lincRNAs Are Regulated by Distinct TF Regulatory Patterns

Since the distinct sequence features and chromatin modifications around TSS intervals of elincRNAs and canonical lincRNAs, we interrogated the TF regulatory functions on their TSS intervals. Based on known TF motifs obtained from the HOCOMOCO Mouse v11 CORE and JASPAR CORE (2018), the enrichment TF motifs in TSS intervals for elincRNAs and canonical lincRNAs were detected by using AME (Details were shown in Methods).

As a result, 52 TF motifs were enriched in elincRNA TSS intervals, while 90 motifs were enriched in canonical lincRNA TSS intervals with statistical significance (E-value < 0.05) (**Figure 6A**). There were four TFs named as NANOG (Loh et al., 2006), POU5F1 (OCT4) (Zhang X. et al., 2008), SOX2 (Kim et al., 2008) and ESRRB (Festuccia et al., 2018), which were the acknowledged proteins or regulators related to cell differentiation and embryonic development, were enriched in the elincRNA TSS regions. NANOG, POU5F1, and SOX2 were the core markers for stem cells, which were essential to maintain mouse embryonic stem cell pluripotency (Loh et al., 2006; Kim et al., 2008). They bound 97.01, 86.21, 98.22, and 97.47% elincRNAs with the statistical significance for NANOG, POU5F1, SOX2, and ESRRB (**Figure 6B**). And, KLF4 existed among the canonical lincRNA enriched TF sets with statistical significance, which was a critical regulator for cell reprogramming and early embryonic development in mouse (Ye et al., 2018).

Furthermore, only one TF motif ZN143 was common in both elincRNAs and canonical lincRNAs enriched TF sets (**Figure 6A**), indicating that elincRNAs and canonical lincRNAs had specific regulatory patterns by binding the distinct TF sets. By integrated the available ChIP-Seq data, we further identified the regulatory relations of the TFs for the elincRNAs and canonical linRNAs, and constructed regulatory network for elincRNAs and caonical lincRNAs, respectively. An elincRNA or a canonical lincRNA was considered to be regulated by a certain TF, if the TSS intervals existed the predicted binding sites and were covered by the corresponding ChIP-Seq peak. The elincRNA regulatory network comprised 662 elincRNAs (**Figure 6C**), while the canonical lincRNA regulatory network comprised 567 canonical lincRNAs (**Figure 6D**).

Moreover, function enrichment results found that elincRNAs and canonical lincRNAs were usually involved in the different biological processes, via GREAT method based on the GO BP annotation (**Figure 6E**, FDR$<$0.05). For these elincRNAs regulated by ESC markers, they are significantly involved in well-known functions related with stemness maintenance and cell differentiation, including ormation of primary germ layer, gastrulation, the morphogenesis, formation and development of mesoderm, endoderm formation, stem cell differentiation and so on. The elincRNAs were also enriched in several GO BP terms related with Notch signaling pathway, which have been discovered contributing to the formation, growth (Rowan et al., 2008), and development of embryos (Rowan et al., 2008; Fernandez-Valdivia et al., 2011; Djabrayan et al., 2012), even could play crucial functions in the embryonic cell differentiation (Ben-Shushan et al., 2015). While, the canonical lincRNAs regulated by KLF4 tended to regulate basic biological functions, such as nucleus organization, posttranscriptional regulation of gene expression, which were also important for life function maintenance (**Figure 6E**). Summarizing the results above, elincRNAs and canonical lincRNAs had the specific regulatory patterns, and elincRNAs might be involved in development specific biological processes while canonical lincRNAs played the basic biological functions.

DISCUSSION

In this study, we developed a novel approach to identify elincRNAs by integrating multi-omic data. We first revealed that expressed lincRNAs could be marked by two common active chromatin modifications H3K4me1 and H3K4me3. And then, several epigenetic features were identified as the signatures for elincRNAs and canonical lincRNAs by lasso regression model. Besides the common acknowledged features H3K4me1 and H3K4me3, more specific features were recognized in our predictive model. For example, DNA methylation and H3K122ac could be the novel signatures to mark elincRNAs, and H3K9ac and H3K36me3 could be the makers for canonical lincRNAs. Unexpectedly, DNA methylation contributed much more than H3K4me1 in the TSS intervals for mouse ESCs. TSS_ and body_DNA methylation were both significant features for elincRNAs, which was consistent with the previous study that the DNA methylation level of active enhancers were commonly higher than active promoters (Roadmap Epigenomics et al., 2015). And in Charlet's study, it was shown that DNA methylation could co-exist with H3K27ac at enhancers and super-enhancers, but not at promoters (Charlet et al., 2016). DNA methylation play important roles in cell differentiation, embryonic development and complex diseases (Su et al., 2018; Yu et al., 2019). Thus, DNA methylation might play unexpected

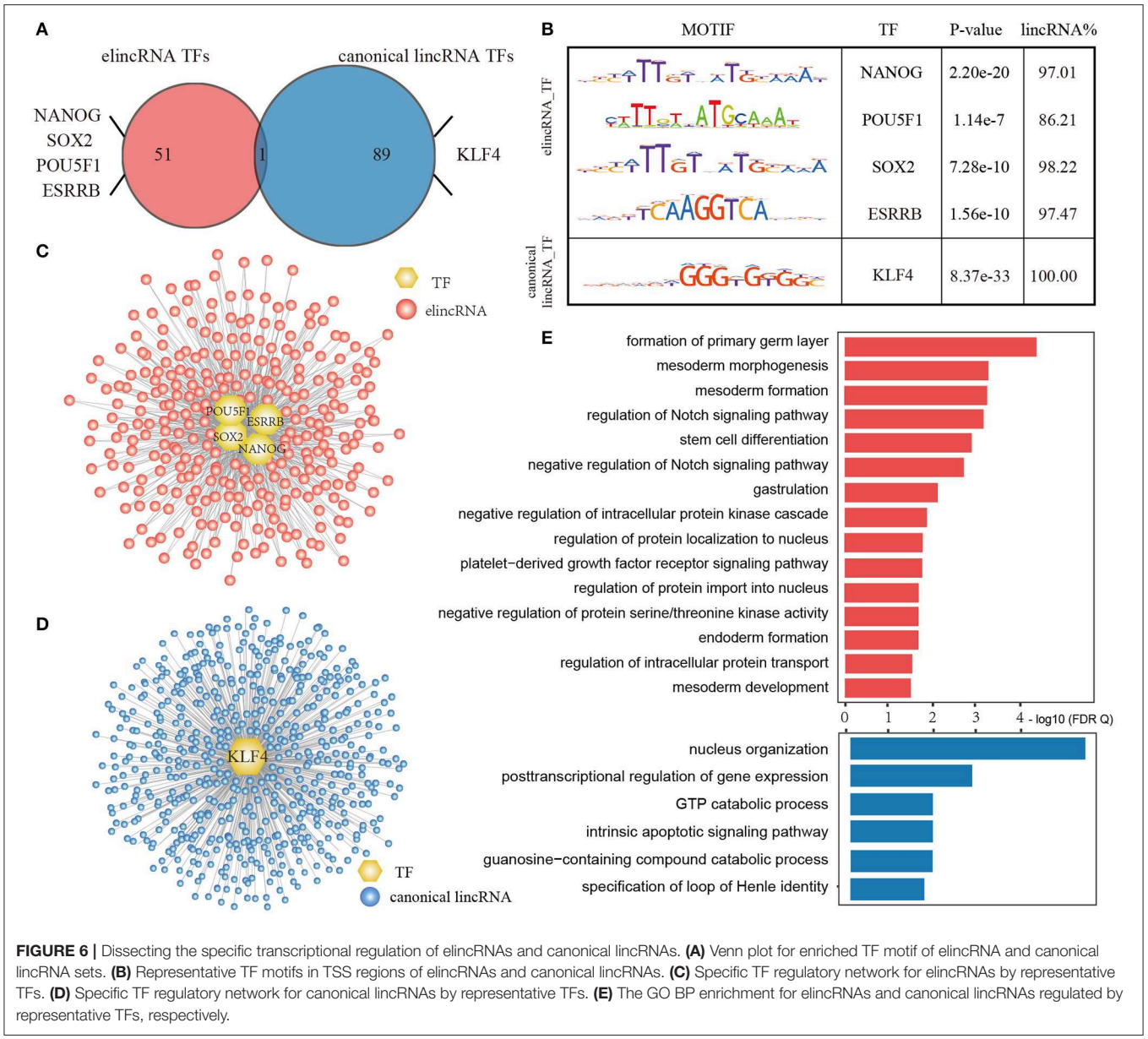


FIGURE 6 | Dissecting the specific transcriptional regulation of elincRNAs and canonical lincRNAs. **(A)** Venn plot for enriched TF motif of elincRNA and canonical lincRNA sets. **(B)** Representative TF motifs in TSS regions of elincRNAs and canonical lincRNAs. **(C)** Specific TF regulatory network for elincRNAs by representative TFs. **(D)** Specific TF regulatory network for canonical lincRNAs by representative TFs. **(E)** The GO BP enrichment for elincRNAs and canonical lincRNAs regulated by representative TFs, respectively.

roles at enhancer regions. In our predictive model, H3K27ac was not identified as the significant feature to mark elincRNAs, However, H3K122ac was recognized to enriched in the body regions of expressed elincRNAs, which was consistent with the previous study (Pradeepa et al., 2016). In Pradeepa's study, H3K122ac was identified as a novel signature for active enhancers which were enriched with H3K27ac, and also could mark a subset of active enhancers without H3K27ac enriched (Pradeepa et al., 2016). Thus, H3K122ac might play the important roles in elincRNAs. However, when removing DNA methylation from the predictive model, TSS_H3k27ac could be identified as the contributing factor for elincRNA, despite its contributing score (coefficient) was less than H3K4me1 and H3K122ac. Thus, the combination of features TSS_methylation, TSS_H3K4me1, Body_DNA methylation and Body_H3K122ac enrichment could be the effective markers for elincRNAs. Therefore, the signature sets for elincRNAs and canonical lincRNAs were much more complemented and perfected, than the features H3K4me1 and H3K4m3 alone.

When constructing the predictive model, we used the high confident elincRNA sets identified by the H3K4me1/H3K4me3 ratios. Through the predictive model with identified epigenetic features, we identified 589 expressed elincRNAs from the known lincRNAs, and identified 3140 expressed elincRNAs from the expressed lincRNA set in mouse ESCs (Liu et al., 2016). Along with 224 elincRNAs in training set and 37 elincRNAs in testing set, total 3990 elincRNAs were collected in mouse ESCs. When interrogated the FANTOM5 annotated enhancers, 1179 enhancers were found covered by our identified elincRNAs. We also compared the 3990 elincRNAs with the 147 lncRNAs

that were demonstrated to affect the global gene expression in Guttman's study (Guttman et al., 2011), and 48 lncRNAs could be covered by our identified elincRNAs, including 36 known lincRNAs and 58 predicted lincRNAs. Thus, the atlas of elincRNAs depicted by the predictive model provided essential for insight into the regulatory function roles of elincRNAs during the embryonic development.

In the analysis for the specific regulatory patterns of elincRNAs and canonical licnRNAs, we found that, there were specific TF motifs enriched in elincRNAs or canonical lincRNA TSS intervals. Further, the regulatory relationship of several specific enriched TFs were validated by the corresponding ChIP-Seq peaks in mouse ESCs, including NANOG, POU5F1, SOX2 and ESRRB, which were crucial regulators in circuitry controlling ES cell state (Young, 2011). Thus, elincRNAs might be essential component of the TF regulatory circuitry, which were involved in the key regulatory functions of stem cells. Further, the analysis of the elincRNAs and canonical lincRNAs regulated by the specific TFs showed that elincRNAs tended to be involved in the biological processes related with cell differentiation and embryonic development. This indicated that elincRNAs might play the crucial roles in mouse embryonic development.

In conclusion, this work provides a novel approach to identify elincRNAs and canonical lincRNAs by combination of genomic, epigenomic and regulatomic features based on the regularization regression model. Specific epigenetic features were recognized to mark elincRNAs and canonical lincRNAs, respectively. This would help to supplement and improve the atlas of elincRNAs, and dissect the crucial roles of elincRNAs in mouse embryonic development and complex diseases.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found in the Gene Expression Omnibus, accession numbers: GSE39619, GSE11172, GSE12241, GSE66023, GSE40951, GSE31039, GSE30202, GSE20530, GSE11431, GSE22562.

AUTHOR CONTRIBUTIONS

QioW and JX conceived and designed the experiments. HL, TJ, SW, and XC acquired the experiment data. HL, XJ, QiW, and XL performed the study. HL and JY carried out the data analysis. HL wrote this manuscript. JX, YL, and TS revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China [61803130, 61801150, 61873075, 31871338, 31771601, 31571331, 31970646, 31601065]; the National Key R&D Program of China (2018YFC2000100); Natural Science Foundation for Distinguished Young Scholars of Heilongjiang Province [JQ2019C004]; and University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province [UNPYSCT-2016189, UNPYSCT-2017060].

ACKNOWLEDGMENTS

We thank the ENCODE consortia and FANTOM for generating and sharing the data used in this paper. We thank National Natural Science Foundation of China for funding.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00427/full#supplementary-material

REFERENCES

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME suite. *Nucleic Acids Res.* 43, W39–W49. doi: 10.1093/nar/gkv416

Ben-Shushan, E., Feldman, E., and Reubinoff, B. E. (2015). Notch signaling regulates motor neuron differentiation of human embryonic stem cells. *Stem Cells* 33, 403–415. doi: 10.1002/stem.1873

Bogu, G. K., Vizan, P., Stanton, L. W., Beato, M., Di Croce, L., and Marti-Renom, M. A. (2015). Chromatin and RNA maps reveal regulatory long noncoding RNAs in mouse. *Mol. Cell Biol.* 36, 809–819. doi: 10.1128/MCB.00955-15

Bose, D. A., Donahue, G., Reinberg, D., Shiekhattar, R., Bonasio, R., and Berger, S. L. (2017). RNA binding to CBP stimulates HISTONE Acetylation and transcription. *Cell* 168, 135–149. e22. doi: 10.1016/j.cell.2016.12.020

Catarino, R. R., and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.* 32, 202–223. doi: 10.1101/gad.310367.117

Charlet, J., Duymich, C. E., Lay, F. D., Mundbjerg, K., Dalsgaard Sorensen, K., Liang, G., et al. (2016). Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. *Mol. Cell* 62, 422–431. doi: 10.1016/j.molcel.2016.03.033

Cinghu, S., Yang, P., Kosak, J. P., Conway, A. E., Kumar, D., Oldfield, A. J., et al. (2017). Intragenic enhancers attenuate host gene expression. *Mol. Cell* 68, 104–117. e6. doi: 10.1016/j.molcel.2017.09.010

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Dinger, M. E., Amaral, P. P., Mercer, T. R., Pang, K. C., Bruce, S. J., Gardiner, B. B., et al. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* 18, 1433–1445. doi: 10.1101/gr.078378.108

Djabrayan, N. J., Dudley, N. R., Sommermann, E. M., and Rothman, J. H. (2012). Essential role for notch signaling in restricting developmental plasticity. *Genes Dev.* 26, 2386–2391. doi: 10.1101/gad.199588.112

Dreos, R., Ambrosini, G., Groux, R., Cavin Perier, R., and Bucher, P. (2017). The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.* 45, D51–D55. doi: 10.1093/nar/gkw1069

Emera, D., Yin, J., Reilly, S. K., Gockley, J., and Noonan, J. P. (2016). Origin and evolution of developmental enhancers in the mammalian neocortex. *Proc. Natl. Acad. Sci. U.S.A.* 113, E2617–E2626. doi: 10.1073/pnas.1603718113

Fernandez-Valdivia, R., Takeuchi, H., Samarghandi, A., Lopez, M., Leonardi, J., Haltiwanger, R. S., et al. (2011). Regulation of mammalian notch signaling and embryonic development by the protein O-glucosyltransferase rumi. *Development* 138, 1925–1934. doi: 10.1242/dev.060020

Festuccia, N., Owens, N., and Navarro, P. (2018). Esrrb, an estrogen-related receptor involved in early development, pluripotency, and reprogramming. *FEBS Lett.* 592, 852–877. doi: 10.1002/1873-3468.12826

Fleischer, T., Tekpli, X., Mathelier, A., Wang, S., Nebdal, D., Dhakal, H. P., et al. (2017). DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.* 8:1379. doi: 10.1038/s41467-017-00510-x

Flynn, R. A., and Chang, H. Y. (2014). Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell* 14, 752–761. doi: 10.1016/j.stem.2014.05.014

Gong, C., and Maquat, L. E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via alu elements. *Nature* 470, 284–288. doi: 10.1038/nature09701

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477, 295–300. doi: 10.1038/nature10398

Hamazaki, N., Uesaka, M., Nakashima, K., Agata, K., and Imamura, T. (2015). Gene activation-associated long noncoding RNAs function in mouse preimplantation development. *Development* 142, 910–920. doi: 10.1242/dev.116996

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318. doi: 10.1038/ng1966

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., et al. (2013). Super-enhancers in the control of cell identity and disease. *Cell* 155, 934–947. doi: 10.1016/j.cell.2013.09.053

Hon, C. C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* 543, 199–204. doi: 10.1038/nature21374

Ilott, N. E., Heward, J. A., Roux, B., Tsitsiou, E., Fenwick, P. S., Lenzi, L., et al. (2014). Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nat. Commun.* 5:3979. doi: 10.1038/ncomms4979

Karnuta, J. M., and Scacheri, P. C. (2018). Enhancers: bridging the gap between gene control and human disease. *Hum. Mol. Genet.* 27, R219–R227. doi: 10.1093/hmg/ddy167

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* 132, 1049–1061. doi: 10.1016/j.cell.2008.02.039

Kim, T. K., Hemberg, M., and Gray, J. M. (2015). Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harb. Perspect. Biol.* 7:a018622. doi: 10.1101/cshperspect.a018622

King, A. D., Huang, K., Rubbi, L., Liu, S., Wang, C. Y., Wang, Y., et al. (2016). Reversible regulation of promoter and enhancer histone landscape by DNA methylation in mouse embryonic stem cells. *Cell Rep.* 17, 289–302. doi: 10.1016/j.celrep.2016.08.083

Krueger, F., and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* 27, 1571–1572. doi: 10.1093/bioinformatics/btr167

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 46, D252–D259. doi: 10.1093/nar/gkx1106

Lam, M. T., Li, W., Rosenfeld, M. G., and Glass, C. K. (2014). Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* 39, 170–182. doi: 10.1016/j.tibs.2014.02.007

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lee, J. T., and Bartolomei, M. S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* 152, 1308–1323. doi: 10.1016/j.cell.2013.02.016

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

Liu, H., Lyu, J., Liu, H., Gao, Y., Guo, J., He, H., et al. (2016). Computational identification of putative lincRNAs in mouse embryonic stem cell. *Sci. Rep.* 6:34892. doi: 10.1038/srep34892

Loh, Y. H., and Shen, L. (2016). Analysis and visualization of ChIP-seq and RNA-seq sequence alignments using ngs.plot. *Methods Mol. Biol.* 1415, 371–383. doi: 10.1007/978-1-4939-3572-7_18

Loh, Y. H., Wu, Q., Chew, J. L., Vega, V. B., Zhang, W., Chen, X., et al. (2006). The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* 38, 431–440. doi: 10.1038/ng1760

Long, H. K., Prescott, S. L., and Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* 167, 1170–1187. doi: 10.1016/j.cell.2016.09.018

Mousavi, K., Zare, H., Dell'orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., et al. (2013). eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol. Cell* 51, 606–617. doi: 10.1016/j.molcel.2013.07.022

Natoli, G., and Andrau, J. C. (2012). Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* 46, 1–19. doi: 10.1146/annurev-genet-110711-155459

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, stringtie and ballgown. *Nat. Protoc.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Plank, J. L., and Dean, A. (2014). Enhancer function: mechanistic and genome-wide insights come together. *Mol. Cell* 55, 5–14. doi: 10.1016/j.molcel.2014.06.015

Pradeepa, M. M., Grimes, G. R., Kumar, Y., Olley, G., Taylor, G. C., Schneider, R., et al. (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.* 48, 681–686. doi: 10.1038/ng.3550

Ramirez, F., Ryan, D. P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44, W160–W165. doi: 10.1093/nar/gkw257

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Rowan, S., Conley, K. W., Le, T. T., Donner, A. L., Maas, R. L., and Brown, N. L. (2008). Notch signaling regulates growth and differentiation in the mammalian lens. *Dev. Biol.* 321, 111–122. doi: 10.1016/j.ydbio.2008.06.002

Rutenberg-Schoenberg, M., Sexton, A. N., and Simon, M. D. (2016). The properties of long noncoding RNAs that regulate chromatin. *Annu. Rev. Genom. Hum. Genet.* 17, 69–94. doi: 10.1146/annurev-genom-090314-024939

Sakabe, N. J., Savic, D., and Nobrega, M. A. (2012). Transcriptional enhancers in development and disease. *Genome Biol.* 13:238. doi: 10.1186/gb-2012-13-1-238

Signal, B., Gloss, B. S., and Dinger, M. E. (2016). Computational approaches for functional prediction and characterisation of long noncoding RNAs. *Trends Genet.* 32, 620–637. doi: 10.1016/j.tig.2016.08.004

St Laurent, G., Wahlestedt, C., and Kapranov, P. (2015). The landscape of long noncoding RNA classification. *Trends Genet.* 31, 239–251. doi: 10.1016/j.tig.2015.03.007

Su, J., Huang, Y. H., Cui, X., Wang, X., Zhang, X., Lei, Y., et al. (2018). Homeobox oncogene activation by pan-cancer DNA hypermethylation. *Genome Biol.* 19:108. doi: 10.1186/s13059-018-1492-3

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, 2074–2081. doi: 10.1039/C3MB70608G

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505, 344–352. doi: 10.1038/nature12986

Vance, K. W., and Ponting, C. P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet.* 30, 348–355. doi: 10.1016/j.tig.2014.06.001

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.* 35(Database issue), D88–D92. doi: 10.1093/nar/gkl822

Ye, B., Liu, B., Hao, L., Zhu, X., Yang, L., Wang, S., et al. (2018). Klf4 glutamylation is required for cell reprogramming and early embryonic development in mice. *Nat. Commun.* 9:1261. doi: 10.1038/s41467-018-03008-2

Young, R. A. (2011). Control of the embryonic stem cell state. *Cell* 144, 940–954. doi: 10.1016/j.cell.2011.01.032

Yu, F., Li, K., Li, S., Liu, J., Zhang, Y., Zhou, M., et al. (2019). CFEA: a cell-free epigenome atlas in human diseases. *Nucleic Acids Res*. doi: 10.1093/nar/gkz715. [Epub ahead of print].

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515, 355–364. doi: 10.1038/nature13992.

Zhang, X., Zhang, J., Wang, T., Esteban, M. A., and Pei, D. (2008). Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J. Biol. Chem.* 283, 35825–35833. doi: 10.1074/jbc.M803481200

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2019). Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in alzheimer's disease. *Brief Bioinform.* 20, 598–608. doi: 10.1093/bib/bby021

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

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 © 2019 Liu, Jiang, Wang, Chen, Jin, Wang, Li, Yin, Shao, Li, Xu and Wu. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 09 January 2020
doi: 10.3389/fbioe.2019.00439

Integrative Analysis of Methylation and Transcriptome Identified Epigenetically Regulated lncRNAs With Prognostic Relevance for Thyroid Cancer

Qiuying Li[1], *Peng Wang*[1], *Chuanhui Sun*[2], *Chao Wang*[1]* *and Yanan Sun*[1]*

[1] Department of Otorhinolaryngology, Head and Neck Surgery, The Second Affiliated Hospital, Harbin Medical University, Harbin, China, [2] Department of Otorhinolaryngology, The First Affiliated Hospital, Guizhou University of Traditional Chinese Medicine, Guiyang, China

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Hengqiang Zhao,
Peking Union Medical College Hospital (CAMS), China
Yang Wang,
Hubei University, China

***Correspondence:**
Chao Wang
hljwangchao@163.com
Yanan Sun
syn2767@126.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 27 October 2019
Accepted: 10 December 2019
Published: 09 January 2020

Citation:
Li Q, Wang P, Sun C, Wang C and Sun Y (2020) Integrative Analysis of Methylation and Transcriptome Identified Epigenetically Regulated lncRNAs With Prognostic Relevance for Thyroid Cancer. Front. Bioeng. Biotechnol. 7:439. doi: 10.3389/fbioe.2019.00439

Emerging evidence has shown that epigenetic changes in DNA methylation, an important regulator of long non-coding RNA (lncRNA) expression, can disturb the expression patterns of lncRNAs and contribute to carcinogenesis. However, knowledge about crosstalk effects between DNA methylation and lncRNA regulation in thyroid cancer (THCA) remain largely unknown. In this study, we performed an integrated analysis of methylation and the transcriptome and identified 483 epigenetically regulated lncRNAs (EpilncRNAs) associated with the development and progression of THCA. These EpilncRNAs can be divided into two categories based on their methylation and expression patterns: 228 HyperLncRNAs and 255 HypoLncRNAs. Then, we identified a methylation-driven 5-lncRNA-based signature (EpiLncPM) to improve prognosis prediction using the random survival forest and multivariate Cox analysis, which were then validated using the training dataset [Hazard ratio (HR) = 50.097, 95% confidence interval (CI): 10.231-245.312, $p < 0.001$] and testing dataset (HR = 4.395, 95% CI: 0.981-19.686, $p = 0.053$). Multivariate analysis suggested that the EpiLncPM is an independent prognostic factor. By performing a functional enrichment analysis of GO and KEGG for mRNAs co-expressed with the EpiLncPM, we found that the EpiLncPM was involved in immune and inflammatory-related biological processes. Finally, *in situ* hybridization analysis in 119 papillary thyroid carcinoma (PTC) tissues and paired adjacent normal tissues revealed that selected candidate lncRNA *AC110011* has significantly higher expression of PTC compared to adjacent non-neoplastic tissues, and was closely related to the tumor size, lymph node metastasis, and extrathyroidal extension. In summary, our study characterized the crosstalk between DNA methylation and lncRNA, and provided novel biomarkers for the prognosis of THCA.

Keywords: thyroid cancer, long non-coding RNAs, DNA methylation, biomarkers, prognosis

INTRODUCTION

Thyroid cancer (THCA) constitutes one of the most frequently diagnosed types of head and neck tumors (Heroiu Cataloiu et al., 2013). Incidence rates of THCA had a significant upward trend in recent decades worldwide. It is estimated that the newly diagnosed cases of thyroid cancer are almost 50,000 in the United States in 2019 (Siegel et al., 2019). In China, the incidence rate of THCA in women has ranked fourth among the 10 most common cancers (Chen et al., 2016). Although patients diagnosed with THCA tended to have a favorable prognosis through surgery and radioiodine therapy, more than 10% of THCA patients will have a recurrence leading to considerable morbidity (Soares et al., 2014). Therefore, identifying reliable and accurate biomarkers for diagnosis and prognosis of THCA remains a challenge.

Long non-coding RNAs (lncRNAs) have been recognized as the major class of RNAs with more than 200 nucleotides that do not encode protein (Kopp and Mendell, 2018). A large body of literature has shown that lncRNAs play essential roles in a wide variety of biological processes, such as developmental and differentiation processes by lncRNA-mediated execution of gene expression programs (Fatica and Bozzoni, 2014; Marchese et al., 2017). High-throughput sequencing and profile analysis have identified a large number of differentially expressed lncRNAs in a multitude of cancers compared to normal tissues. In addition, aberrant lncRNA expression contributes to the development and maintenance of human cancers, thereby demonstrating the potential of lncRNAs as novel biomarkers in cancer diagnosis, prognosis, and as therapeutic targets (Spizzo et al., 2012; Sun et al., 2014; Huarte, 2015; Zhou et al., 2015a,b; Sanchez Calle et al., 2018; Zhou M. et al., 2018a,c; Bao et al., 2019). DNA methylation is one of the most common epigenetic mechanisms and is essential for the regulation of gene expression (Moore et al., 2013). It has been reported that epigenetic alterations in DNA methylation are also associated with various human cancers, including THCA (Stephen et al., 2011; Mancikova et al., 2014). Recent studies have reported aberrant DNA methylation of lncRNA promoters leading to perturbations of gene regulatory network, implying complex interplay between lncRNAs and DNA methylation (Zhao et al., 2016; Morlando and Fatica, 2018). However, crosstalk effects between DNA methylation and lncRNA regulation in THCA remain largely unknown.

In this study, we performed a genome-wide integrated analysis of methylation and the transcriptome to characterize the crosstalk between DNA methylation and lncRNA regulation, and identify epigenetically regulated lncRNAs. We further investigated the potential clinical value of these epigenetically regulated lncRNAs in a large number of THCA patients through bioinformatics analysis and experimental methods.

METHODS AND MATERIALS

Patients and Tissue Samples

The clinical information of 507 THCA patients was downloaded from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/) database.

TABLE 1 | The relation of *AC110011* expression and clinicopathological characteristics of PTC patients.

Characteristics	Number	AC110011 expression		*P*-value
		High	Low	
Gender				
Male	13	7	6	0.7935
Female	106	53	53	
Age (Years)				
<45	67	30	37	0.1622
≥45	52	30	22	
Multifocality				
Yes	22	11	11	0.9652
No	97	49	48	
Tumor size (cm)				
≤2	98	44	54	**0.0092***
>2	21	16	5	
Lymph node metastasis				
Yes	56	36	20	**0.0043***
No	63	24	39	
Extral thyroidal extension				
Yes	23	19	4	**0.0006***
No	96	41	55	
TNM stage				
I+II	106	51	55	0.1506
III+IV	13	9	4	

*$P \leq 0.05$ was considered significant.
The bold values indicates that they have statistical significance.

Another in-house dataset including 119 PTC tissues and paired adjacent normal tissues were collected from patients who underwent surgery at the Second Affiliated Hospital of Harbin Medical University (HMU). No patients received any local or systemic treatments before the operation. The adjacent non-cancerous tissues were collected >2 cm from the tumor margins on the same or another lobe. All tissue samples were confirmed independently by two pathologists, blocked of formalin-fixed paraffin-embedded material and stored at 2–8°C with desiccation until use for later experiments. The clinicopathological characteristics of these patients are listed in **Table 1**. Informed consent was obtained for each patient, and the experiments were allowed by the Research Ethics Committee of HMU.

Processing and Analysis of lncRNA Expression Profiles

RNA-seq data of 510 THCA tumor tissues and 58 non-cancer tissues based on the IlluminaHiSeq_RNASeq platform were retrieved from the UCSC Xena Browser (https://xena.ucsc.edu/). We annotate 15,873 lncRNAs from the RNA-seq data based on GENCODE (v23) annotations. Differential expression analysis was conducted using the R package "DESeq2." Those lncRNAs with |log2(fold change)| >1 and False Discovery Rate (FDR)-corrected *p*-values < 0.05 were identified as differentially

expressed lncRNAs. Hierarchical clustering analysis of samples based on lncRNA expression was performed using the R package "pheatmap" with the "ward.D2" method.

Processing and Analysis of DNA Methylation Profiling

DNA methylation data of 515 THCA tumor tissues and 56 non-cancer tissues based on the Illumina Human Methylation 450 platform were retrieved from the UCSC Xena Browser (https://xena.ucsc.edu/). After removing probes with the missing value in more than 10% of samples, a total of 372,978 probes were stored for further analysis. Differentially methylated CpG loci between paired tumor tissues and non-cancer tissues were identified using a paired *t*-test with FDR-corrected *p*-values < 0.05 and absolute differences between group methylation mean (DGMB) > 0.04. Differentially methylated CpG loci between tumors with and without recurrence were determined using the R package "minfi" with FDR-corrected *p*-values < 0.05 and DGMB > 0.04.

Survival Analysis

The random survival forest was carried out to identify the optimal combination of epigenetically regulated lncRNAs as novel signature for survival prediction through the R package "randomforestSRC" (Taylor, 2011). Kaplan-Meier survival curves and log-rank tests were used to compare survival differences between high-risk group and low-risk group with the R package "survival." Univariate and multivariate Cox proportional hazards regression analyses were conducted through the R package "survival." The time-dependent Receiver Operating Characteristic (ROC) curve was used to evaluate the performance of the signature for survival prediction using the R package "survivalROC."

Function Enrichment Analysis

Function enrichment analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) was carried out using the R package "clusterProfiler" (Yu et al., 2012). GO terms or KEGG pathways with adjusted *p*-value <0.05 were taken to be significantly enriched.

RNAscope *in situ* Hybridization

Paraffin sections were disposed according to the agreement developed by the Advanced Cell Diagnostics. ISH was disposed using the RNAscope® 2.5 Assay (ACD, Inc. Catalog No. 322335) and RNAscope® 2.5HD Detection Kit -BROWN (ACD, Inc.Cat. No. 322310). RNAscope Probe-Hs-AC110011 and positive and negative control probes were ordered from ACD. These probes included positive control probes PPIB (positive control, Cat. No. 313901), and negative control probe DapB (negative control, Cat. No. 310043). The percentages of positive cancer cells were scored as follows: 0: none; 1: <10%; 2: 10–50%; and 3: >50%. A score of 2 was used to distinguish between low (<2) and high (≥2) levels of AC110011 gene expression. Images of the slides were analyzed using an OLYMPUS Dual-CCD microscope digital camera, and relevant semi-quantitative scores were acquired by estimating the punctate staining. In addition, half of the quantitative fraction was acquired by estimating the dot dyeing.

RESULTS

Identification of Altered lncRNAs in the Development and Progression of THCA

We first performed differential expression analysis for lncRNAs in 58 pairs of tumors and non-cancer tissues of THCA. A total of 1,969 lncRNAs were found to be differentially expressed, including 868 up-regulated and 1,101 down-regulated lncRNAs in tumors compared to non-cancer tissues (**Figure 1A**). Hierarchical clustering analysis showed that the expression patterns of 1,969 lncRNAs were capable of distinguishing tumor samples from non-cancer tissues (**Figure 1B**). To identify lncRNAs associated with tumor recurrence, differential expression analysis for lncRNAs was undertaken between 32 recurrent tumors and 401 recurrence-free tumors. Finally, we identified 72 lncRNAs that were significantly altered in recurrent tumors vs. recurrence-free tumors. Of them, 21 lncRNAs were up-regulated and 51 lncRNAs were downregulated in recurrent tumors compared to recurrence-free tumors (**Figure 1C**). Hierarchical clustering analysis showed that all samples fell into two clusters that are significantly associated with recurrence status ($p = 0.0008$, Chi-square test; **Figure 1D**).

Differential DNA Methylation Profiling During THCA Development and Progression

Further comparison of genome-wide DNA methylation on 56 pairs of tumor and adjacent normal tissues of THCA identified a total of 41,157 (11.03%) differentially methylated CpG sites, including 18,848 (45.8%) hypermethylated CpG sites (FDR < 0.05 and DGMB > 0.04) and 22,309 (54.2%) hypomethylated CpG sites (FDR < 0.05 and DGMB <-0.04) in tumors compared to adjacent normal tissue. To identify novel CpG islands aberrantly methylated in THCA recurrence, we also performed differential DNA methylation analysis between tumors with and without recurrence. We identified 20,283 (5.44%) differentially methylated CpG sites including 14,654 (72.25%) hypermethylated CpG sites (FDR < 0.05 and DGMB > 0.04) and 5,629 (27.75%) hypomethylated CpG sites (FDR < 0.05 and DGMB <-0.04) in tumors with recurrence compared to recurrence-free tumors. A Manhattan plot showed that these differentially methylated CpG sites are significantly associated with THCA development and progression is distributed across the methylome (**Figures 2A,B**).

The Integrated Analysis Identifies Epigenetically Regulated lncRNAs in THCA

To identify epigenetically regulated lncRNAs (EpilncRNAs) in THCA, these differentially methylated CpG sites were mapped to 3 kb upstream and 3 kb downstream of differentially expressed lncRNAs. A total of 7,765 differentially methylated CpG sites were located in 689 up-regulated lncRNAs and 10,349 differentially methylated CpG sites were located in 873 down-regulated lncRNAs in tumors compared to adjacent

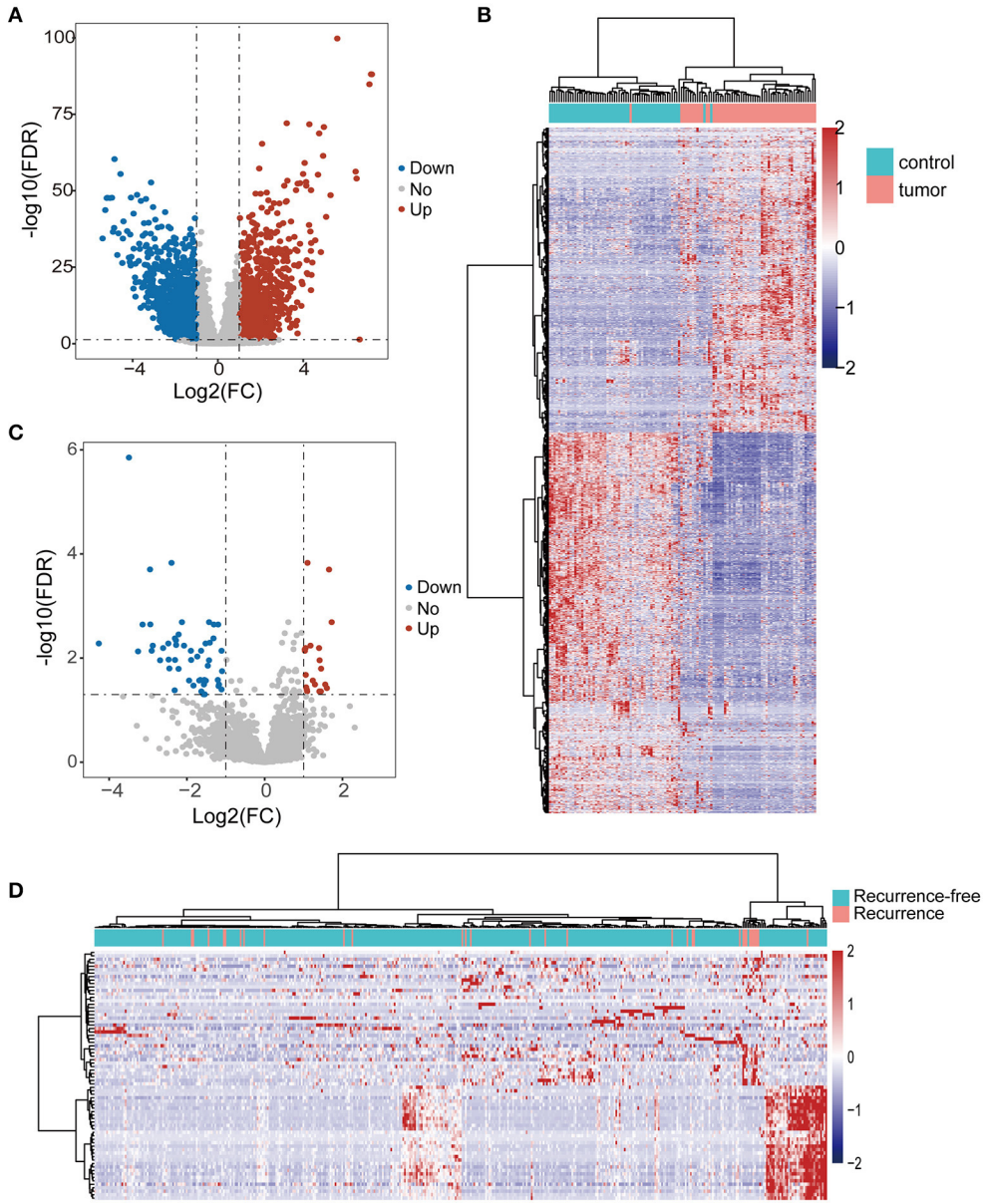


FIGURE 1 | Identification of differentially expressed lncRNAs. **(A)** Volcano plot of differentially expressed lncRNAs between tumors and normal tissues. Blue dots indicate down-regulated lncRNAs and red dots denote up-regulated lncRNAs. **(B)** The hierarchical clustering heat map of tumor samples and non-cancer tissues based on differentially expressed lncRNAs. **(C)** Volcano plot of differentially expressed lncRNAs between tumors with and without recurrence. Blue dots indicate down-regulated lncRNAs and red dots denote up-regulated lncRNAs. **(D)** The hierarchical clustering heat map of tumor with and without recurrence based on differentially expressed lncRNAs.

normal tissues. For tumors with recurrence, there were 427 differentially methylated CpG sites located in 20 up-regulated lncRNAs and 334 differentially methylated CpG sites located in 40 down-regulated lncRNAs. Then we grouped these lncRNAs into categories based on DNA methylation in lncRNA: HyperLncRNAs and HypoLncRNAs. HyperLncRNAs are down-regulated lncRNAs with high DNA methylation levels, whereas HypoLncRNAs are up-regulated lncRNAs with low DNA methylation levels. Finally, we identified 475 methylation–driven lncRNAs (224 HyperLncRNAs and 251 HypoLncRNAs) associated with tumor development and 8 methylation-driven lncRNAs (4 HyperLncRNAs and 4 HypoLncRNAs) associated with tumor recurrence. These results demonstrated that methylation-mediated dysregulated lncRNA expression patterns involved in THCA development and progression. Furthermore, we found that 228 HyperLncRNAs displayed dissimilar DNA methylation patterns with 255 HypoLncRNAs (**Figures 2C,D**).

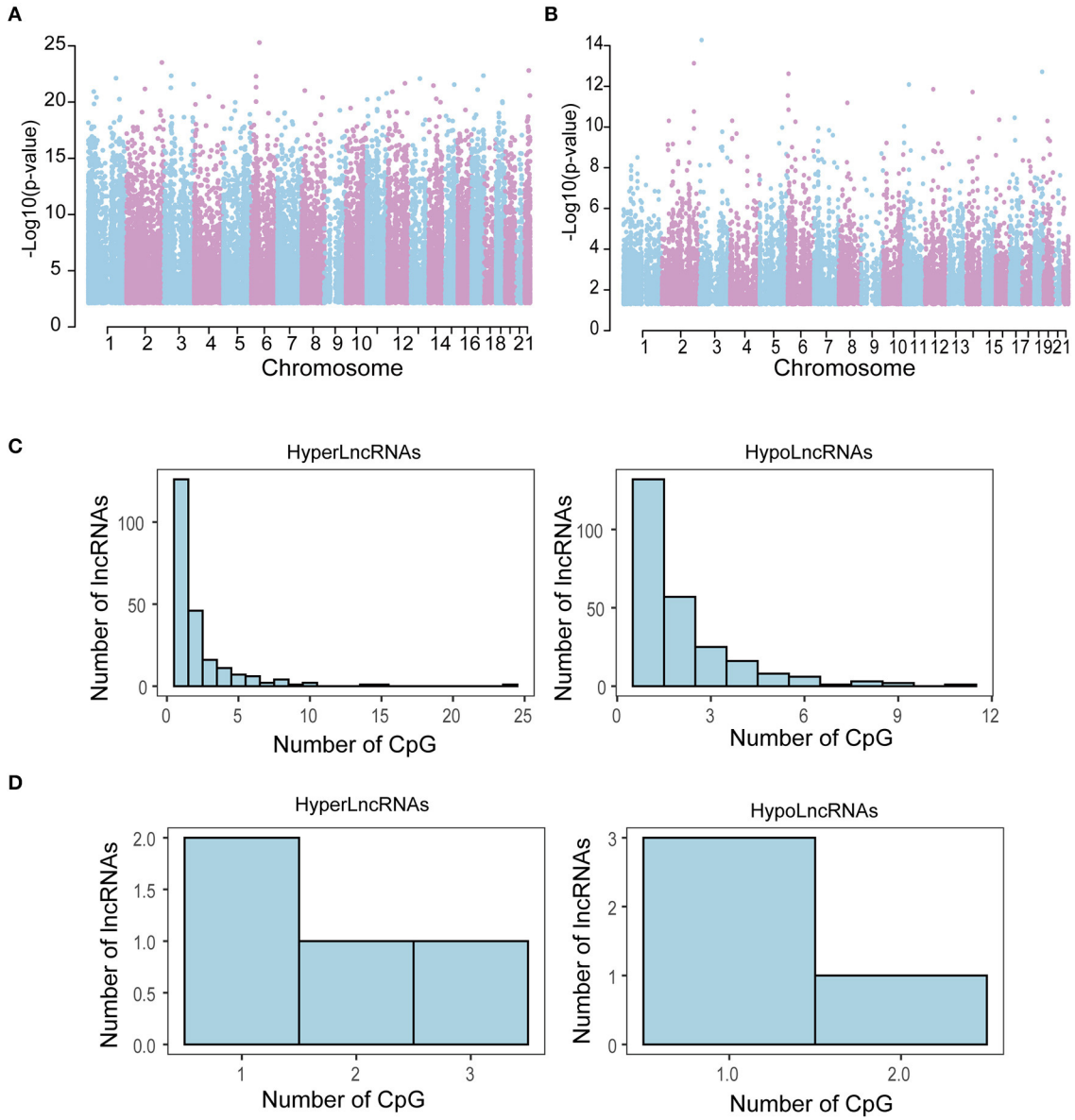


FIGURE 2 | Chromosome distribution of differentially methylated CpG sites. A Manhattan plot was drawn for differentially methylated CpG sites between tumors and adjacent normal tissue **(A)** and between tumors with and without recurrence **(B)**. Distribution of HyperLncRNAs and HypoLncRNAs possessing different differentially methylated CpG sites associated with tumor development **(C)** and tumor recurrence **(D)**.

Construction and Validation of an Epigenetically Regulated lncRNA-Based Prognostic Model (EpiLncPM) in THCA

To examine whether epigenetically regulated lncRNAs have prognostic value for THCA patients, we first performed a hierarchical clustering analysis of all THCA patients based on the 483 methylation-driven lncRNAs. As shown in **Figure 3A**, all THCA were grouped into two patient clusters with significantly different survival ($p = 0.034$, log-rank test; **Figure 3B**), suggesting that epigenetically regulated lncRNAs may be used as biomarkers in survival prediction of THCA. Therefore, all THCA patients were randomly split into the training dataset ($n = 250$) and testing dataset ($n = 249$) and performed univariate Cox regression analysis for 475 methylation-driven lncRNAs in the training dataset. A total of 42 methylation-driven lncRNAs were significantly associated with survival. Thereafter, we performed the RSF on 42 candidate methylation-driven lncRNA biomarkers and identified an optimal combination of 5 methylation-driven lncRNAs as a prognostic signature. Finally an epigenetically regulated lncRNA-based prognostic scoring model (EpiLncPM) was developed in the training dataset using the expression levels of five optimal methylation-driven

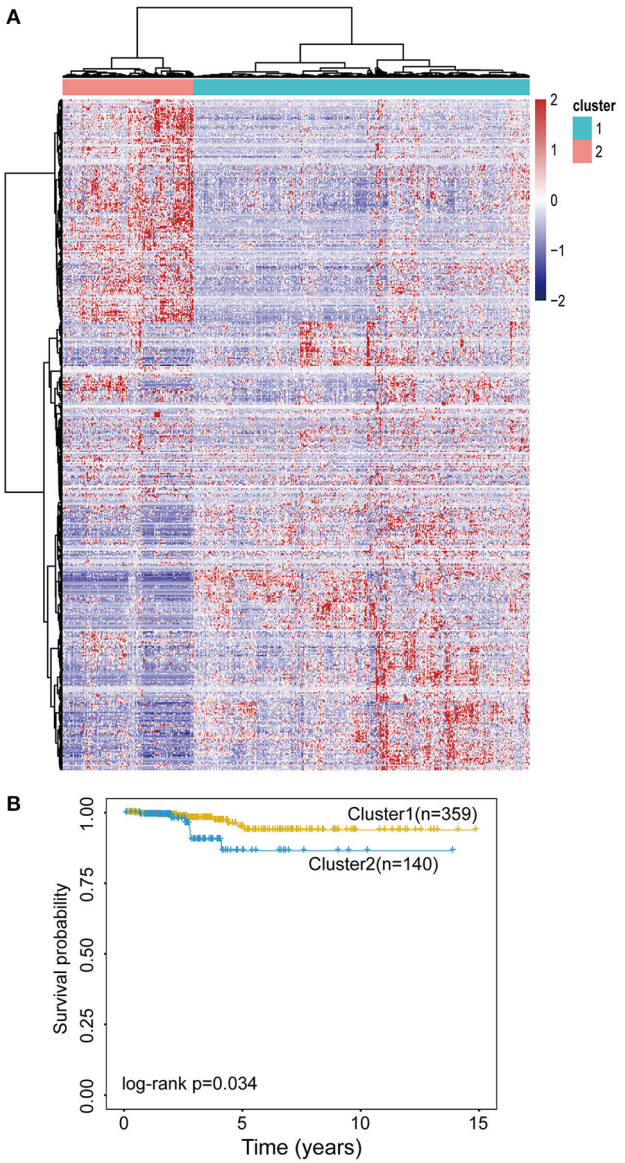


FIGURE 3 | Evaluation of the prognostic value of the EpilncRNAs. **(A)** The hierarchical clustering heat map of THCA patients based on 483 EpilncRNAs. **(B)** Kaplan-Meier survival curves for two patient clusters.

lncRNAs weighted by their coefficients from multivariate Cox regression analysis as follows: EpiLncPM = 13.1* *AP006248.2*+2.53**AC068580.3*+33.2**AC016396.2*+3.12**LINC 01140*+1.19**LINC01135* (**Table 2**). Patients in the training dataset were divided into the high-risk group (n = 26) and low-risk group (n = 224) according to the optimal risk cutoff point from the R package "maxstat." As expected, the survival analysis showed that patients with a low-risk score have a better prognosis than those with a high-risk score ($p < 0.001$, log-rank test; **Figure 4A**). The time-dependent ROC analysis showed that the AUC of the performance of the EpiLncPM for survival prediction at three, 5 and 10-year survival rates in the training dataset reached 0.948, 0.965, and 0.949, respectively (**Figure 4B**).

TABLE 2 | Detailed information of five optimal methylation–driven prognostic lncRNAs.

Ensembl version	Ensembl name	Genomic location	*P*-value
ENSG00000279271.1	AP006248.2	Chromosome 8: 17,498,647-17,499,238(–)	0.008
ENSG00000235027.1	AC068580.3	Chromosome 11: 1,760,348-1,762,486 (+)	0.009
ENSG00000261076.1	AC016396.2	Chromosome 10: 58,325,614-58,327,030 (–)	0.002
ENSG00000267272.5	LINC01140	Chromosome 1: 87,129,765-87,169,198 (+)	0.023
ENSG00000234807.7	LINC01135	Chromosome 1: 58,785,128-58,901,109 (+)	0.030

When the EpiLncPM was applied to the testing dataset, patients in the high-risk group had significantly shorter survival than those in the low-risk group ($p = 0.034$, log-rank test; **Figure 4C**). The time-dependent ROC analysis showed that the AUC of the performance of the EpiLncPM for survival prediction at three, 5 and 10-year survival rates in the testing dataset reached 0.542, 0.625 and 0.688, respectively (**Figure 4D**).

Independence of the EpiLncPM From Other Clinical Factors

Univariate Cox regression analysis showed that the EpiLncPM and stage are both significantly correlated with survival in the training and testing datasets (HR = 50.097, 95% CI 10.231-245.312, $p < 0.001$ for training dataset; HR = 4.395, 95% CI 0.981-19.686, $p = 0.053$; **Table 3**). Therefore, to investigate whether the EpiLncPM is an independent factor in predicting survival, we performed multivariate analysis including age, gender, stage, and the EpiLncPM. Results of multivariate analyses showed that the EpiLncPM still maintained a significant association with survival adjusting by other clinical factors (**Table 3**). These observations showed that EpiLncPM is an independent factor in predicting the survival of THCA patients.

Functional Implication of the EpiLncPM

To gain insights into the functional role of the EpiLncPM, we calculated the Pearson correlation coefficient between expression levels of lncRNAs and mRNAs using the EpiLncPM and identified 537 mRNAs that correlated with the EpiLncPM (Pearson correlation coefficient > 0.5). GO functional enrichment analysis revealed that the correlated mRNAs were significantly clustered in immune and inflammatory-related biological processes (**Figure 5A**). In addition, the KEGG enrichment analysis showed that the correlated mRNAs were enriched in known cancer-related pathways (**Figure 5B**).

In situ Hybridization Analysis

The expression of AC110011 in 119 paraffin-embedded tissue samples of PTC and adjacent non-neoplastic tissues was semi-quantitatively examined by ISH. ISH showed that AC110011 was expressed in the nucleus (**Figure 6**). PTC tissues showed a significant increase in AC110011 expression as compared to that observed in adjacent non-neoplastic tissues. The AC110011

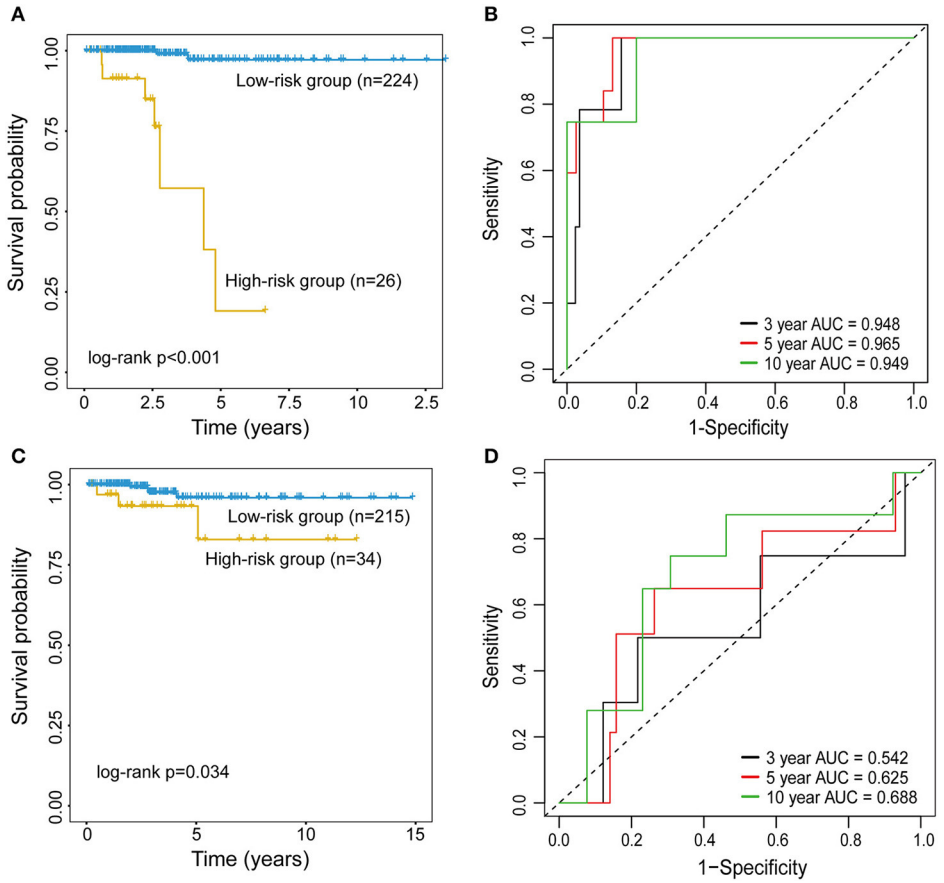


FIGURE 4 | Performance evaluation of the EpiLncPM for survival prediction. Kaplan-Meier survival curves for the high- and low-risk groups in the training **(A)** and testing **(C)** datasets. The time-dependent receiver operating characteristic (ROC) curve analysis for survival prediction at three, 5 and 10-year survival rates in the training **(B)** and testing **(D)** datasets.

TABLE 3 | Univariate and multivariate analysis for survival in the training and testing datasets.

Variables		Univariate analysis			Multivariate analysis		
		HR	95% CI	*p*-value	HR	95% CI	*p*-value
Training dataset (*n* = 250)							
EpiLncPM	High vs. low	50.097	10.231–245.312	<0.001	29.050	5.94–142.074	<0.001
Age	Old vs. young	Inf	0-Inf	0.998	Inf	0-Inf	0.990
Gender	Female vs. male	1.307	0.271–6.305	0.739	0.237	0.247–6.509	0.776
Stage	III, IV vs. I, II	9.377	1.933–45.492	0.005	0.302	0.26–7.051	0.720
Testing dataset (*n* = 249)							
EpiLncPM	High vs. low	4.395	0.981–19.686	0.053	3.699	0.813–16.833	0.091
Age	Old vs. young	Inf	0-Inf	0.998	Inf	0-Inf	0.999
Gender	Female vs. male	0.202	0.045–0.915	0.038	0.268	0.058–1.233	0.091
Stage	III, IV vs. I, II	5.416	1.048–27.997	0.044	1.083	0.207–5.662	0.924

expression was statistically associated with tumor size (cm), lymph node metastasis, and extrathyroidal extension (**Table 1**).

DISCUSSION

Recent multi-omics analysis has demonstrated that THCA is a heterogeneous disease characterized by a high degree of molecular heterogeneity (Killock, 2014; Yoo et al., 2019), implying the potential of molecular profiling as a predictive factor in the diagnosis, prognosis and treatment response for THCA which will overcome the limitations of conventional clinical and histopathological features. mRNA and miRNA expression patterns have been widely investigated for a long time and many mRNA- or miRNA-based signatures have been identified

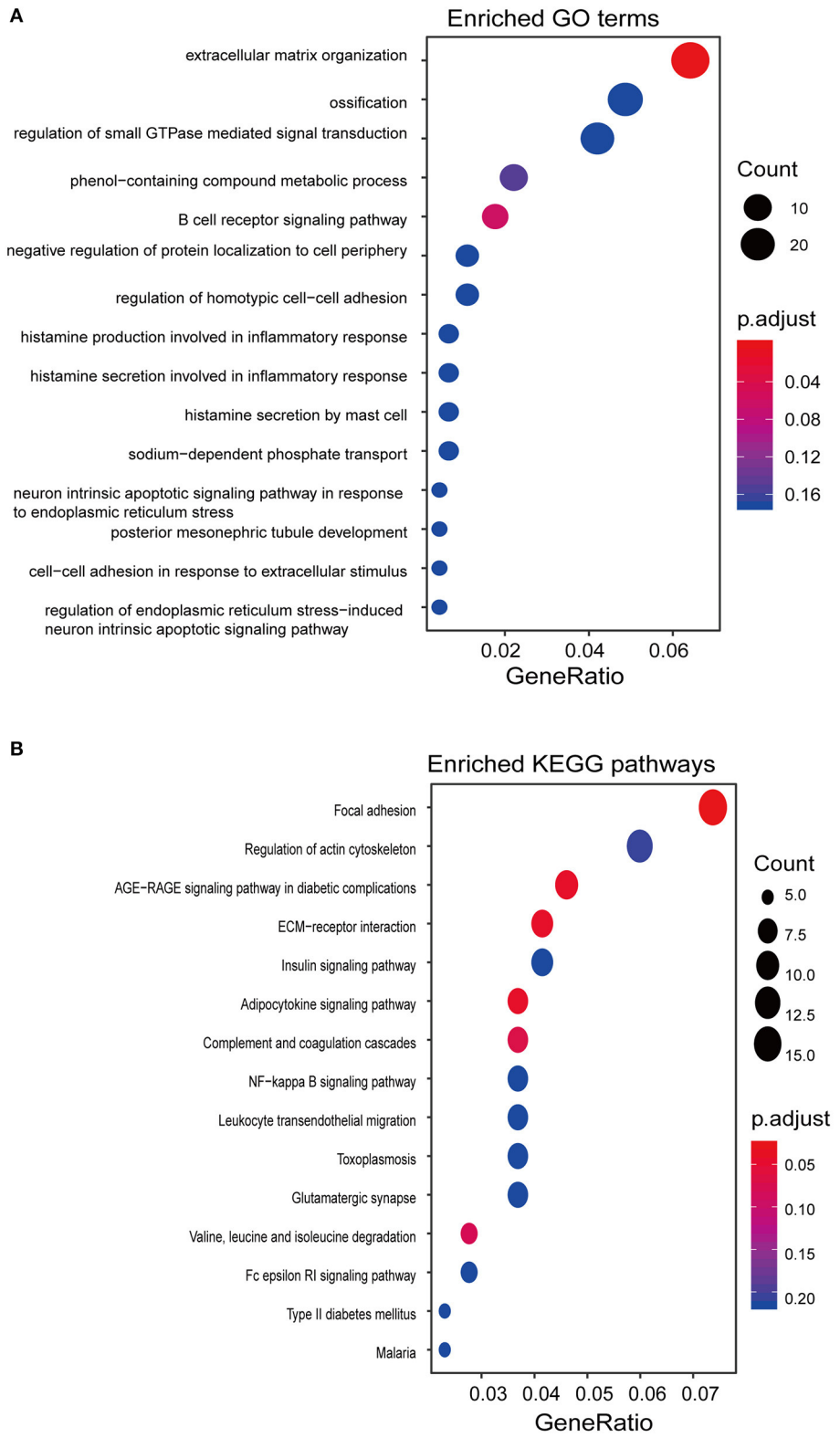


FIGURE 5 | Functional prediction of the EpiLncPM. **(A)** Functional enrichment analysis of GO. **(B)** Functional enrichment analysis of KEGG.

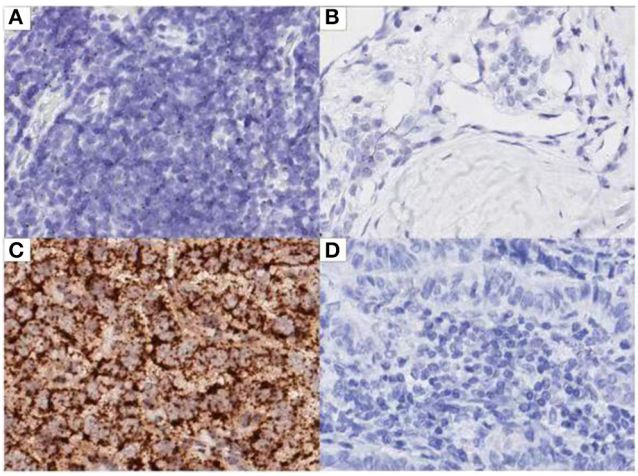


FIGURE 6 | ISH analysis for lncRNA AC110011 in patient samples. **(A)** PTC tissue. **(B)** Adjacent non-neoplastic tissues. **(C)** Positive control. **(D)** Negative control.

to improve the diagnosis and prognosis of THCA (Santarpia et al., 2013; Han et al., 2018; Kim et al., 2018; Teng et al., 2018; Wang et al., 2019). In recent years, a new ncRNA class, lncRNA, has gradually become a research hotspot in diverse cancer fields. Numerous studies have demonstrated the superiority of lncRNA as diagnostic and predictive biomarkers in a diverse range of cancers compared to mRNAs and miRNAs (Zhou et al., 2015c, 2017, 2019; Zhou M. et al., 2018b). Aberrant lncRNA expression has also been observed in the development and progression of THCA (Yang et al., 2016; Lu et al., 2018), and several lncRNA biomarkers have been identified for THCA diagnosis and prognosis (Li et al., 2017; Liu et al., 2018; Liang and Sun, 2019).

Emerging evidence has shown that DNA methylation is an important epigenetic regulator of lncRNA expression and epigenetic changes in DNA methylation can disturb the expression pattern of lncRNAs and contributes to carcinogenesis (Heilmann et al., 2017; Tang, 2018; Zhou Z. et al., 2018; Bao et al., 2019). However, the interplay between lncRNA regulation and DNA methylation in the development and progression of THCA is still largely unknown. In this study, we performed an integrated analysis of methylation and the transcriptome to characterize the dysregulated DNA methylation pattern of lncRNAs. By performing a differential expression analysis, we observed genome-wide changes in lncRNA expression and DNA methylation during THCA development and progression. By directly mapping altered DNA methylation to the flaking regions of lncRNAs, we found that a total of 483 differentially expressed lncRNAs were epigenetically deregulated, which were defined as methylation-driven lncRNAs. These lncRNAs can be separated into two categories based on their methylation patterns and expression levels, between tumor and normal samples or between tumors with and without recurrence.

Despite the fact that several previous studies have identified a host of lncRNAs with prognostic roles in THCA, knowledge about the clinical value of methylation-driven lncRNAs remains limited. We performed clustering analysis for THCA patients based on the expression patterns of methylation-driven lncRNAs, which were found to be able to distinguish patients with different prognoses, thereby demonstrating a potential function in the prognosis of THCA. Therefore, we identified a methylation-driven 5-lncRNA-based signature (EpiLncPM) to improve the prognosis prediction utilizing the RSF and multivariate Cox analysis. This EpiLncPM was validated by the training and testing datasets. Moreover, the prognostic capacity of the EpiLncPM is independent of other clinical and pathological factors for the survival of patients with THCA. By performing function enrichment analysis for mRNAs co-expressed with the EpiLncPM, we gained insight into the potential functional relevance of methylation-driven lncRNAs in THCA. We found that the EpiLncPM is involved in immune and inflammatory-related biological processes, such as B cell receptor signaling pathways, histamine production involved in inflammatory responses, histamine secretion by mast cells, and known cancer-related pathways. It has been reported that chronic inflammation and the tumor microenvironment play critical roles in cancer development and progression, including THCA (Cunha et al., 2014; Rotondi et al., 2018; Ferrari et al., 2019). For example, mast cell infiltrates are linked to extrathyroidal extension and invasiveness (Visciano et al., 2015). NF-κB is associated with inflammatory and immune responses and its increased activity correlate with a more aggressive phenotype of THCA (Giuliani et al., 2018). Finally, we validated the functional roles of one selected methylation-driven lncRNA signature by *in situ* hybridization analysis in 119 PTC tissues and paired with adjacent normal tissues.

In summary, we performed a genome-wide integrated analysis of methylation and the transcriptome to characterize the crosstalk between DNA methylation and lncRNA, and identify epigenetically regulated lncRNAs. Additionally, we identified a methylation-driven 5-lncRNA-based signature (EpiLncPM) with potential clinical application in predicting the prognosis of THCA.

DATA AVAILABILITY STATEMENT

Clinical information of THCA patients was downloaded from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/) database. RNA-seq data and DNA methylation data of tumor tissues and non-cancer tissues were retrieved from the UCSC Xena Browser (https://xena.ucsc.edu/).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee of Harbin medical university. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CW and YS conceived and designed the experiments. QL, PW, and CS performed the experiments and analyzed the data. QL, CW, and YS wrote the paper. All authors read and approved the final manuscript.

FUNDING

This study was supported by the scientific research project of Heilongjiang health and Family Planning Commission (Grant no. 2017-081); the Innovation Research Fund for young and middle-aged in the Second Affiliated Hospital of Harbin Medical University: (Grant no. kycx2018-04); the scientific funds of National Natural Science Foundation of China (81772874 and 81272965).

REFERENCES

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief. Bioinform.* doi: 10.1093/bib/bbz118. [Epub ahead of print].

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

Cunha, L. L., Marcello, M. A., and Ward, L. S. (2014). The role of the inflammatory microenvironment in thyroid carcinogenesis. *Endocr. Relat. Cancer* 21, R85–R103. doi: 10.1530/ERC-13-0431

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606

Ferrari, S. M., Fallahi, P., Galdiero, M. R., Ruffilli, I., Elia, G., Ragusa, F., et al. (2019). Immune and inflammatory cells in thyroid cancer microenvironment. *Int. J. Mol. Sci.* 20:4413. doi: 10.3390/ijms20184413

Giuliani, C., Bucci, I., and Napolitano, G. (2018). The role of the transcription factor nuclear factor-kappa b in thyroid autoimmunity and cancer. *Front. Endocrinol.* 9:471. doi: 10.3389/fendo.2018.00471

Han, L. O., Li, X. Y., Cao, M. M., Cao, Y., and Zhou, L. H. (2018). Development and validation of an individualized diagnostic signature in thyroid cancer. *Cancer Med.* 7, 1135–1140. doi: 10.1002/cam4.1397

Heilmann, K., Toth, R., Bossmann, C., Klimo, K., Plass, C., and Gerhauser, C. (2017). Genome-wide screen for differentially methylated long noncoding RNAs identifies Esrp2 and lncRNA Esrp2-as regulated by enhancer DNA methylation with prognostic relevance for human breast cancer. *Oncogene* 36, 6446–6461. doi: 10.1038/onc.2017.246

Heroiu Cataloiu, A. D., Danciu, C. E., and Popescu, C. R. (2013). Multiple cancers of the head and neck. *Maedica* 8, 80–85.

Huarte, M. (2015). The emerging role of lncRNAs in cancer. *Nat. Med.* 21, 1253–1261. doi: 10.1038/nm.3981

Killock, D. (2014). Genetics: the cancer genome atlas maps papillary thyroid cancer. *Nat. Rev. Clin. Oncol.* 11:681. doi: 10.1038/nrclinonc.2014.193

Kim, K., Jeon, S., Kim, T. M., and Jung, C. K. (2018). Immune gene signature delineates a subclass of papillary thyroid cancer with unfavorable clinical outcomes. *Cancers* 10:E494. doi: 10.3390/cancers10120494

Kopp, F., and Mendell, J. T. (2018). Functional classification and experimental dissection of long noncoding RNAs. *Cell* 172, 393–407. doi: 10.1016/j.cell.2018.01.011

Li, Q., Li, H., Zhang, L., Zhang, C., Yan, W., and Wang, C. (2017). Identification of novel long non-coding RNA biomarkers for prognosis prediction of papillary thyroid cancer. *Oncotarget* 8, 46136–46144. doi: 10.18632/oncotarget.17556

Liang, W., and Sun, F. (2019). Identification of pivotal lncRNAs in papillary thyroid cancer using lncRNA-mRNA-miRNA ceRNA network analysis. *PeerJ* 7:e7441. doi: 10.7717/peerj.7441

Liu, G., Zheng, J., Zhuang, L., Lv, Y., Zhu, G., Pi, L., et al. (2018). A prognostic 5-lncRNA expression signature for head and neck squamous cell carcinoma. *Sci. Rep.* 8:15250. doi: 10.1038/s41598-018-33642-1

Lu, W., Xu, Y., Xu, J., Wang, Z., and Ye, G. (2018). Identification of differential expressed lncRNAs in human thyroid cancer by a genome-wide analyses. *Cancer Med.* 7, 3935–3944. doi: 10.1002/cam4.1627

Mancikova, V., Buj, R., Castelblanco, E., Inglada-Pérez, L., Diez, A., de Cubas, A. A., et al. (2014). DNA methylation profiling of well-differentiated thyroid cancer uncovers markers of recurrence free survival. *Int. J. Cancer* 135, 598–610. doi: 10.1002/ijc.28703

Marchese, F. P., Raimondi, I., and Huarte, M. (2017). The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* 18:206. doi: 10.1186/s13059-017-1348-2

Moore, L. D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology* 38, 23–38. doi: 10.1038/npp.2012.112

Morlando, M., and Fatica, A. (2018). Alteration of epigenetic regulation by long noncoding RNAs in cancer. *Int. J. Mol. Sci.* 19:E570. doi: 10.3390/ijms19020570

Rotondi, M., Coperchini, F., Latrofa, F., and Chiovato, L. (2018). Role of chemokines in thyroid cancer microenvironment: is CXCL8 the main player? *Front. Endocrinol.* 9:314. doi: 10.3389/fendo.2018.00314

Sanchez Calle, A., Kawamura, Y., Yamamoto, Y., Takeshita, F., and Ochiya, T. (2018). Emerging roles of long non-coding RNA in cancer. *Cancer Sci.* 109, 2093–2100. doi: 10.1111/cas.13642

Santarpia, L., Calin, G. A., Adam, L., Ye, L., Fusco, A., Giunti, S., et al. (2013). A miRNA signature associated with human metastatic medullary thyroid carcinoma. *Endocr. Relat. Cancer* 20, 809–823. doi: 10.1530/ERC-13-0357

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Soares, P., Celestino, R., Melo, M., Fonseca, E., and Sobrinho-Simões, M. (2014). Prognostic biomarkers in thyroid cancer. *Virchows Arch.* 464, 333–346. doi: 10.1007/s00428-013-1521-2

Spizzo, R., Almeida, M. I., Colombatti, A., and Calin, G. A. (2012). Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene* 31, 4577–4587. doi: 10.1038/onc.2011.621

Stephen, J. K., Chitale, D., Narra, V., Chen, K. M., Sawhney, R., and Worsham, M. J. (2011). DNA methylation in thyroid tumorigenesis. *Cancers* 3, 1732–1743. doi: 10.3390/cancers3021732

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, 2074–2081. doi: 10.1039/C3MB70608G

Tang, B. (2018). Inference of crosstalk effects between DNA methylation and lncRNA regulation in NSCLC. *Biomed Res. Int.* 2018:7602794. doi: 10.1155/2018/7602794

Taylor, J. M. (2011). Random survival forests. *J. Thorac. Oncol.* 6, 1974–1975. doi: 10.1097/JTO.0b013e318233d835

Teng, H., Mao, F., Liang, J., Xue, M., Wei, W., Li, X., et al. (2018). Transcriptomic signature associated with carcinogenesis and aggressiveness of papillary thyroid carcinoma. *Theranostics* 8, 4345–4358. doi: 10.7150/thno.26862

Visciano, C., Prevete, N., Liotti, F., and Marone, G. (2015). Tumor-associated mast cells in thyroid cancer. *Int. J. Endocrinol.* 2015:705169. doi: 10.1155/2015/705169

Wang, Z., Lv, J., Zou, X., Huang, Z., Zhang, H., Liu, Q., et al. (2019). A three plasma microRNA signature for papillary thyroid carcinoma diagnosis in Chinese patients. *Gene* 693, 37–45. doi: 10.1016/j.gene.2019.01.016

Yang, M., Tian, J., Guo, X., Yang, Y., Guan, R., Qiu, M., et al. (2016). Long noncoding RNA are aberrantly expressed in human papillary thyroid carcinoma. *Oncol. Lett.* 12, 544–552. doi: 10.3892/ol.2016.4653

Yoo, S. K., Song, Y. S., Lee, E. K., Hwang, J., Kim, H. H., Jung, G., et al. (2019). Integrative analysis of genomic and transcriptomic characteristics associated with progression of aggressive thyroid cancer. *Nat. Commun.* 10:2764. doi: 10.1038/s41467-019-10680-5

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

Zhao, Y., Sun, H., and Wang, H. (2016). Long noncoding RNAs in DNA methylation: new players stepping into the old game. *Cell Biosci.* 6:45. doi: 10.1186/s13578-016-0109-3

Zhou, M., Guo, M., He, D., Wang, X., Cui, Y., Yang, H., et al. (2015a). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J. Transl. Med.* 13:231. doi: 10.1186/s12967-015-0556-3

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018a). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., et al. (2015b). Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. *Mol. Biosyst.* 11, 760–769. doi: 10.1039/C4MB00511B

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018c). An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. *Mol. Neurobiol.* 55, 3684–3697. doi: 10.1007/s12035-017-0572-9

Zhou, M., Zhang, Z., Zhao, H., Bao, S., and Sun, J. (2018b). A novel lncRNA-focus expression signature for survival prediction in endometrial carcinoma. *BMC Cancer* 18:39. doi: 10.1186/s12885-017-3983-0

Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2019). Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Brief. Bioinform.* 20, 598–608. doi: 10.1093/bib/bby021

Zhou, M., Zhao, H., Wang, Z., Cheng, L., Yang, L., Shi, H., et al. (2015c). Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. *J. Exp. Clin. Cancer Res.* 34:102. doi: 10.1186/s13046-015-0219-5

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

Zhou, Z., Lin, Z., Pang, X., Tariq, M. A., Ao, X., Li, P., et al. (2018). Epigenetic regulation of long non-coding RNAs in gastric cancer. *Oncotarget* 9, 19443–19458. doi: 10.18632/oncotarget.23821

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 © 2020 Li, Wang, Sun, Wang and Sun. 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 January 2020
doi: 10.3389/fcell.2020.00020

Co-occurrence and Mutual Exclusivity Analysis of DNA Methylation Reveals Distinct Subtypes in Multiple Cancers

Wubin Ding[1†], Guoshuang Feng[2†], Yige Hu[1], Geng Chen[1] and Tieliu Shi[1,2,3]*

[1] Center for Bioinformatics and Computational Biology, Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai, China, [2] Big Data and Engineering Research Center, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China, [3] Biological Targeting Diagnosis and Therapy Research Center, Guangxi Medical University, Nanning, China

Co-occurrence and mutual exclusivity (COME) of DNA methylation refer to two or more genes that tend to be positively or negatively correlated in DNA methylation among different samples. Although COME of gene mutations in pan-cancer have been well explored, little is known about the COME of DNA methylation in pan-cancer. Here, we systematically explored the COME of DNA methylation profile in diverse human cancer. A total of 5,128,332 COME events were identified in 14 main cancers types in The Cancer Genome Atlas (TCGA). We also identified functional epigenetic modules of the zinc finger gene family in six cancer types by integrating the gene expression and DNA methylation data and the frequently occurred COME network. Interestingly, most of the genes in those functional epigenetic modules are epigenetically repressed. Strikingly, those frequently occurred COME events could be used to classify the patients into several subtypes with significant different clinical outcomes in six cancers as well as pan-cancer (p-value $\leq = 0.05$). Moreover, we observed significant associations between different COME subtypes and clinical features (e.g., age, gender, histological type, neoplasm histologic grade, and pathologic stage) in distinct cancers. Taken together, we identified millions of COME events of DNA methylation in pan-cancer and detected functional epigenetic COME events that could separate tumor patients into different subtypes, which may benefit the diagnosis and prognosis of pan-cancer.

Keywords: DNA methylation, co-methylation, mutual exclusivity, survival analysis, pan-cancer, cancer diagnosis

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Xuexin Yu,
The University of Texas Southwestern Medical Center, United States
You Guo,
First Affiliated Hospital of Gannan Medical University, China

***Correspondence:**
Tieliu Shi
tieliushi@yahoo.com

† These authors have contributed equally to this work

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

Received: *27 November 2019*
Accepted: *10 January 2020*
Published: *29 January 2020*

Citation:
Ding W, Feng G, Hu Y, Chen G and Shi T (2020) Co-occurrence and Mutual Exclusivity Analysis of DNA Methylation Reveals Distinct Subtypes in Multiple Cancers. Front. Cell Dev. Biol. 8:20. doi: 10.3389/fcell.2020.00020

INTRODUCTION

DNA methylation (DNAm) is a major epigenetic modification, which is considered as an approach for disease diagnosis. An increasing number of studies have indicated that aberrant DNAm plays an important role in diverse diseases, especially cancers (Delpu et al., 2013; Stefansson et al., 2015; Tahara and Arisawa, 2015). For example, the hypermethylation of CpG island in promoter region

Abbreviations: BLCA, Bladder Urothelial Carcinoma; BRCA, Breast invasive carcinoma; COAD, Colon adenocarcinoma; ESCA, Esophageal carcinoma; HNSC, Head and Neck squamous cell carcinoma; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; PAAD, Pancreatic adenocarcinoma; PRAD, Prostate adenocarcinoma; THCA, Thyroid carcinoma; UCEC, Uterine Corpus Endometrial Carcinoma.

of tumor suppressor genes have been observed in pediatric acute myeloid leukemia (Tao et al., 2014), bladder (Garcia-Baquero et al., 2014) and adult brain tumors (Hill et al., 2011) as well as hepatocellular carcinoma (Revill et al., 2013), which may lead to proliferative advantages and aggressive phenotypes during tumorigenesis (Suva et al., 2013).

Previous studies showed that the co-occurrence of gene mutations is frequently observed in two or more genes that tend to have mutations simultaneously in cancer patients (Kang et al., 2008; Zhang et al., 2017). Genes that have mutually exclusive mutations are generally involved in the same biological process (Szczurek and Beerenwinkel, 2014). Genomic alterations targeting similar biological processes could be mutually redundant, with one alteration being able to disrupt the affected process, thus identifying mutual exclusive events may facilitate discovering unknown functional interactions (Zhang et al., 2017). Detecting such patterns is crucial for identifying related novel cancerous pathways and potential treatment targets (Szczurek and Beerenwinkel, 2014). However, to date, co-occurrence (CO) and mutually exclusivity (ME) of DNA methylation in human cancers are less explored. Co-methylation has been reported as a new indicator for discovering functional associations between gene pairs in breast cancer (Akulenko and Helms, 2013). Recently, a number of algorithms have been developed for estimating the significance of ME and CO patterns between two genes (Canisius et al., 2016; Hua et al., 2016; Kim et al., 2017). Some of those tools can be used on DNA methylation data (Canisius et al., 2016), making it possible to comprehensively investigate the CO and ME events of DNAm in diverse The Cancer Genome Atlas (TCGA) cancers.

In this study, we first detected the CO and ME events of DNAm in 14 distinct cancers and explored the relationship between related gene pairs at gene expression and DNA methylation level. Then, we constructed a pan-cancer network with filtered Co-occurrence and mutual exclusivity (COME) events and identified functional epigenetic modules consisting of genes in the zinc finger family. We also found that the selected CO and ME events could be used to classify different types of tumors including pan-cancer into several subtypes with significantly different progression-free interval (PFI). Interestingly, the subtypes determined by COME events are significantly correlated with distinct clinical features, including age, gender, histological type, neoplasm histologic grade, and pathologic stage. Our results suggest that the COME events of DNA methylation could play important roles in tumorigenesis and may benefit the prognosis of different cancers.

MATERIALS AND METHODS

Data Source

The gene expression and DNA methylation data of the TCGA project were downloaded from UCSC Xena[1] and preprocessed as we previously described (Ding et al., 2019; Ji et al., 2019). The clinical data matrix of TCGA cancers was downloaded from the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) (Liu et al., 2018). The statistics of clinical information in 14 cancer types are listed in **Supplementary Table S1**.

Definition of Methylated and Unmethylated Genes

We first assigned the DNAm values for each gene with the average beta value of the probes mapped to promoter region, including TSS1500 (from −1,500 to −200 bp upstream of the TSS), TSS200 (region from −200 bp upstream to the transcription start site (TSS) itself), 1stExon (the first exon), and 5'UTR in order as previously described (Jiao et al., 2014; Sharma et al., 2016). According to previous studies (Sproul et al., 2011, 2012; Heyn et al., 2016), a beta value threshold of 0.3 was used to separate methylated from unmethylated probes. In this study, we defined methylated (average CpG DNAm beta values within gene promoter >0.3) and unmethylated (average CpG DNAm beta values within gene promoter <0.3) genes at the threshold of 0.3 in the lack of a better way to dichotomize continuous DNA methylation beta values.

Cancer Genes and Tumor Suppressor Genes

Cancer genes (CGs) were obtained from the database of CCGD (Abbott et al., 2015), DriverDB (Cheng et al., 2014) and CGC (v84) (Futreal et al., 2004), and tumor suppressor genes (TPG) were downloaded from TSGene (Zhao et al., 2016) database.

Identification of Co-occurrence and Mutual Exclusive Gene Pairs

We first convert the DNA methylation profile to a binary matrix, in which methylated genes were set to 1 in corresponding patients, and unmethylated genes were set to 0. Then we used DISCOVER (Canisius et al., 2016), a novel statistical independence test that assesses both COME gene pairs by counting how many samples have an alteration in both genes and comparing this to the number of samples expected to have such an overlap by chance if these alterations were independent. DISCOVER algorithm accepts a binary matrix that each row represents a gene and each column represents a patient as an input, then output the result of significant CO or ME gene pairs.

Filtration of COME Events

Firstly, Fisher's exact test was performed, for each COME event E_i, a contingency table (a, b, c, d) was created as bellow:

	Occurred	Not Occurred
Tumor	a	b
Normal	c	d

In the table, *a* and *b* denote the number of tumor samples in which event E_i occurred and not occurred, respectively, whereas *c* and *d* separately represent the number of normal samples in which event E_i occurred and not occurred respectively.

[1] http://xena.ucsc.edu

Then Fisher's exact test (SciPy package in Python) p-value was calculated to evaluate whether E_i was significantly differentially occurred in this cancer type. Finally, frequently occurred COME events were defined as the events that were significantly differentially occurred in at least three different cancer types.

Construction of FEM Models

The FEM algorithm (Jiao et al., 2014) is a functional supervised algorithm, which uses a network of relations between genes (in our case, is frequently occurred COME network) to identify subnetworks where a significant number of genes are associated with a phenotype of interest (POI, in our case, is the differential expression and differential methylation). Differential expression and differential methylation analysis were implemented inside the FEM algorithm.

Unsupervised Consensus Clustering and Survival Analysis

K-means clustering in R package ConsensusClusterPlus (Wilkerson and Hayes, 2010) was used to perform consensus clustering. The optimal cluster number k was chosen depending on the elbow and CDF curve (Senbabaoglu et al., 2014). For survival analysis of the pan-cancer, the best cluster number was chosen as the one with the maximum average silhouette coefficient. Python package lifelines[2] was implemented in survival analysis, and the log-rank test was used to estimate the significance of different groups.

Gene Ontology and KEGG Pathway Enrichment Analysis

Gene Ontology (GO) biological process and KEGG pathway enrichment analysis were performed using the web-based gene annotation tools DAVID (Huang da et al., 2009a,b) and ToppGene (Chen et al., 2009), the terms with FDR $\leq$ = 0.05 were considered as significant.

Statistical Analysis

All statistical analyses were performed with Python3.5.2 on anaconda3-4.0.0. Kruskal-Wallis H-test and Chi-square test were performed with Python package SciPy (Jones et al., 2014).

RESULTS

Overview of Co-occurrence and Mutual Exclusivity Network of DNA Methylation in Different Cancers

To construct the COME network of DNA methylation in cancers, we first dichotomized the DNA methylation beta values in every sample with threshold of 0.3, the genes with average beta value $\geq$ 0.3 in promoter region are designated as methylated while the genes with average beta value lower than 0.3 in promoter regions were considered as unmethylated (see section "Materials and Methods"). Thus, a binary matrix was built for each of 14 different cancers, in which 1 represents methylated and 0 for unmethylated. Then DISCOVER algorithm (Canisius et al., 2016) was employed to detect the CO and ME events based on the binary alteration matrix of DNA methylation. A total of 2,670,651 CO and 2,457,681 ME events that were identified as significant by DISCOVER in 14 cancers (q-value $\leq$ 0.05, q-value was calculated by DISCOVER). The expression correlation between genes pair of CO is significantly higher than those of ME (**Figure 1A**, p-value <0.001, independent Student's t-test). Moreover, gene pairs of CO events were mainly positively correlated at the DNA methylation level, whereas gene pairs of ME events were negatively correlated (**Figure 1B**). In addition, co-methylated gene pairs tend to co-expressed (**Figure 1C**, Pearson's correlation = 0.32, p-value = 0).

Then, to screen COME events that are associated with the tumor, Fisher exact test was performed in each cancer to screen the COME events that were significantly enriched in tumor or normal samples (p-value $\leq$ 0.05, see section "Materials and Methods"). After filtration, a total of 1,385,366 COME events (involving 7334 unique genes) were retained for further analyses [including 1,029,686 CO (involving 6894 genes) and 355,680 ME events (involving 6924 genes), the distribution of COME events in 14 cancers is shown in **Figure 1D**]. To explore the associations between COME events and cancers, we calculated the fraction of tumor suppressor genes (TSGs) and disease genes against all genes set involved in COME events. Strikingly, over 45% of the COME genes are cancer genes or TSGs (**Figure 1E**). We further constructed pan-cancer networks based on CO and ME events. Interestingly, 5 out of the top 10 hub genes in the network are cancer genes, including ELF3, SLC10A4, ANXA9, DEFB118, KRT8 (**Figure 1F**). Specifically, ELF3 was annotated as a cancer gene for rectum adenocarcinoma, colorectal neoplasms, hematologic diseases and breast neoplasms in the database of DriverDB (Cheng et al., 2014), CoReCG (Agarwal et al., 2016), and DDMGD (Bin Raies et al., 2015) respectively. Moreover, aberrant methylation of hub gene AGR2 was reported to be associated with ovarian cancer (Sung et al., 2014), while MT3 was a putative tumor suppressor gene in pediatric acute myeloid leukemia (Tao et al., 2014).

Gene Pairs of COME Events Tend to Be Significantly Correlated Between DNA Methylation and Gene Expression

To build a reliable network of COME events in each cancer, we further screened frequently occurred events from the above 1,385,366 COME events that were significantly enriched in tumor or normal samples, as frequently occurred events, we considered events that were differentially occurred in at least three different cancer types. After filtering, we found that gene expression correlation and DNA methylation correlation between gene pairs of COME events tend to be more correlated than that of before-filtering (Pearson's correlation r = 0.55, p-value = 0, **Supplementary Figure S1A**). The correlations between gene expression and DNA methylation of most genes involved in COME gene pairs tend to be more negatively correlated than that of randomly generated random gene sets

[2]http://lifelines.readthedocs.io/en/latest/index.html

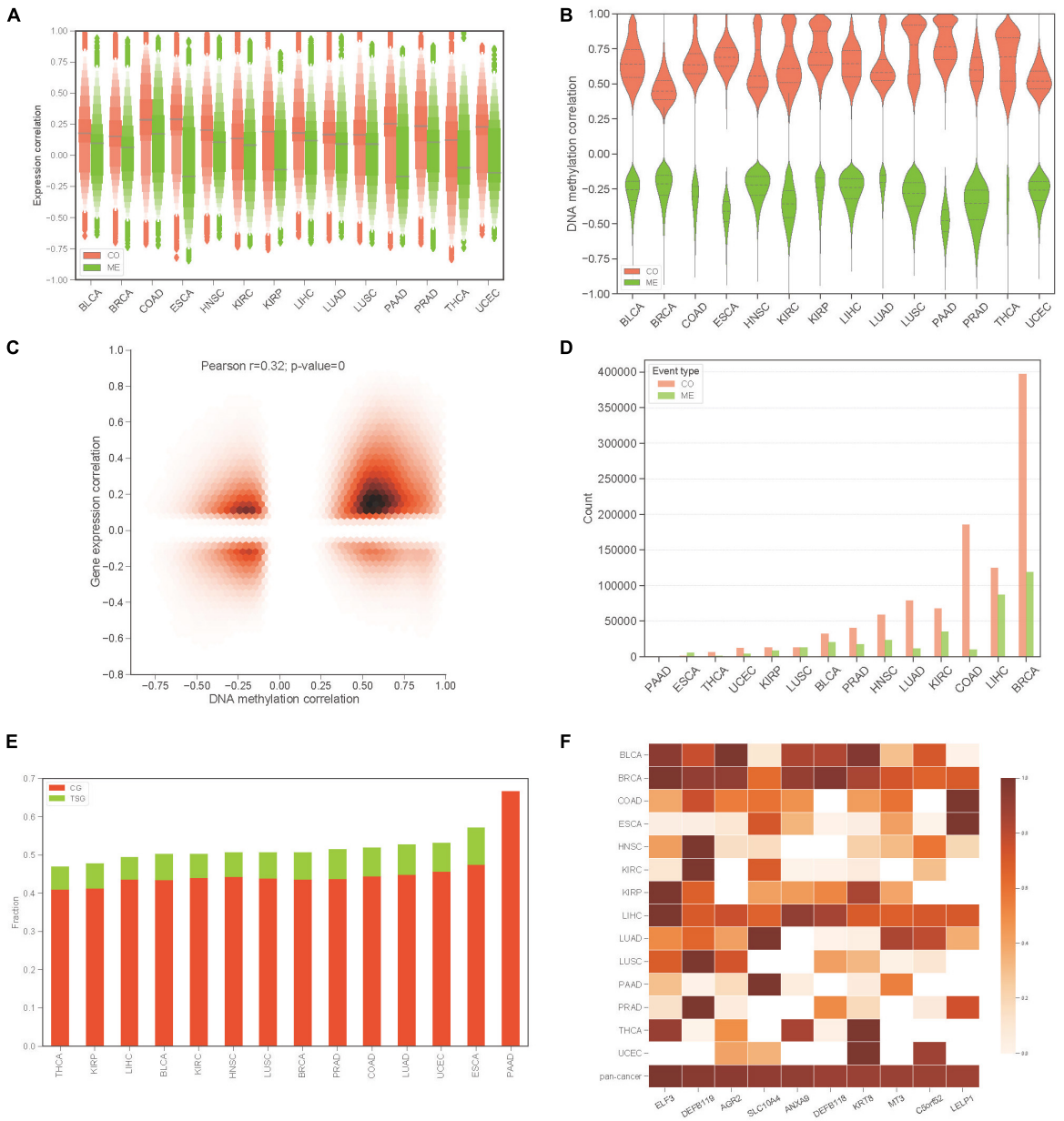


FIGURE 1 | Identification of co-occurrence and mutual exclusivity events of methylation. **(A)** Letter value (LV) plot of Pearson's correlation coefficient of gene expression level between gene pairs for CO and ME events. **(B)** Pearson's correlation plot of DNA methylation between gene pairs. **(C)** Jointplot of Pearson's correlation coefficient between gene expression and DNA methylation for COME gene pairs. **(D)** Countplot of COME events in 14 cancers. **(E)** The proportion of cancer genes and tumor suppressor genes against all of the genes involving in COME events in different cancers. **(F)** Degree distribution of top 10 genes in pan-cancer network for each cancer and pan-cancer.

(**Figure 2A**, random gene set was generated randomly with the same number of genes in each cancer). Gene functional enrichment analysis showed that the genes involved in CO events were mainly enriched in the pathways of Neuroactive ligand-receptor interaction, Nicotine addiction, Morphine addiction, cAMP signaling, Calcium signaling and the biological processes of chemical synaptic transmission, cell adhesion, neuropeptide signaling pathway and so on (**Figure 2B** and **Supplementary Table S2**). While the genes of ME events were mainly associated with the development of the central nervous system and brain (**Figure 2C** and **Supplementary Table S2**). We further built a pan-cancer cooperative network by merging the networks of each cancer (**Supplementary Figures S1B–D**). Intriguingly, most of the genes with a high degree (have the largest number of links) in the pan-cancer network showed enrichment for known cancer-related genes or tumor suppressor genes (**Figure 2D**). Aberrant DNA methylation of some of those top 10 genes with the highest degree has been reported to be associated with neoplasms, such

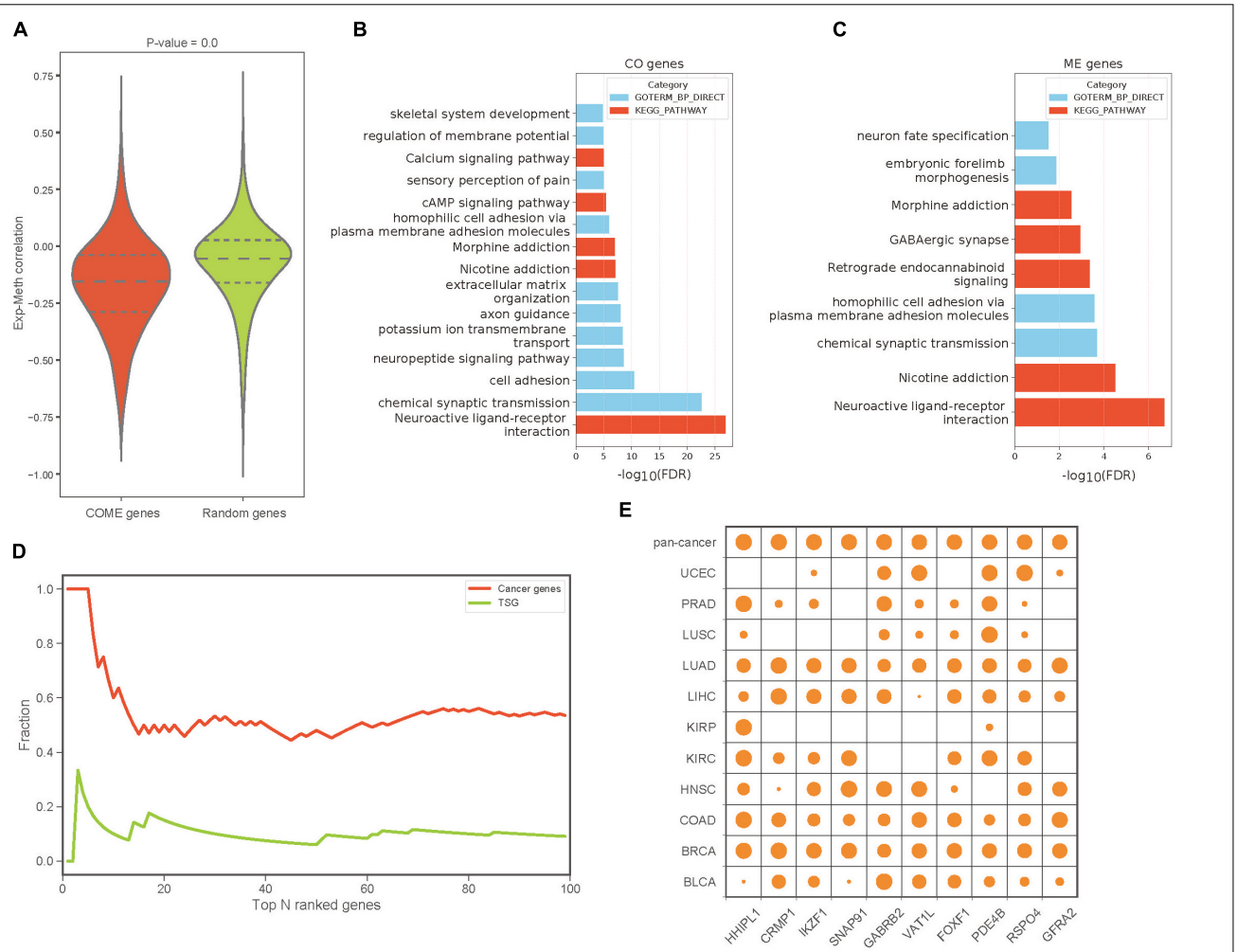


FIGURE 2 | The features and functions of gene pairs in COME events. **(A)** Box plot of Pearson's correlation coefficient between gene expression and DNA methylation for the genes involved in COME events compared with that of random genes set (*p*-value was calculated by independent Student's *t*-test). **(B,C)** Enriched KEGG pathways and biological processes for the genes involved in CO and ME events (DAVID online web server). **(D)** Fraction of cancer genes and tumor suppressor genes (TSG) against the top N genes ranked by the degree in pan-cancer network. **(E)** Degree distribution of top 10 genes in pan-cancer network for each cancer and pan-cancer.

as HHIPL1 (Duong et al., 2012), GABRB2 (Beltrami et al., 2017), FOXF1 (Lo et al., 2010) and RSPO4 (Oka et al., 2009: **Figure 2E**).

Zinc Fingers Gene Family Is Enriched in Functional Epigenetic Modules

To identify functional epigenetic modules, we integrated the gene expression and DNA methylation data from TCGA and frequently occurred COME networks constructed in each cancer using the FEM algorithm (Jiao et al., 2014), which can be used to effectively identify gene modules of coordinated differential methylation and differential expression in the context of a network. Many functional epigenetic modules were identified by FEM. Remarkably, six modules enriched in the zinc fingers gene family were identified in 6 distinct cancer types (**Figures 3A–F**). Most genes in these modules were hypermethylated and down-regulated, indicating that genes of zinc fingers family may tend to be co-methylated and transcriptionally suppressed.

To further explore the associations between aberrant DNA methylation of zinc fingers gene family and neoplasms, we found that those genes were significantly enriched in regulation of transcription, DNA binding transcription factor activity, RNA polymerase II regulatory region sequence-specific DNA binding, Neuroactive ligand-receptor interaction, and so on (**Supplementary Table S3**). Besides, many of the genes in these modules were enriched in cytoband of 19q13.43, transcription factor binding sites of ZNF274 and they tend to have similar DNA methylation patterns.

We also examined whether those genes in the zinc fingers gene family can be used to distinguish tumor samples from normal samples in the above 14 cancers. Six genes (ZIK1, ZNF471, ZNF229, ZFP28, ZNF677, and ZNF582) shared among 6 modules were selected to build a logistic regression module. Compared with the models based on gene expression or DNA methylation along (AUC = 0.76), the module integrated both

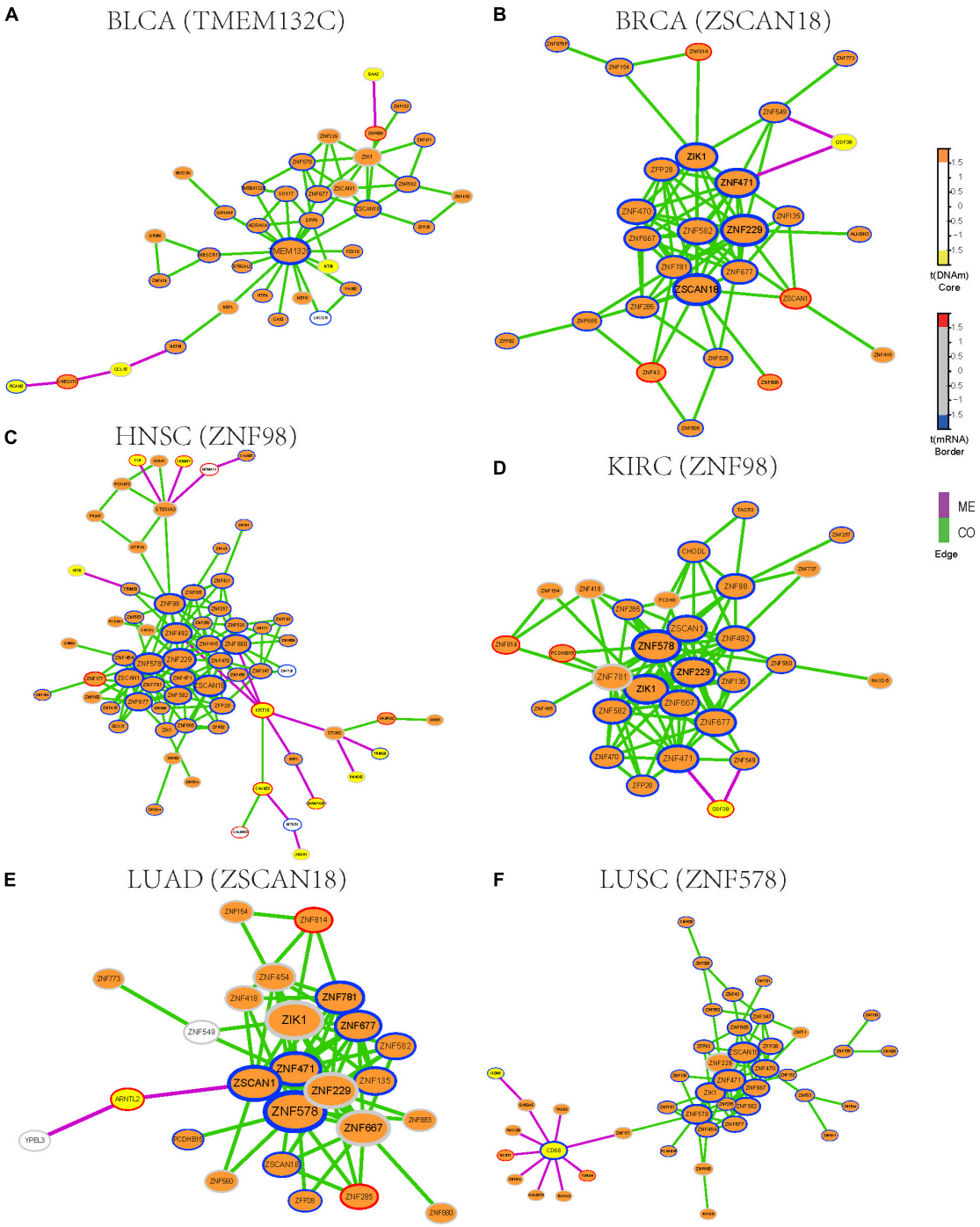


FIGURE 3 | Functional epigenetic modules identified from CO and ME network. The color of core indicate significant DNA methylation changes, color of border represent significant gene expression changes, edges represent COME events between two genes, and edge color indicates different event type (co-occurrence or mutual exclusivity). **(A–F)** Functional epigenetic modules identified from corresponding cancer type.

DNA methylation and gene expression data of the 6 genes showed better performance in distinguishing tumor samples from normal (AUC = 0.86, **Supplementary Figure S2**). Moreover, the clustering result also indicates that gene expression and DNA methylation profile of those 6 genes can effectively separate tumor samples from normal samples (**Supplementary Figure S3**). Most of those 6 genes are involved in function of nucleic acid binding, and some of those genes have been reported to be aberrantly methylated in tumors [such as ZIK1 (Borinstein et al., 2010), ZNF471 (Bhat et al., 2017)] and may serve as a marker of

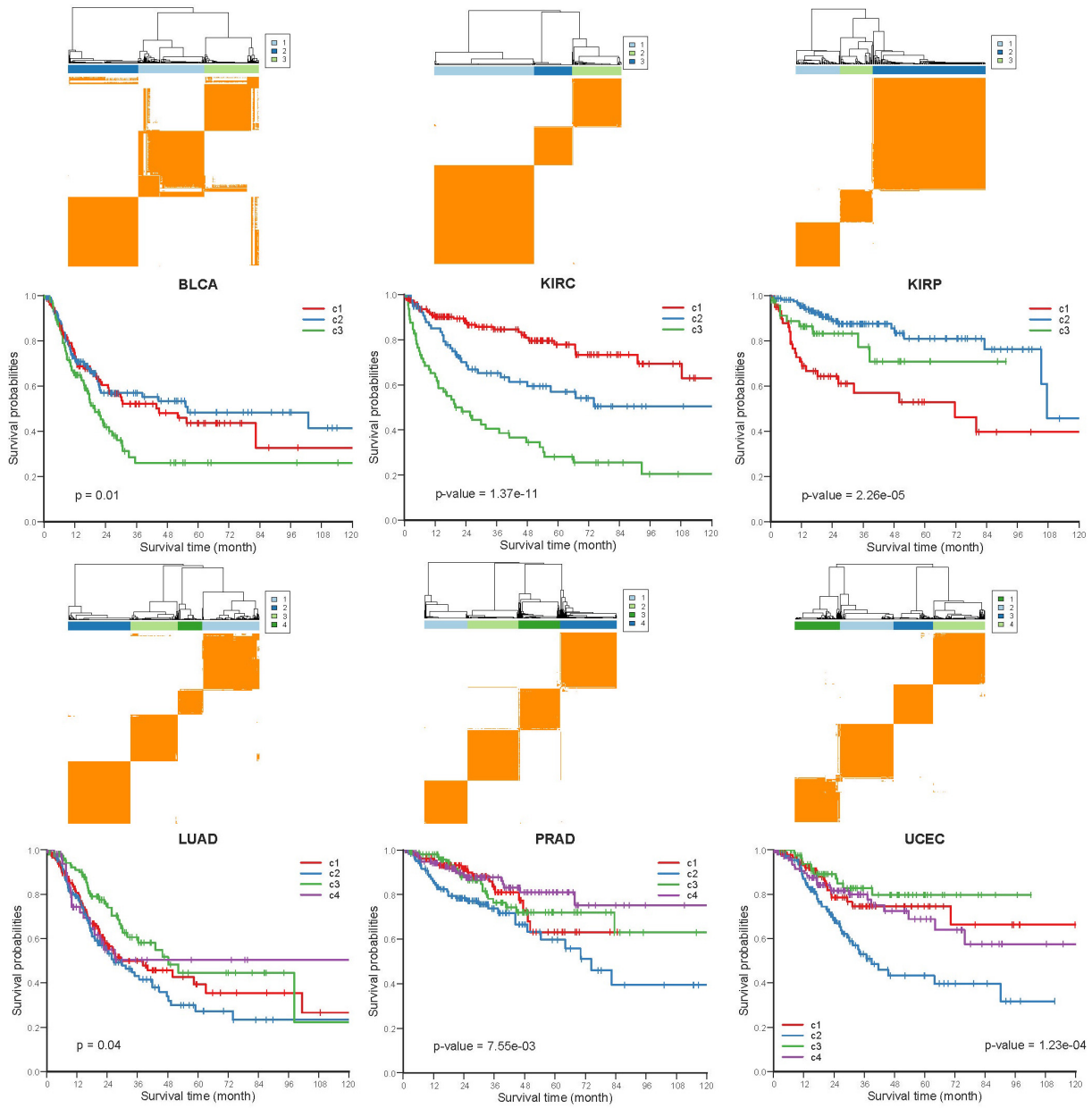


FIGURE 4 | Kaplan-Meier plot of PFI for distinct subtypes of 6 different cancers. Consensus cluster plot (top) and Kaplan-Meier survival plots (bottom) were separately shown for 6 disparate cancers, c1-c4 represent cluster 1 – cluster 4, and *p*-values were calculated by log rank test.

cancer [e.g., ZNF677 (Heller et al., 2015), ZNF582 (Lin et al., 2014)]. Consequently, our finding demonstrates that combing DNA methylation and gene expression data of those genes from zinc fingers family may be associated with tumorigenesis.

Co-occurrence and Mutual Exclusive Events Contribute to Prognosis in Human Cancers

To investigate whether the COME events are associated with cancer prognosis, Frequently occurred COME events (occurred in at least 3 distinct cancer types, **Supplementary Table S4**) were used to perform consensus cluster and Kaplan-Meier survival analysis in each of those 13 cancers (no frequently occurred COME event was left in pancreatic adenocarcinoma, PAAD). Strikingly, we observed significantly different PFI among disparate subtypes in 6 different cancer types (p-value <0.05, log-rank test, **Figure 4**), as well as in BRCA and lung squamous cell carcinoma (LUSC) (p-value = 0.07, **Supplementary Figure S4**).

Next, we performed Kruskal-Wallis H-test and Chi-square test to explore the association between COME subtypes and different clinical information (including age, gender, histological type,

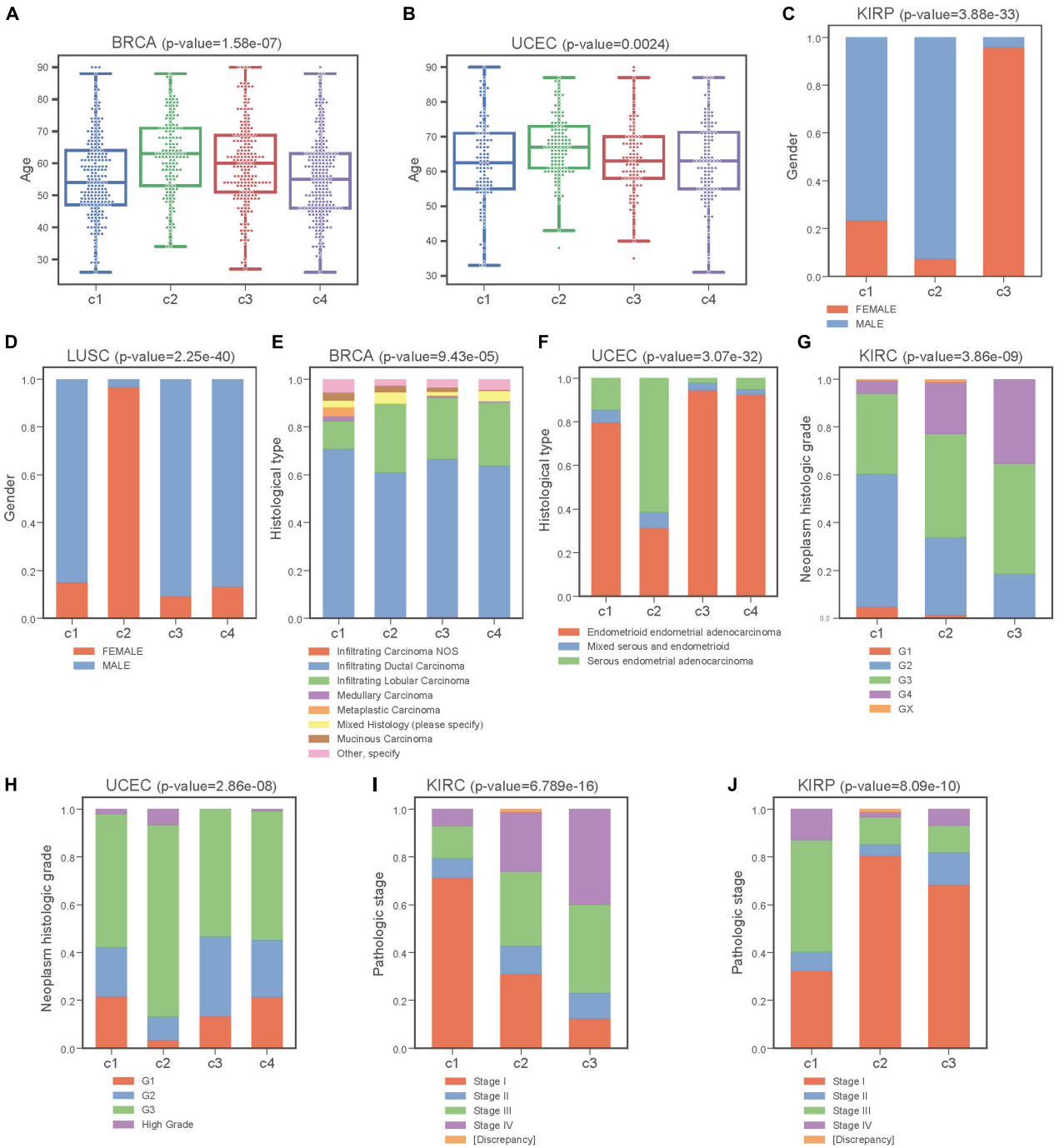


FIGURE 5 | COME subtypes correlate with distinct clinical features. **(A,B)** Age distribution of different COME subtypes identified in BRCA and UCEC, respectively. **(C–J)** Distribution of gender, histological type, neoplasm histologic grade and pathologic stage against COME subtypes in corresponding TCGA cancers. *p*-values for continuous variable were calculated by using Kruskal-Wallis *H*-test **(A,B)**, and *p*-values for categorical variable were calculated by performing Chi-square test **(C–J)**. *p*-value were calculated between all different groups.

neoplasm histologic grade, and pathologic stage). Interestingly, significant difference was observed between different clinical features and different subtypes in 5 cancers (**Figure 5**). For example, the distribution of age was significantly different in different COME subtypes in BRCA and uterine corpus endometrial carcinoma (UCEC) (**Figures 5A,B**, Kruskal-Wallis *H*-test, *p*-value <0.001), which is in accordance with previous studies that DNA methylation is associated with age (Horvath, 2013; Johnson et al., 2014). In kidney renal papillary cell carcinoma (KIRP), cluster 3 is enriched with female whereas

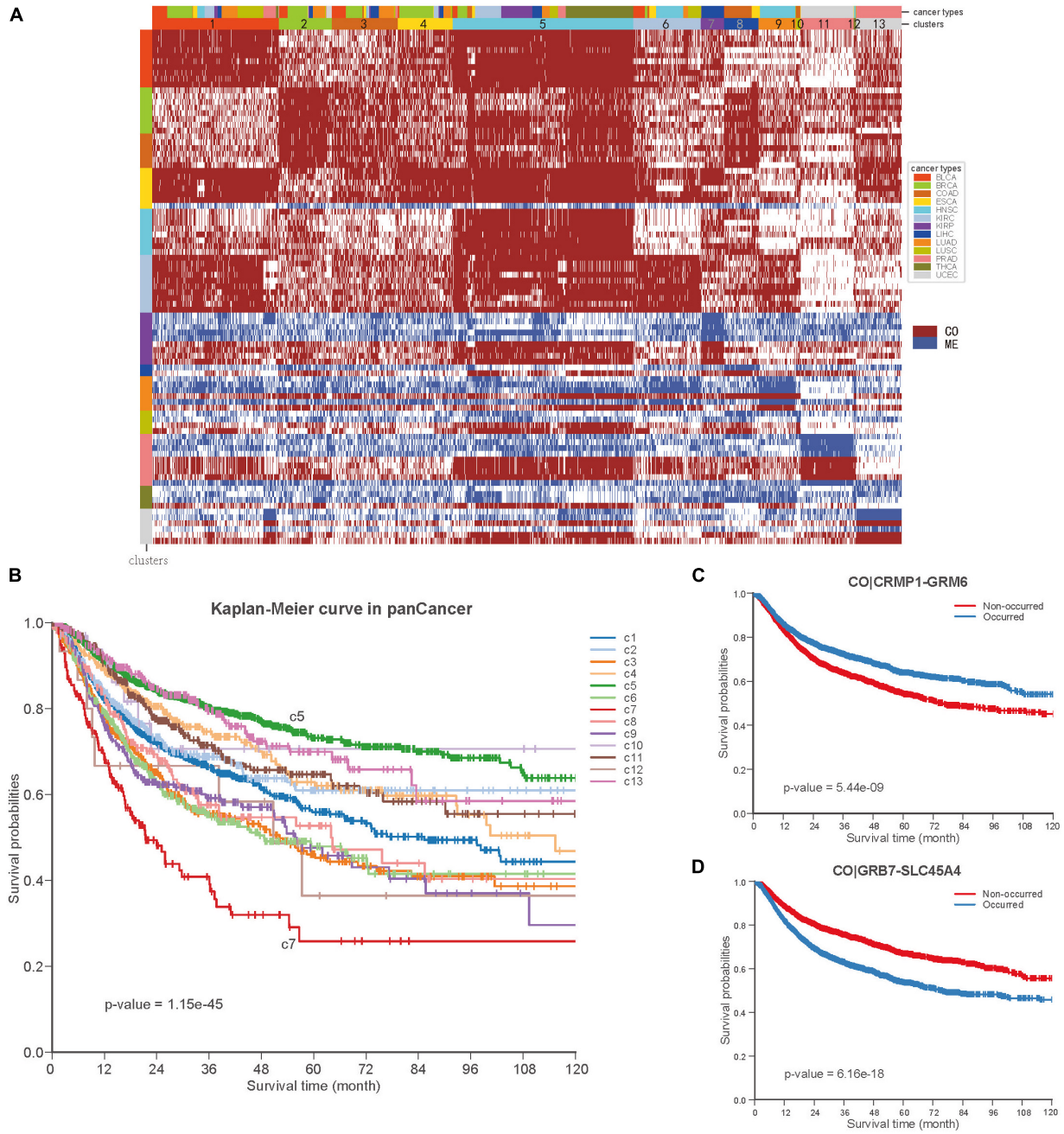


FIGURE 6 | Co-occurrence and mutual exclusive events contribute to prognosis of pan-cancer. **(A)** Clustering heatmap based on top 15 enriched COME events in 13 clusters. The clusters are denoted by number and color in the second bar, and tissue of origin specified in the first color bar. The color in the left identified top 15 COME pairs significantly enriched in corresponding cluster. **(B)** Kaplan-Meier plot for the PFI of 13 clusters identified in pan-cancer. **(C,D)** Kaplan-Meier plot of PFI for co-methylation event of CRMP1-GRM6 and GRB7-SLC45A4 in pan-cancer, respect.

cluster 2 is enriched with male (**Figure 5C**). Moreover, cluster 2 of LUSC is significantly enriched with female, which is opposite to clusters 1, 3, and 4 (**Figure 5D**). We also found that different COME subtypes have distinct distribution of histological types in BRCA and UCEC (**Figures 5E,F**). Furthermore, cluster 2 of UCEC is enriched with the histological type of serous endometrial adenocarcinoma and cluster 2 has a shorter PFI compared to other clusters (**Figure 4**). As for neoplasm histologic grade (**Figures 5G,H**), cluster 3 of kidney renal clear cell carcinoma (KIRC) is enriched in grade G3 and G4, and poorer prognosis was observed in this cluster (**Figure 4**). Similarly, G3 and high-grade patients are enriched in cluster 2 of UCEC, while cluster 2 has poorer survival probability compared to other subtypes (**Figure 4**). Patients of stage III and stage IV are enriched

in cluster 3 and has the poorest clinical outcome in KIRC, while cluster 1 is enriched with the patients of stage III and stage IV and has the shortest PFI in KIRP (**Figures 4**, **5I,J**). Collectively, the subtypes determined by COME pattern are correlated with various clinical features (age, gender, histological type, neoplasm histologic grade, and pathologic stage), which may explain why distinct COME subtypes have significantly different PFI.

To further explore the COME events in pan-cancer, we performed consensus clustering based on the frequently occurred COME events in 5442 tumor samples of 13 types of cancer. Thirteen clusters were identified by maximizing the average silhouette coefficient. The top 15 significantly enriched COME events in each cluster are shown in **Figure 6A**. Most of these clusters were significantly correlated with cancer tissue of origin (p-value <0.0001, Chi-square test, **Supplementary Table S5**). For example, clusters of C7, C9, C10, C11, and C13 were significantly enriched with patients from LIHC (liver hepatocellular carcinoma), HNSC (head and neck squamous cell carcinoma), COAD (colon adenocarcinoma), UCEC and PRAD (prostate adenocarcinoma), respectively (p-value <0.0001, hypergeometric test). While some other clusters, such as C1, contained a mixture type of BLCA (bladder urothelial carcinoma), BRCA, LUAD (lung adenocarcinoma), and LUSC (p-value <0.0001, hypergeometric test). Furthermore, those 13 clusters exhibited significantly different PFI via Kaplan-Meier analysis (p-value <0.0001, log-rank test, **Figure 6B**). Among the 13 clusters, cluster C5 exhibited the best prognosis and the co-occurrence of CRMP1-GRM6 was the most significantly enriched event in this cluster (p-value <0.001, hypergeometric test). CRMP1 (collapsin response mediator protein 1) has been reported to be associated with medulloblastoma (Li et al., 2015) and gliomas (Mukherjee et al., 2009). Hypermethylation of the CpG sites on GRM6 (glutamate metabotropic receptor 6) was reported to be a hallmark of CIMP in clear cell renal cell carcinomas (Arai et al., 2012). In contrast, cluster C7 was associated with the poorest prognosis and co-methylation of GRB7-SLC45A4 was enriched in this group (p-value <0.001, hypergeometric test). GRB7 (growth factor receptor bound protein 7) was reported to play an important role in breast cancer progression (Lim et al., 2014). We further performed Kaplan-Meier survival analysis in the pan-cancer to verify whether the co-methylation of CRMP1-GRM6 and GRB7-SLC45A4 was associated with clinical outcome in pan-cancer. The result shows that patients with co-methylation of CRMP1-GRM6 have better outcomes (p-value <0.0001, log-rank test, **Figure 6C**), whereas patients with co-methylation of GRB7-SLC45A4 exhibit significantly poorer prognosis (p-value <0.0001, log-rank test, **Figure 6D**).

DISCUSSION

In our study, we first identified 2,670,651 CO and 2,457,681 ME gene pairs in 14 different cancers based on the methylation profile from the TCGA project. Interestingly, the genes in functional epigenetic modules identified in six cancer types were mainly from the zinc finger gene family, and most of those genes were epigenetically repressed. Although several studies have reported the epigenetic silencing of the zinc finger gene family (Lleras et al., 2011; Severson et al., 2013; Gaykalova et al., 2015), we are the first to identify functional epigenetic modules of the zinc finger gene family in six cancer types by integrating gene expression and DNA methylation data in the context of COME networks. Methylation was reported to be the main mechanism for downregulation of tumor cell growth suppressor ZNF677 in non-small cell lung cancers (NSCLCs) and the methylation of ZNF677 could be used in the prognosis of NSCLCs (Heller et al., 2015). Furthermore, we identified a set of COME events that can divide tumor patients into different subtypes with significantly different clinical outcomes. Different COME subtypes were found to be significantly associated with distinct clinical features, such as age, gender, histological type, neoplasm histologic grade and pathologic stage. We also found that COME events could be used to divide tumor samples of pan-cancer into different subtypes with significantly different outcomes, which may benefit the prediction of the prognosis for pan-cancer.

This study is just the beginning to investigate and characterize the roles of COME of DNA methylation in human cancers. Our findings may contribute to the diagnosis and prognosis of human pan-cancer. The underlying mechanism and function of COME events in diverse cancers are still needed to be further studied in the future.

DATA AVAILABILITY STATEMENT

The TCGA data analyzed in this study was obtained from UCSC Xena (http://xena.ucsc.edu). This study fully complies with the TCGA publication requirements (http://cancergenome.nih.gov/publications/publicationguidelines).

AUTHOR CONTRIBUTIONS

TS and WD designed the study and wrote the manuscript. WD and YH conducted the data analysis. WD, GC, and TS revised and finalized the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Science Foundation of China (31671377 and 91629103), the Beihang University and Capital Medical University Advanced Innovation Center for Big Data-Based Precision Medicine Plan (BHME-201801), Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01), and Shanghai 111 Project (B14019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.00020/full#supplementary-material

REFERENCES

Abbott, K. L., Nyre, E. T., Abrahante, J., Ho, Y. Y., Isaksson Vogel, R., and Starr, T. K. (2015). The candidate cancer gene database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.* 43, D844–D848. doi: 10.1093/nar/gku770

Agarwal, R., Kumar, B., Jayadev, M., Raghav, D., and Singh, A. (2016). CoReCG: a comprehensive database of genes associated with colon-rectal cancer. *Database* 2016:baw059. doi: 10.1093/database/baw059

Akulenko, R., and Helms, V. (2013). DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Hum. Mol. Genet.* 22, 3016–3022. doi: 10.1093/hmg/ddt158

Arai, E., Chiku, S., Mori, T., Gotoh, M., Nakagawa, T., Fujimoto, H., et al. (2012). Single-CpG-resolution methylome analysis identifies clinicopathologically aggressive CpG island methylator phenotype clear cell renal cell carcinomas. *Carcinogenesis* 33, 1487–1493. doi: 10.1093/carcin/bgs177

Beltrami, C. M., Dos Reis, M. B., Barros-Filho, M. C., Marchi, F. A., Kuasne, H., Pinto, C. A. L., et al. (2017). Integrated data analysis reveals potential drivers and pathways disrupted by DNA methylation in papillary thyroid carcinomas. *Clin. Epigenet.* 9:45. doi: 10.1186/s13148-017-0346-2

Bhat, S., Kabekkodu, S. P., Jayaprakash, C., Radhakrishnan, R., Ray, S., and Satyamoorthy, K. (2017). Gene promoter-associated CpG island hypermethylation in squamous cell carcinoma of the tongue. *Virchows Arch.* 470, 445–454. doi: 10.1007/s00428-017-2094-2

Bin Raies, A., Mansour, H., Incitti, R., and Bajic, V. B. (2015). DDMGD: the database of text-mined associations between genes methylated in diseases from different species. *Nucleic Acids Res.* 43, D879–D886. doi: 10.1093/nar/gku1168

Borinstein, S. C., Conerly, M., Dzieciatkowski, S., Biswas, S., Washington, M. K., Trobridge, P., et al. (2010). Aberrant DNA methylation occurs in colon neoplasms arising in the azoxymethane colon cancer model. *Mol. Carcinog.* 49, 94–103. doi: 10.1002/mc.20581

Canisius, S., Martens, J. W., and Wessels, L. F. (2016). A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol.* 17:261. doi: 10.1186/s13059-016-1114-x

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi: 10.1093/nar/gkp427

Cheng, W. C., Chung, I. F., Chen, C. Y., Sun, H. J., Fen, J. J., Tang, W. C., et al. (2014). DriverDB: an exome sequencing database for cancer driver gene identification. *Nucleic Acids Res.* 42, D1048–D1054. doi: 10.1093/nar/gkt1025

Delpu, Y., Cordelier, P., Cho, W. C., and Torrisani, J. (2013). DNA methylation and cancer diagnosis. *Int. J. Mol. Sci.* 14, 15029–15058. doi: 10.3390/ijms140715029

Ding, W., Chen, G., and Shi, T. (2019). Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. *Epigenetics* 14, 67–80. doi: 10.1080/15592294.2019.1568178

Duong, C. V., Emes, R. D., Wessely, F., Yacqub-Usman, K., Clayton, R. N., and Farrell, W. E. (2012). Quantitative, genome-wide analysis of the DNA methylome in sporadic pituitary adenomas. *Endocr. Relat. Cancer* 19, 805–816. doi: 10.1530/ERC-12-0251

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Garcia-Baquero, R., Puerta, P., Beltran, M., Alvarez-Mujica, M., Alvarez-Ossorio, J. L., and Sanchez-Carbayo, M. (2014). Methylation of tumor suppressor genes in a novel panel predicts clinical outcome in paraffin-embedded bladder tumors. *Tumour Biol.* 35, 5777–5786. doi: 10.1007/s13277-014-1767-6

Gaykalova, D. A., Vatapalli, R., Wei, Y., Tsai, H. L., Wang, H., Zhang, C., et al. (2015). Outlier analysis defines zinc finger gene family DNA methylation in tumors and saliva of head and neck cancer patients. *PLoS One* 10:e0142148. doi: 10.1371/journal.pone.0142148

Heller, G., Altenberger, C., Schmid, B., Marhold, M., Tomasich, E., Ziegler, B., et al. (2015). DNA methylation transcriptionally regulates the putative tumor cell growth suppressor ZNF677 in non-small cell lung cancers. *Oncotarget* 6, 394–408. doi: 10.18632/oncotarget.2697

Heyn, H., Vidal, E., Ferreira, H. J., Vizoso, M., Sayols, S., Gomez, A., et al. (2016). Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol.* 17:11. doi: 10.1186/s13059-016-0879-2

Hill, V. K., Underhill-Day, N., Krex, D., Robel, K., Sangan, C. B., Summersgill, H. R., et al. (2011). Epigenetic inactivation of the RASSF10 candidate tumor suppressor gene is a frequent and an early event in gliomagenesis. *Oncogene* 30, 978–989. doi: 10.1038/onc.2010.471

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* 14:R115. doi: 10.1186/gb-2013-14-10-r115

Hua, X., Hyland, P. L., Huang, J., Song, L., Zhu, B., Caporaso, N. E., et al. (2016). MEGSA: a powerful and flexible framework for analyzing mutual exclusivity of tumor mutations. *Am. J. Hum. Genet.* 98, 442–455. doi: 10.1016/j.ajhg.2015.12.021

Huang da, 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, 1–13. doi: 10.1093/nar/gkn923

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc* 4, 44–57. doi: 10.1038/nprot.2008.211

Ji, X., Tong, W., Ning, B., Mason, C. E., Kreil, D. P., Labaj, P. P., et al. (2019). QuaPra: efficient transcript assembly and quantification using quadratic programming with Apriori algorithm. *Sci. China Life Sci.* 62, 937–946. doi: 10.1007/s11427-018-9433-3

Jiao, Y., Widschwendter, M., and Teschendorff, A. E. (2014). A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* 30, 2360–2366. doi: 10.1093/bioinformatics/btu316

Johnson, K. C., Koestler, D. C., Cheng, C., and Christensen, B. C. (2014). Age-related DNA methylation in normal breast tissue and its relationship with invasive breast tumor methylation. *Epigenetics* 9, 268–275. doi: 10.4161/epi.27015

Jones, E., Oliphant, T., and Peterson, P. (2014). *{SciPy}: Open Source scientific Tools for {Python}*.

Kang, S., Seo, S. S., Chang, H. J., Yoo, C. W., Park, S. Y., and Dong, S. M. (2008). Mutual exclusiveness between PIK3CA and KRAS mutations in endometrial carcinoma. *Int. J. Gynecol. Cancer* 18, 1339–1343. doi: 10.1111/j.1525-1438.2007.01172.x

Kim, Y. A., Madan, S., and Przytycka, T. M. (2017). WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* 33, 814–821. doi: 10.1093/bioinformatics/btw242

Li, K. K., Qi, Y., Xia, T., Yao, Y., Zhou, L., Lau, K. M., et al. (2015). CRMP1 inhibits proliferation of medulloblastoma and is regulated by HMGA1. *PLoS One* 10:e0127910. doi: 10.1371/journal.pone.0127910

Lim, R. C., Price, J. T., and Wilce, J. A. (2014). Context-dependent role of Grb7 in HER2+ve and triple-negative breast cancer cell lines. *Breast Cancer Res. Treat.* 143, 593–603. doi: 10.1007/s10549-014-2838-5

Lin, H., Chen, T. C., Chang, T. C., Cheng, Y. M., Chen, C. H., Chu, T. Y., et al. (2014). Methylated ZNF582 gene as a marker for triage of women with Pap smear reporting low-grade squamous intraepithelial lesions - a Taiwanese Gynecologic Oncology Group (TGOG) study. *Gynecol. Oncol.* 135, 64–68. doi: 10.1016/j.ygyno.2014.08.012

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.e11. doi: 10.1016/j.cell.2018.02.052

Lleras, R. A., Adrien, L. R., Smith, R. V., Brown, B., Jivraj, N., Keller, C., et al. (2011). Hypermethylation of a cluster of Kruppel-type zinc finger protein genes on chromosome 19q13 in oropharyngeal squamous cell carcinoma. *Am. J. Pathol.* 178, 1965–1974. doi: 10.1016/j.ajpath.2011.01.049

Lo, P. K., Lee, J. S., Liang, X., Han, L., Mori, T., Fackler, M. J., et al. (2010). Epigenetic inactivation of the potential tumor suppressor gene FOXF1 in breast cancer. *Cancer Res.* 70, 6047–6058. doi: 10.1158/0008-5472.CAN-10-1576

Mukherjee, J., DeSouza, L. V., Micallef, J., Karim, Z., Croul, S., Siu, K. W., et al. (2009). Loss of collapsin response mediator Protein1, as detected by iTRAQ analysis, promotes invasion of human gliomas expressing mutant EGFRvIII. *Cancer Res.* 69, 8545–8554. doi: 10.1158/0008-5472.CAN-09-1778

Oka, D., Yamashita, S., Tomioka, T., Nakanishi, Y., Kato, H., Kaminishi, M., et al. (2009). The presence of aberrant DNA methylation in noncancerous esophageal mucosae in association with smoking history: a target for risk diagnosis and prevention of esophageal cancers. *Cancer* 115, 3412–3426. doi: 10.1002/cncr.24394

Revill, K., Wang, T., Lachenmayer, A., Kojima, K., Harrington, A., Li, J., et al. (2013). Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. *Gastroenterology* 145, e1–e25. doi: 10.1053/j.gastro.2013.08.055

Senbabaoglu, Y., Michailidis, G., and Li, J. Z. (2014). Critical limitations of consensus clustering in class discovery. *Sci. Rep.* 4:6207. doi: 10.1038/srep06207

Severson, P. L., Tokar, E. J., Vrba, L., Waalkes, M. P., and Futscher, B. W. (2013). Coordinate H3K9 and DNA methylation silencing of ZNFs in toxicant-induced malignant transformation. *Epigenetics* 8, 1080–1088. doi: 10.4161/epi.25926

Sharma, P., Bhunia, S., Poojary, S. S., Tekcham, D. S., Barbhuiya, M. A., Gupta, S., et al. (2016). Global methylation profiling to identify epigenetic signature of gallbladder cancer and gallstone disease. *Tumour Biology* 37, 14687–14699. doi: 10.1007/s13277-016-5355-9

Sproul, D., Kitchen, R. R., Nestor, C. E., Dixon, J. M., Sims, A. H., Harrison, D. J., et al. (2012). Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol.* 13:R84.

Sproul, D., Nestor, C., Culley, J., Dickson, J. H., Dixon, J. M., Harrison, D. J., et al. (2011). Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proc. Natl. Acad. Sci. U.S.A* 108, 4364–4369. doi: 10.1073/pnas.1013224108

Stefansson, O. A., Moran, S., Gomez, A., Sayols, S., Arribas-Jorba, C., Sandoval, J., et al. (2015). A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Mol. Oncol.* 9, 555–568. doi: 10.1016/j.molonc.2014.10.012

Sung, H. Y., Choi, E. N., Lyu, D., Park, A. K., Ju, W., and Ahn, J. H. (2014). Aberrant hypomethylation-mediated AGR2 overexpression induces an aggressive phenotype in ovarian cancer cells. *Oncol. Rep.* 32, 815–820. doi: 10.3892/or.2014.3243

Suva, M. L., Riggi, N., and Bernstein, B. E. (2013). Epigenetic reprogramming in cancer. *Science* 339, 1567–1570. doi: 10.1126/science.1230184

Szczurek, E., and Beerenwinkel, N. (2014). Modeling mutual exclusivity of cancer mutations. *PLoS Computat. Biol.* 10:e1003503. doi: 10.1371/journal.pcbi.1003503

Tahara, T., and Arisawa, T. (2015). DNA methylation as a molecular biomarker in gastric cancer. *Epigenomics* 7, 475–486. doi: 10.2217/epi.15.4

Tao, Y. F., Xu, L. X., Lu, J., Cao, L., Li, Z. H., Hu, S. Y., et al. (2014). Metallothionein III (MT3) is a putative tumor suppressor gene that is frequently inactivated in pediatric acute myeloid leukemia by promoter hypermethylation. *J. Transl. Med.* 12:182. doi: 10.1186/1479-5876-12-182

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

Zhang, H., Deng, Y., Zhang, Y., Ping, Y., Zhao, H., Pang, L., et al. (2017). Cooperative genomic alteration network reveals molecular classification across 12 major cancer types. *Nucleic Acids Res.* 45, 567–582. doi: 10.1093/nar/gkw1087

Zhao, M., Kim, P., Mitra, R., Zhao, J., and Zhao, Z. (2016). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* 44, D1023–D1031. doi: 10.1093/nar/gkv1268

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 © 2020 Ding, Feng, Hu, Chen and Shi. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 February 2020
doi: 10.3389/fbioe.2020.00040

A Computational Study of Potential miRNA-Disease Association Inference Based on Ensemble Learning and Kernel Ridge Regression

Li-Hong Peng[1], Li-Qian Zhou[1], Xing Chen[2]* and Xue Piao[3]**

[1] School of Computer Science, Hunan University of Technology, Zhuzhou, China, [2] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China, [3] School of Medical Informatics, Xuzhou Medical University, Xuzhou, China

As increasing experimental studies have shown that microRNAs (miRNAs) are closely related to multiple biological processes and the prevention, diagnosis and treatment of human diseases, a growing number of researchers are focusing on the identification of associations between miRNAs and diseases. Identifying such associations purely via experiments is costly and demanding, which prompts researchers to develop computational methods to complement the experiments. In this paper, a novel prediction model named Ensemble of Kernel Ridge Regression based MiRNA-Disease Association prediction (EKRRMDA) was developed. EKRRMDA obtained features of miRNAs and diseases by integrating the disease semantic similarity, the miRNA functional similarity and the Gaussian interaction profile kernel similarity for diseases and miRNAs. Under the computational framework that utilized ensemble learning and feature dimensionality reduction, multiple base classifiers that combined two Kernel Ridge Regression classifiers from the miRNA side and disease side, respectively, were obtained based on random selection of features. Then average strategy for these base classifiers was adopted to obtain final association scores of miRNA-disease pairs. In the global and local leave-one-out cross validation, EKRRMDA attained the AUCs of 0.9314 and 0.8618, respectively. Moreover, the model's average AUC with standard deviation in 5-fold cross validation was 0.9275 ± 0.0008. In addition, we implemented three different types of case studies on predicting miRNAs associated with five important diseases. As a result, there were 90% (Esophageal Neoplasms), 86% (Kidney Neoplasms), 86% (Lymphoma), 98% (Lung Neoplasms), and 96% (Breast Neoplasms) of the top 50 predicted miRNAs verified to have associations with these diseases.

Keywords: miRNA, disease, association prediction, ensemble, kernel ridge regression

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Zhaohui Steve Qin,
Emory University, United States
Edwin Wang,
University of Calgary, Canada

****Correspondence:***
Li-Qian Zhou
zhoulq11@163.com
Xing Chen
xingchen@cumt.edu.cn
Xue Piao
px@xzhmu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *20 November 2019*
Accepted: *17 January 2020*
Published: *06 February 2020*

Citation:
Peng L-H, Zhou L-Q, Chen X and Piao X (2020) A Computational Study of Potential miRNA-Disease Association Inference Based on Ensemble Learning and Kernel Ridge Regression. Front. Bioeng. Biotechnol. 8:40. doi: 10.3389/fbioe.2020.00040

INTRODUCTION

MicroRNAs (miRNAs), known as the member of short non-coding RNA family, are found in eukaryotic organisms including viruses, plants and animals. They negatively regulate the expression of messenger RNA (mRNA) and the protein translation of their target genes (Bartel, 2004). In addition, miRNAs could also play a role of positive regulators demonstrated in some previous

studies (Jopling et al., 2005; Vasudevan et al., 2007). Under normal physiological conditions, miRNAs function in feedback mechanisms by safeguarding key biological processes including cell proliferation, differentiation and apoptosis (Bruce et al., 2015; Reddy, 2015). Many researchers have studied and validated the dysregulation of miRNA expression in various disease conditions (Esquelakerscher and Slack, 2006; Latronico et al., 2007; Lynam-Lennon et al., 2009; Meola et al., 2009; Wilczynska and Bushell, 2015). For example, Jeong et al. (2011) clarified that let-7a was under-expressed in the blood, cells and tissues of non-small cell lung cancer (NSCLC) patients compared to normal controls; and that the possibility of using let-7a as a serologic marker for lung cancer detection needed further study. Liang et al. (2012) reported that 66 miRNAs were differentially expressed in denatured dermis compared with those in normal skin; and the most significantly up-regulated miRNA was miR-663, while miR-203 was the most significantly down-regulated one. They further pointed out that identifying different miRNA expressions could enhance the understanding the mechanisms behind the functional recovery of the denatured dermis. Besides, miR-23/27/24 cluster has been shown by experiments to be involved in angiogenesis and endothelial apoptosis in cardiac ischemia and retinal vascular development (Bang et al., 2012). Hence, it is necessary and urgent to discover more miRNA-disease associations, contributing the prevention, diagnosis and treatment of complex human diseases. Nevertheless, it costs much time and money to discover true disease-related miRNAs from a mass of candidates by traditional biological experiments (Calin and Croce, 2006). Nowadays, many computational models of predicting miRNA-disease associations were developed based on some biological datasets, which could be used as an important complement to biological experiments (Chen et al., 2015).

Considering the hypothesis that functionally similar miRNAs tend to be related to similar diseases (Wang et al., 2010; Chen et al., 2019), some scoring function-based methods have been established to reveal new miRNA-disease associations. For example, Jiang et al. (2010) designed an initial computational model to infer potential disease-associated miRNAs. The approach integrated miRNA functional similarity network, the disease phenotype similarity network and the known miRNA-disease association network to prioritize the entire human miRNAome for the investigated disease with a cumulative hypergeometric distribution. But, the model failed to achieve excellent results because only neighbor information of miRNAs was used in the model. Moreover, Xuan et al. (2013) presented a prediction method of Human Disease-MiRNA association Prediction (HDMP) based on weighted k most similar neighbors of unlabeled miRNAs that have no known associations with disease d. However, HDMP could not predict related miRNAs for new diseases having no known association information. Additionally, the model also only used local network similarity. Furthermore, Chen et al. (2016b) proposed a new computational approach of Within and Between Score for MiRNA-Disease Association prediction (WBSMDA), which integrated known miRNA-disease associations, the miRNA functional similarity, the disease semantic similarity and the Gaussian interaction profile (GIP) kernel similarity for diseases and miRNAs. The authors first defined Within-Score and Between-Score in the side of miRNAs and diseases, respectively. Then the association score of the investigated miRNA-disease pair could be obtained by combining the corresponding Within-Score and Between-Score. However, WBSMDA did not exhibit outstanding performance because it was difficult to integrate Within-Score and Between-Score in reasonable way. Pasquier and Gardes (2016) introduced the MiRAI model that concatenated multiple miRNA-related association networks. Then dimensionality reduction technique was conducted for the combined network with Singular Value Decomposition (SVD). The final miRNA-disease association scores were attained by calculating cosine similarity between miRNA vectors in the miRNA space and disease vectors in the disease space.

Moreover, some network-based models were put forward. For example, Chen et al. (2012) proposed the model of Random Walk with Restart for MiRNA-Disease Association prediction (RWRMDA). The model was the first to adopt global network similarity measures and carried out random walk on the miRNA functional similarity network. However, the model also had the important limitation that it was not applicable to new diseases having no known association information. Another model named MIDP was presented by Xuan et al. (2015) for miRNA-disease association prediction. The model also performed random walk on the miRNA functional similarity network. To predict miRNA-disease associations for new diseases having no association information, the miRNA-disease bilayer network was built and MIDP could implement walk on this network. Furthermore, Chen et al. (2016c) presented the computational method of Heterogeneous Graph Inference for MiRNA-Disease Association prediction (HGIMDA), the inputs of which was same as WBSMDA. HGIMDA implemented an iterative process on the constructed heterogeneous graph to predict potential miRNA-disease associations. The model's performance was better than many previous models. In order to obtain better performance, Chen et al. (2018f) constructed Matrix Decomposition and Heterogeneous Graph Inference (MDHGI) to infer disease-related miRNAs. Before implementing heterogeneous graph inference similar to HGIMDA, the authors employed matrix factorization for miRNA-disease adjacent matrix to remove redundant information. Gu et al. (Gu et al., 2016) proposed another method named Network Consistency Projection for miRNA-Disease Associations (NCPMDA). Firstly, the authors constructed similarity network for miRNAs and diseases, respectively, by integrating multiple heterogeneous biological data. Secondly, the authors performed network consistency projection from the miRNA (disease) similarity network to the miRNA-disease association network, respectively. Lastly, scores of both network consistency projections were combined as the final miRNA-disease association score. Yu et al. (2017) introduced a prediction method named MaxFlow for miRNA-disease association prediction. In this method, miRNAome-phenome network was constructed by combining multiple heterogeneous network. For an investigated disease, the authors adopted push-relabel maximum flow algorithm to compute the maximum information flow from the source node over all links to sink node, and used the flow quantity leaving a miRNA node

as the association score between the investigated disease and miRNA. In addition, You et al. (2017) developed the model of Path-Based MiRNA-Disease Association prediction (PBMDA). This model constructed a heterogeneous graph, adopted depth-first search algorithm to traverse all paths between a miRNA node and a disease node, and used the product of all the edges' weights as path score in each path. The miRNA-disease association score could be obtained by summing all path scores between the miRNA node and disease node. In this model, distance-decay function was used for further weakening the score contribution of longer path. In addition, Chen et al. (2018e) presented the method of Bipartite Network Projection for MiRNA–Disease Association prediction (BNPMDA), which built bias ratings for miRNAs (diseases) by using agglomerative hierarchical clustering and improved traditional bipartite network recommendation. Although BNPMDA obtained better prediction accuracy than many previous models, it is not applicable for new diseases without known related miRNAs.

In fact, many previous studies were carried out with the addition of other association networks. For instance, Shi et al. (2013) developed a method to exploit associations between miRNAs and diseases by implementing a random walk analysis that focused on the functional links between the miRNA targets in the protein-protein interaction (PPI) network and the disease genes. Because of involving the PPI network, the model's prediction performance was improved and better than that of many previous models. In addition, Mørk et al. (2014) proposed a model named miRPD that integrated the protein-disease relations and the miRNA-protein interactions, and could effectively predict new associations between miRNAs and diseases.

With the development of machine learning algorithms, many researchers have begun to use this technology to solve various biological problems, such as prediction of drug-target interactions (Chen et al., 2016d), synergistic drug combinations (Chen et al., 2016a), disease related long non-coding RNAs (Chen et al., 2017c), miRNA-small molecule associations (Chen et al., 2018b), genome-wide features (Xu et al., 2018) and functional impact of variants (Milanese et al., 2019; Rojano et al., 2019; Xu et al., 2019). Of course, some machine learning models were developed for predicting potential associations between miRNAs and diseases (Chen et al., 2018a,c,g; Wang et al., 2019). For example, the model of Restricted Boltzmann Machine for Multiple types of MiRNA-Disease Association prediction (RBMMMDA) was proposed by Chen et al. (2015). It was worth noting that RBMMMDA could reveal association types for predicted miRNA-disease associations, which was different from other prediction models. However, RBMMMDA only took the advantage of the information of known multiple types of miRNA-disease associations, which hindered it from achieving an excellent performance. Xu et al. (2011) introduced a model based on a heterogeneous MiRNA-Target Dysregulated Network (MTDN). Features were extracted based on the network and a support vector machine (SVM) classifier was constructed to differentiate positive miRNA–disease associations from negative associations. However, the performance of this method was significantly influenced by an inaccurate selection of negative samples. In addition, another computational model of Regularized Least Squares for MiRNA-Disease Association prediction (RLSMDA) was proposed by Chen and Yan (2014). RLSMDA could implemented prediction for new miRNAs and new diseases without known association information. Moreover, negative samples were not needed in the model because RLSMDA was based on a semi-supervised learning-based model. However, the selection of parameter values limited the performance of RLSMDA. For a further improvement, Chen et al. (2017a) developed a computational model based on Super-Disease and MiRNA for potential MiRNA–Disease Association prediction (SDMMDA). In order to obtain more accurate miRNA (disease) similarity measures, the concepts of "super-miRNA" and "super-disease" were introduced into the model. Furthermore, Chen et al. (2017b) constructed a computational method of Ranking-based K-Nearest-Neighbors (KNN) for MiRNA-Disease Association prediction (RKNNMDA). The KNN of miRNA and diseases were obtained from their similarity scores. Then these KNN were reranked with a Support Vector Machine (SVM) ranking model, and finally, voting was weighted and final ranking of all possible miRNA-disease pairs was obtained. However, RKNNMDA did not show excellent prediction performance with an AUC (Area Under the ROC Curve) of 0.8221 in leave-one-out cross validation (LOOCV). Li et al. (2017) proposed the method of Matrix Completion for MiRNA-Disease Association prediction (MCMDA). In the model, the adjacency matrix of known miRNA-disease associations could be updated based on matrix completion technology, without requiring negative associations as needed by several previous models. Furthermore, Chen et al. (2018d) proposed another matrix completion-based method named Inductive Matrix Completion for MiRNA-Disease Association prediction (IMCMDA), which utilized miRNA (disease) similarity as features to train the model and complete the missing miRNA-disease associations. Furthermore, Chen and Huang (2017) developed the novel prediction method of Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction (LRSSLMDA). First, the model extracted feature profiles for miRNAs/diseases and formed graph Laplacian matrices. Second, a common subspace for the miRNA/disease feature profiles, a L_1-norm constraint and Laplacian regularization terms were joint to construct the objective function from miRNA and disease perspective, respectively. Third, the projection matrices in objective functions were iteratively updated and we obtained the final project matrix. Fourth, the association score between the miRNA and disease was computed using final projection matrix and feature profiles from miRNA and disease perspective respectively, and then the average of these two scores was the final prediction result. In addition, Zhao et al. (2019) further proposed the model named Adaptive Boosting for MiRNA-Disease Association prediction (ABMDA). In order to balance positive samples and negative samples, all unknown samples were divided into k clusters with k-means clustering and the same amount of negative samples were randomly selected from each cluster, and the number of total negative samples was almost equal to the positive. Then the authors integrated multiple weak classifiers (decision

trees) to build a strong classifier based on corresponding weights for prediction.

As mentioned above, there were various limitations on previous prediction methods. Developing new and effective computational methods for potential miRNA-disease association prediction is in urgent need. Some computational methods have been proposed based on the assumption that functionally similar miRNAs tend to relate to similar diseases (Wang et al., 2010; Chen et al., 2019). Therefore, we considered using miRNA functional (disease semantic) similarity as miRNA (disease) features to develop a machine learning-based method for miRNA-disease association prediction. In addition, given that miRNA functional and disease semantic similarity was not complete, GIP kernel similarity for miRNAs and diseases could be utilized to supplement similarity information. Therefore, we obtained integrated miRNA (disease) similarity features and introduced a framework based on ensemble learning and feature dimensionality reduction to construct prediction model. In each base learning process, a feature subspace was firstly built by randomly choosing a set of integrated similarity features. Secondly, a dimensionality reduction method called Truncated Singular Value Decomposition (TSVD) was used to reduce the number of features in the feature subspace. Finally, we used Kernel Ridge Regression (KRR) to construct two classifiers in the miRNA space and the disease space, respectively, and they were integrated as the base classifier. The above base learning process was conducted repeatedly to yield many base classifiers based on random selection of features. The average of all the association scores from base classifiers was computed to obtain the final prediction results. This new model was named as Ensemble of Kernel Ridge Regression based on MiRNA-Disease Association prediction (EKRRMDA). In our work, EKRRMDA showed sound performance in cross validation and case studies. The AUCs were 0.9314 and 0.8618 in global and local LOOCV, respectively, and in 5-fold cross validation, the average and standard deviation of AUCs was 0.9275 $\pm$ 0.0008. Furthermore, we implemented three types of case studies: (1) using known miRNA-disease associations in the HMDD V2.0 database, (2) simulating new diseases that have no known association information by removing known associations for the investigated disease in the HMDD V2.0 database, and (3) using known miRNA-disease associations in HMDD V1.0 database to test model's prediction performance in different datasets. The results showed that most miRNAs in top 50 predicted list were confirmed by experimental literature in case studies, which indicated that reliable prediction performance for the model.

MATERIALS AND METHODS

Human miRNA-Disease Associations

In our study, the dataset of human miRNA-disease associations came from HMDD V2.0 database (Li et al., 2014), covering 5430 known miRNA-disease associations between 495 miRNAs and 383 diseases. An adjacency matrix $A \in R^{m \times n}$ (Variables m and n represent the number of miRNAs and diseases, respectively) was used to describe all information of miRNA-disease associations. If miRNA m_i was associated with disease d_j, then $A(m_i, d_j)$ was equal to 1, and 0 otherwise.

miRNA Functional Similarity

Under the assumption that functionally similar miRNAs tend to be relate to semantically similar diseases, the method for calculating the miRNA functional similarity was proposed by Wang et al. (2010). MiRNA functional similarity could be obtained from http://www.cuilab.cn/files/images/cuilab/misim.zip and matrix *FS* was constructed to represent it.

Disease Semantic Similarity Model 1

Disease semantic similarity were computed according to the methodology adopted in the literature (Xuan et al., 2013). At first, we obtained the relationship among various diseases from the Mesh database[1] (Lipscomb, 2000; Wang et al., 2010). Then, we could use a graph $DAG(d) = [d, T(d), E(d)]$ to describe disease d. Here, $T(d)$ represented node set of all ancestor nodes of d and d itself, and $E(d)$ was the corresponding direct edges set. Each disease t in $DAG(d)$ has the contribution to the semantic value of disease d and we calculated the contribution as follows:

$$\begin{cases} D1_d(t) = 1 & \text{if } t = d \\ D1_d(t) = \max\{\Delta^* D1_d(t') \,|\, t' \in \text{children of } d\} & \text{if } t \neq d \end{cases} \quad (1)$$

The semantic value of disease d could be defined as follows:

$$DV1(d) = \sum_{t \in T(d)} D1_d(t) \quad (2)$$

where Δ was the semantic contribution decay factor. The above formula shows that diseases in different layers of $DAG(d)$ had different contributions to the semantic value of disease d. For diseases that locate in different layers, their contributions to the semantic value of disease d decreased as distance between these diseases and disease d increased. Specially, it is easy to understand that we defined the contribution of disease d to semantic value of itself as 1. Based on the assumption that two diseases sharing a larger part of their DAGs have a larger similarity score, the semantic similarity score between disease d_i and d_j was defined as follows:

$$SS1(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} \left(D1_{d_i}(t) + D1_{d_j}(t)\right)}{DV1(d_i) + DV1(d_j)} \quad (3)$$

where $SS1$ was disease semantic similarity matrix.

Disease Semantic Similarity Model 2

The point that diseases in the same layers of $DAG(d)$ have the same contribution to semantic value of disease d was adopted in disease semantic similarity model 1, however, it was not always reasonable. According to the literature (Xuan et al., 2013), another method of measuring disease semantic similarity was adopted. For example, if two diseases, t_1 and t_2, were located in the same layer of $DAG(d)$ and disease t_1 appeared in less DAGs

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

than t_2, disease t_1 could be considered as a more specific disease and its contribution to semantic value of disease d should be higher than disease t_2. So we defined the contribution of disease t in $DAG(d)$ to the semantic value of disease d as follows:

$$D2_d(t) = -\log\left[\frac{\text{the number of } DAGs \text{ including } t}{\text{the number of diseases}}\right] \quad (4)$$

Similar to disease semantic similarity model 1, we could define the semantic similarity between disease d_i and d_j as follows:

$$SS2\left(d_i, d_j\right) = \frac{\sum_t \in T\left(d_i\right) \bigcap T\left(d_j\right)\left(D2_{d_i}(t) + D2_{d_j}(t)\right)}{DV2\left(d_i\right) + DV2\left(d_j\right)} \quad (5)$$

where $DV2(d_i)$ and $DV2(d_j)$ were semantic value of disease d_i and d_j in semantic similarity model 2, respectively.

Gaussian Interaction Profile Kernel Similarity

Considering that not all miRNAs has functional similarity and so do diseases, the GIP kernel similarity for diseases and miRNAs were calculated according to van Laarhoven et al. (2011). By observing association information between miRNA m_i and each disease, binary vector $IV(m_i)$ was defined to represent the interaction profiles of miRNA m_i. The GIP kernel similarity between miRNA m_i and m_j could be computed as follows:

$$GM\left(m_i, m_j\right) = \exp\left(-\beta_m \| IV\left(m_i\right) - IV\left(m_j\right) \|^2\right) \quad (6)$$

$$\beta_m = \beta'_m \Big/ \left(\frac{1}{m}\sum_{i=1}^{m} \| IV\left(m_i\right) \|^2\right) \quad (7)$$

where adjustment coefficient β_m for the kernel bandwidth and β'_m was the original bandwidth. Similarly, disease GIP kernel similarity between disease d_i and d_j was computed as follows:

$$GD\left(d_i, d_j\right) = \exp\left(-\beta_d \| IV\left(d_i\right) - IV\left(d_j\right) \|^2\right) \quad (8)$$

$$\beta_d = \beta'_d \Big/ \left(\frac{1}{n}\sum_{i=1}^{n} \| IV\left(d_i\right) \|^2\right) \quad (9)$$

Integrated Similarity for miRNAs and Diseases

According to the literature (Chen et al., 2016b), the integrated disease (miRNA) similarity was attained by combining the disease semantic (miRNA functional) similarity with GIP kernel similarity. Taking disease as an example, if there was semantic similarity between disease d_i and d_j, then their integrated similarity was the mean of $SS1(d_i, d_j)$ and $SS2(d_i, d_j)$, otherwise we used GIP kernel similarity as the integrated disease similarity. The final integrated similarity between disease d_i and d_j, was computed as follows:

$$SD\left(d_i, d_j\right) = \begin{cases} \frac{SS1\left(d_i, d_j\right) + SS2\left(d_i, d_j\right)}{2} & d_i \text{ and } d_j \text{ has semantic} \\ GD\left(d_i, d_j\right) & \text{similarityotherwise} \end{cases} \quad (10)$$

where SD represented integrated disease similarity matrix. For miRNAs, we defined the integrated miRNA integrated similarity between miRNA m_i and m_j as follows in the same way:

$$SM\left(m_i, m_j\right) = \begin{cases} FS\left(m_i, m_j\right) & m_i \text{ and } m_j \text{ has functional} \\ GM\left(m_i, m_j\right) & \text{similarityotherwise} \end{cases} \quad (11)$$

where SM was denoted as integrated miRNA similarity matrix.

EKRRMDA

Ensemble of Kernel Ridge Regression based MiRNA-Disease Association prediction was implemented by integrating known miRNA-disease association, miRNA functional similarity, disease semantic similarity and GIP kernel similarity for miRNAs and diseases. As formula (Wilczynska and Bushell, 2015) and (Jeong et al., 2011) showed, GIP kernel similarity was employed to supplement missing miRNA functional similarity and disease semantic similarity so that complete similarity information for miRNAs and diseases was obtained, respectively. Integrated similarity was used for miRNA and disease features that were the inputs to training model. Based on random selection of features, multiple base learnings were carried out to yield many base classifiers. Then average strategy was adopted to integrate these classifiers and get final prediction results [see **Figure 1**, motivated by important study from Ezzat et al. (2017)].

In each base learning, every row of similarity matrix SM (SD) is feature vector for the corresponding miRNA (disease). For example, $SM(m_i,*)$ ($SD(d_i,*)$) represents feature vector miRNA m_i (disease d_i), which reflects similarity information between miRNA m_i (disease d_i) and other each miRNA (disease). During base learning, a set of features were firstly randomly selected for miRNAs and diseases. Here, we used parameter r ($0 < r < 1$) to determine number of selected features, which denotes proportion of selected features among all the features ($r = 0.2$ in our work). Here m_s, ($m_s = \lfloor r \times m \rfloor$ represents the largest integrate that is not larger than $r \times m$) miRNA features and n_s ($n_s = \lfloor r \times n \rfloor$) disease features were randomly sampled for each miRNA and disease, respectively. $SM1 \in R^{m \times m_s}$ and $SD1 \in R^{n \times n_s}$ denotes feature matrix of miRNAs and diseases after random feature selection, respectively.

Secondly, feature dimensionality reduction for miRNAs and diseases was further implemented to eliminate noises, redundancy, or irrelevant information and also improve computation efficiency. For method of dimensionality reduction, we chose TSVD that was developed from standard SVD (Xu, 1998). It took a given matrix $SM1$ and decomposed the matrix into $U \in R^{m \times k_m}$, $S \in R^{k_m \times k_m}$ and $V \in R^{m_s \times k_m}$ such that $SM1 = USV^T$, where k_m is truncation parameter and S was a diagonal matrix containing the largest k_m singular value of $SM1$. In our method, $k_m = \lfloor 0.2 \times m_s \rfloor$ indicated that top 20% larger singular value of $SM1$ was saved and others were ignored. The reduced miRNA feature matrix could then be obtained as $SM2 = US$ which realized the compression of the column for the matrix $SM1$. Similarly, reduced disease feature matrix $SD2 \in R^{n \times k_d}$ could be obtained from $SD1$ ($k_d = \lfloor 0.2 \times m_d \rfloor$). Finally, k_m and

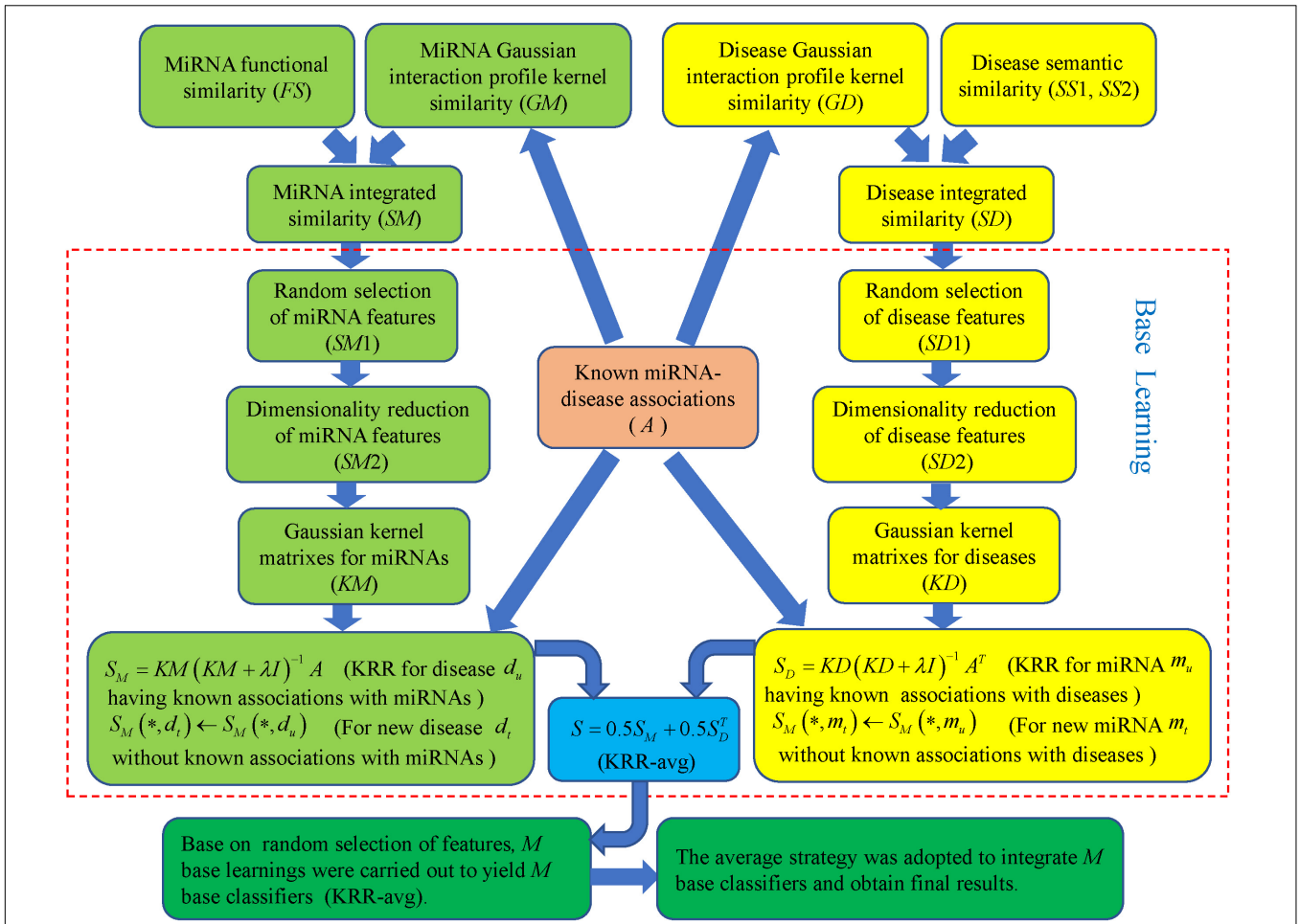


FIGURE 1 | Flowchart of EKRRMDA to predict the potential miRNA-disease associations based on the known associations in HMDD V2.0 database.

k_d is 19 and 15, which represented the number of miRNAs features and disease features after dimensionality reduction.

Thirdly, based on the known miRNA-disease associations and the dimensionally-reduced features, we used KRR to build two classifiers in the miRNA space and disease space, respectively. KRR was kernel-based classifier, where Least Squares was used in the kernel-induced space (Vovk, 2013). At first, we computed Gaussian kernel matrices for the miRNA and disease from *SM*2 and *SD*2, respectively. For example, for a pair of miRNA m_i and m_j, their Gaussian kernel similarity was computed as follows:

$$KM(m_i, m_j) = \exp\left(-\frac{\|SM2(m_i,*) - SM2(m_j,*)\|^2}{k_m}\right) \quad (12)$$

where $SM2(m_i,*)$ and $SM2(m_j,*)$ are reduced feature vectors for miRNA m_i and m_j, respectively. Analogically, Gaussian kernel similarity for the pair of disease d_i and d_j were computed as follows:

$$KD(d_i, d_j) = \exp\left(-\frac{\|SD2(d_i,*) - SD2(d_j,*)\|^2}{k_d}\right) \quad (13)$$

Then two KRR classifiers could be established with Gaussian kernel matrixes in different spaces, respectively. Taking miRNA space as an example, for each the investigated disease, the KRR was trained using the miRNA kernel matrix *KM* and adjacency matrix *A* to obtain association score between every miRNA and the investigated disease. Considering all diseases in the manner of matrix, the least-squares solution could be obtained as follows:

$$S_M = KM(KM + \lambda I)^{-1}A \quad (14)$$

where λ was a regularization parameter and we set its value as λ=1, referring to the previous work (van Laarhoven et al., 2011), and $I \in R^{m\times m}$ is the identity matrix. However, above formula could not work for new diseases that had no known associations with miRNAs. we inferred association scores between the new disease and miRNAs according to integrated disease similarities. For new disease d_t, association scores were recalculated by

$$S_M(*, d_t) = \frac{\sum_{d_u \in D_p} SD(d_t, d_u) \times S_M(*, d_u)}{\sum_{d_u \in D_p} SD(d_t, d_u)} \quad (15)$$

where D_p represented sets of diseases that have at least one known associations with miRNAs: $D_p = \left\{d_j \left| \sum_{i=1}^{m} A\left(i, j\right) \neq 0 \right.\right\}$. Similarly, in the disease space, S_D was calculated as follows:

$$S_D = KD\left(KD + \lambda I\right)^{-1} A^T \tag{16}$$

For new miRNA m_t, association scores were inferred as follows:

$$S_D\left(*, m_t\right) = \frac{\sum_{m_u \in M_p} SM\left(m_t, m_u\right) \times S_D\left(*, m_u\right)}{\sum_{m_u \in M_p} SM\left(m_t, m_u\right)} \tag{17}$$

The prediction scores in each base learning were obtained as follows:

$$S = 0.5S_M + 0.5S_D^T \tag{18}$$

In base learning, the base classifier that combined two classifiers from miRNA and disease spaces was named as KRR-avg. Above base learning containing three steps was implemented M times to yield M KRR-avg. The final miRNA-disease association scores could be obtained with average strategy. **Figure 2** shows pseudocode of EKRRMDA. For all predicted scores of miRNA-disease pairs with unknown associations, we ranked them and thought that pairs with higher scores were more likely to be associated.

RESULTS

Performance Evaluation

In order to assess performance of EKRRMDA, LOOCV and 5-fold cross validation were carried out based on the known miRNA-disease associations from the HMDD V2.0 database (Li et al., 2014), and prediction performance was measured in terms of AUC. Cross validation experiments and AUC measurement were usually used for evaluating methods of miRNA-disease association prediction in many previous important studies (Chen and Huang, 2017; Xiao et al., 2017; You et al., 2017; Chen et al., 2018e; Yang et al., 2018). Particularly, LOOCV was divided into

Algorithm: **EKRRMDA**

Input: SM, SD, A
Output: S
For i from 1 to M:
  $SM1_i$ = Randomly selected miRNA feature subset
  $SM2_i$ = Dimensionality reduction for $SM1_i$
  KM_i = Construct Gaussian kernel from $SM2_i$
  $S_{M(i)}$ = KRR(KM_i, A)
  If disease d_t has no known associations with miRNAs ($A(*, d_t) = 0$):
    $S_{M(i)}(*, d_t)$ = Infer the association score by combining scores of other diseases

  $SD1_i$ = Randomly selected disease feature subset
  $SD2_i$ = Dimensionality reduction for $SD1_i$
  KD_i = Construct Gaussian kernel from $SD2_i$
  $S_{D(i)}$ = KRR(KD_i, A)
  If miRNA m_t has no known associations with diseases ($A(m_t, *) = 0$):
    $S_{D(i)}(*, m_t)$ = Infer the association score by combining scores of other miRNAs

  S_i = Average($S_{M(i)}$, $S_{D(i)}$)
End
Return $S = \frac{1}{M}\sum_{i=1}^{M} S_i$

FIGURE 2 | The pseudocode of EKRRMDA.

global LOOCV and local LOOCV. In global LOOCV, we selected each known miRNA-disease association as test sample in turn and changed its label "1" to "0" in the adjacency matrix so that the association was hidden. Base on predicted association scores given by EKRRMDA, the test sample was ranked with all miRNA-disease pairs without association evidences. While in local LOOCV, for each given disease d, the test sample was one of the miRNAs associated with d, and the test sample was ranked with all the unassociated miRNAs for d. If the ranking of the test sample exceeded a pre-determined threshold, the model was considered to make a correct prediction for the sample. At different thresholds, the true positive rate (TPR) and the false positive rate (FPR) were calculated to plot the Receiver Operating Characteristic (ROC) curve, where TPR was used as the variate for the vertical axis and FPR for the horizontal axis in ROC. We evaluated the performance of EKRRMDA by calculating the area under ROC curve (AUC).

We compared EKRRMDA with several previous prediction methods in terms of AUC measurement. The overview of these methods was showed in **Supplementary Table S1**, which briefly provided the characteristic, input data as well as type (Scoring function-based, network-based or machine learning-based) of these models. **Figure 3** shows the results of performance comparisons in global and local LOOCV. As a result, EKRRMDA, HDMP, MaxFlow, NCPMDA, PBMDA, LRSSLMDA, ABMDA, BNPMDA, MDHGI, and IMCMDA obtained AUCs of 0.9314, 0.8366, 0.8624, 0.9073, 0.9169, 0.9178, 0.9170, 0.9028, 0.8945, and 0.8380 in global LOOCV, respectively. In local LOOCV, they achieved AUCs of 0.8618, 0.7702, 0.7774, 0.8584, 0.8341, 0.8418, 0.8220, 0.8380, 0.8240, and 0.8034, respectively. In addition, MIDP and MiRAI, obtained AUCs of 0.8196 and 0.6299 in local LOOCV, respectively. MIDP was a local approach that could not predict miRNAs for all diseases simultaneously so that global LOOCV could not evaluate performance of the model. In MiRAI, association scores of samples were closely related to the number of miRNAs associated with the diseases and for a disease with more known associated miRNAs, association scores for its candidate miRNAs tend to be higher. So it was not reasonable to implement prediction for all diseases simultaneously in global LOOCV. Additionally, MiRAI had a low AUC of 0.6299 which was worse than the AUC of 0.867 in Pasquier and Gardes (2016) literature, because MiRAI was a collaborative filtering-based model which was impacted by data sparsity problem. Our training data was HMDD V2.0 containing 383 diseases, where the average number of miRNAs related with a disease was 14, which was sparser than in the dataset in Pasquier and Gardes (2016) study containing 83 diseases with at least 20 known associated miRNAs. From the above comparisons, it is obvious that EKRRMDA has a more reliable performance.

Furthermore, we adopted 5-fold cross validation to evaluate performance of EKRRMDA. At First, we randomly partitioned all known miRNA-disease associations into five equal-sized parts. Each part was taken as the test set in turn, and the remaining four were used for model training. Then, samples in the test set were ranked against the miRNA-disease pairs without known association evidences. Finally, we obtained the rankings of all known associations, and TPR and FPR

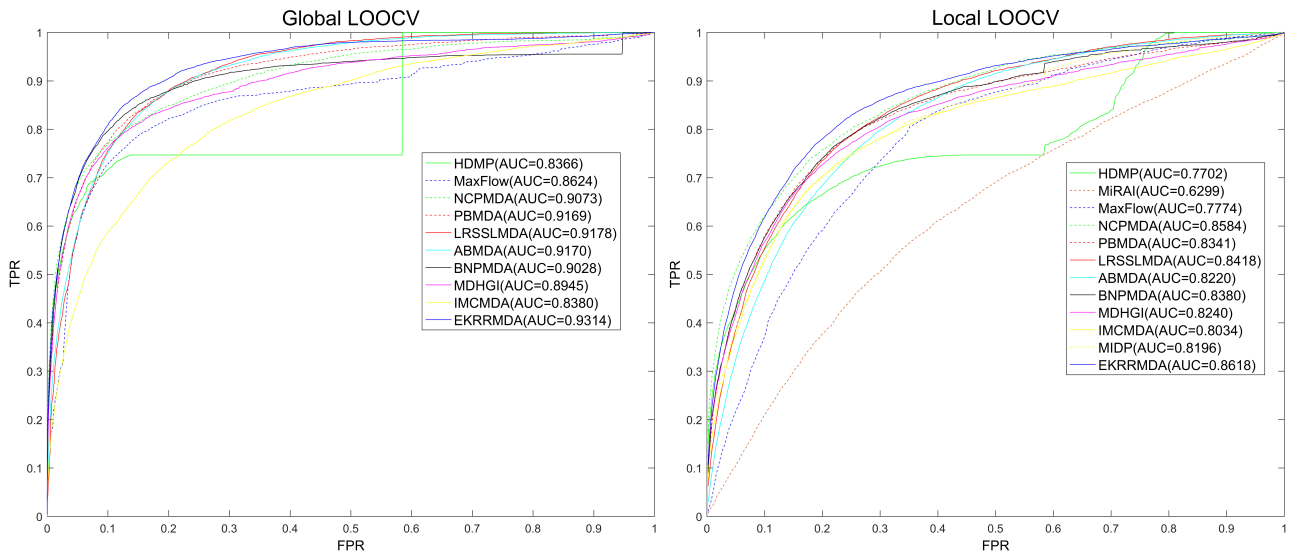


FIGURE 3 | Performance comparisons between EKRRMDA and other 11 prediction models (HDMP, MiRAI, MaxFlow, NCPMDA, PBMDA, LRSSLMDA, ABMDA, BNPMDA, MDHGI, IMCMDA, and MIDP) in terms of ROC curve and AUC based on local and global LOOCV, respectively. As a result, EKRRMDA obtained AUCs of 0.9314 and 0.8618 in the global and local LOOCV, which exceed all the previous classical models.

were calculated at various ranking thresholds to plot ROC and compute AUC. We repeated 5-fold cross-validations 100 times because of random division of known associations. As a result, EKRRMDA, PBMDA, NCPMDA, MaxFlow, HDMP, LRSSLMDA, ABMDA, BNPMDA, MDHGI, and IMCMDA obtained AUCs of 0.9275 ± 0.0008, 0.9127 ± 0.0007, 0.8763 ± 0.0008, 0.8579 ± 0.001, 0.8342 ± 0.0010, 0.9181 ± 0.0004, 0.9023 ± 0.0016, 0.8980 ± 0.0013, 0.8794 ± 0.0021, and 0.8367 ± 0.0005, which further shows the superior performance of our model.

In addition to prediction accuracy, we implemented cumulative distribution function (CDF) for the ranks of predicting samples based on LOOCV results to evaluate the model's prediction ability, which referred to the (Natarajan and Dhillon, 2014) work on predicting gene-disease associations. **Figure 4** showed CDF for the ranks of miRNA-disease associations for different models based on global LOOCV. The vertical axis in the plots gives the probability that a hidden miRNA-disease association is recovered in the top-k predictions for various k values in the horizontal axis. EKRRMDA outperformed most competitive models under global LOOCV. In the **Figure 5**, CDF for the miRNA ranks for different models based on local LOOCV was shown. The vertical axis in the plots gives the probability that a hidden miRNA associated with the investigated disease is recovered in the top-k predictions for various k values in the horizontal axis. EKRRMDA outperformed most competitive models from top 1 to 100 predictions. Specially, the performance of EKRRMDA was weaker than HDMP from top 1 to 10 predictions and NCPMDA from top 1 to 44 predictions, but surpassed HDMP from top 11 to 100 predictions and NCPMDA from top 45 to 100 predictions. However, NCPMDA and HDMP are network-based methods which need reliable similarity measurement for miRNAs and diseases to construct network for prediction. Moreover, it is a significant limitation for HDMP that it could not implement prediction for new diseases having no known association information.

Model Analysis

In this paper, we constructed prediction model by utilizing random selection of features for ensemble learning, TSVD for feature dimensionality reduction, Gaussian kernel for KRR and average strategy for combining two prediction scores in miRNA and disease space. The section was used to evaluate the effect of these steps.

In our work, KRR-avg as base classifier was constructed by introducing Gaussian kernel which is one of the most popular choices for constructing a kernel from feature vectors. We also compared Gaussian kernel with other two kernel

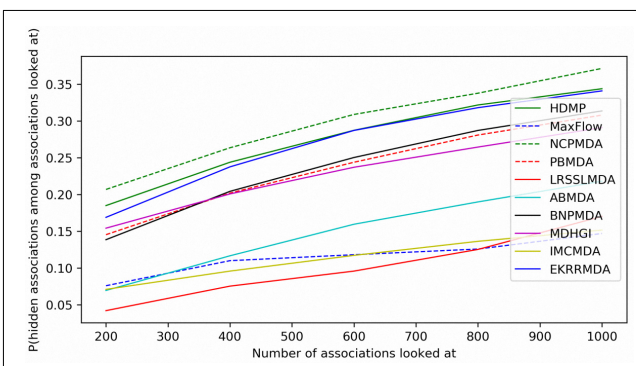


FIGURE 4 | Performance comparisons between EKRRMDA and other prediction models in terms of CDF of ranks based on global LOOCV.

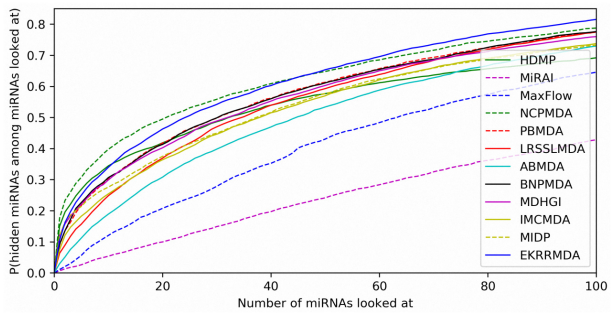


FIGURE 5 | Performance comparisons between EKRRMDA and other prediction models in terms of CDF of ranks based on local LOOCV.

functions used in KRR, Poly(1) and Poly(2) in the literature (Exterkate et al., 2016), both of which are polynomial kernel functions and their corresponding kernel function is $\kappa(x, y) = 1 + x'y$ and $\kappa(x, y) = (1 + x'y)^2$, respectively. The results of comparison were shown in **Table 1**, which indicated that Gaussian kernel performed better than polynomial kernel Poly(1) and Poly(2) in our model.

We constructed Gaussian kernel for KRR in miRNA and disease space, respectively, and averaged two predictions as final result. An alternative is to combine the kernels into a larger kernel that directly relates miRNA-disease pairs. We employed the Kronecker product kernel to realize it, referring to the literature (Chen and Li, 2017). The Kronecker product $KM \otimes KD$ of the miRNA and disease kernel is

$$K\left(\left(m_i, d_j\right), \left(m_k, d_l\right)\right) = KM\left(m_i, m_k\right) KD\left(d_j, d_l\right) \tag{19}$$

With this kernel, we can implement predictions for all pairs as follows:

$$vec\left(S^T\right) = K\left(K + \sigma I\right)^{-1} vec\left(A^T\right) \tag{20}$$

where $vec(\cdot)$ is the vectorization operator that stacks all columns of a matrix into a column vector. To solve the optimization problem more efficiently, some transformations were made and we can get the prediction in the form of Chen and Li (2017).

$$S = V_m Z^T V_d^T \tag{21}$$

where

$$vec(Z) = \left(\Lambda_m \otimes \Lambda_d\right)\left(\Lambda_m \otimes \Lambda_d + \lambda I\right)^{-1} vec\left(V_d^T A^T V_m\right),$$

and Λ_m, Λ_d, V_m and V_d come from the eigen decompositions of the two kernel matrices: $KM = V_m \Lambda_m V_m^T$ and $KD = V_d \Lambda_d V_d^T$. Under the computational framework of EKRRMDA, we used Kronecker product kernel to combine two Gaussian kernels in a single KRR as base classifier, named Ensemble of Kronecker Kernel Ridge Regression based MiRNA-Disease Association prediction (EKKRRMDA) for this method. The results of comparison between EKRRMDA and EKKRRMDA were shown in **Table 2**, which showed that the method of constructing separate KRR in miRNA and disease space, respectively, outperformed the method of combining two kernels for a single KRR.

Ensemble of Kernel Ridge Regression based MiRNA-Disease Association prediction trained multiple classifiers based on random selection of features, which inevitably brought some noise or redundancy. To address the issue, we implemented dimensionality reduction for the feature subset in each base learning. In addition, dimensionality reduction could reduce computation complexity for each base classifier. In order to evaluate the contribution of random selection of features and dimensionality reduction for EKRRMDA, we implemented three experiments including no random selection of features (i.e., all features were training for one classifier in miRNA and disease space, respectively), no dimensionality reduction and no both (no random selection of features and dimensionality reduction). The comparison results were shown in **Table 3**, which indicated that both random selection of features and dimensionality reduction could improve prediction performance and especially, random selection of features for ensemble learning contribute more.

Sensitivity Analysis

Here, we made sensitivity analysis for Gaussian kernel parameter, which was vital to the construction of classifiers in our model. The choice of Gaussian kernel parameter is always important but also tricky problem. Some methodologies for optimizing the kernel parameter have been proposed and used in Gaussian kernel methods. Grid search is often used to optimize the Gaussian kernel parameter, which choose the optimal parameter that show best test precision from candidate grid points. The problem of tuning kernel parameter is also done by minimizing an estimate of the generalization error or some other related performance

TABLE 1 | Comparison of global AUC and local AUC between different kernel functions used in KRR under LOOCV.

Kernel functions	Global AUC	Local AUC
Gaussian kernel	0.9314	0.8618
Poly(1)	0.9163	0.8436
Poly(2)	0.9023	0.8279

TABLE 2 | Comparison of global AUC and local AUC between EKRRMDA and EKKRRMDA under LOOCV.

Methods	Global AUC	Local AUC
EKRRMDA	0.9314	0.8618
EKKRRMDA	0.9093	0.8332

TABLE 3 | Comparison of global AUC and local AUC between EKRRMDA and variants of EKRRMDA under LOOCV.

Methods	Global AUC	Local AUC
EKRRMDA	0.9314	0.8618
No dimensionality reduction	0.9211	0.8568
No random selection of features	0.9015	0.8467
No dimensionality reduction and random selection of features	0.8951	0.8436

measure (Chapelle et al., 2002; Duan et al., 2003). Moreover, the optimization criterion based on kernel target alignment is a widely used method for choice of Gaussian kernel parameter (Cristianini et al., 2006; Fauvel, 2012).

We provided sensitivity analysis for Gaussian kernel parameter, which was implemented to investigate the variation of model's test precision for different parameter values. The results of sensitivity analysis were measured with global AUC and local AUC under the framework of LOOCV, which represented global prediction ability (for all miRNA-disease candidates) and local prediction ability (for miRNA candidates to the investigated disease), respectively. In order to better analyze effect of Gaussian kernel parameter, we adopted other form of Gaussian kernel: $\kappa(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ ($\sigma$is the bandwidth of Gaussian kernel), which was equivalent to $\kappa(x, y) = \exp(-\frac{\|x-y\|^2}{\gamma})$ used in our model, and made sensitivity analysis for parameter σ. **Figure 6** showed that global AUC and local AUC had the same trend and they reached maximum value when σ was about 2.0, and then decreased at a slower rate when increased. As mentioned above, k_m and k_d represented the number of miRNA features and disease features after dimensionality reduction. For Gaussian kernel in our model, we made $2\sigma^2$equal to k_m for miRNAs and k_d for diseases, so their corresponding σ is 3.1 and 2.7, respectively. From the sensitivity analysis results, we think it is sound to choose Gaussian kernel parameter by setting $2\sigma^2$ as number of features in our model.

Case Studies

To demonstrate the prediction accuracy of EKRRMDA, we implemented three different types of case studies on five diseases. The first type of case studies was carried out on three diseases, namely, Esophageal Neoplasms (EN), Kidney Neoplasms (KN) and Lymphoma. The known miRNA-disease associations in HMDD V2.0 were used as the training dataset for our model.

Esophageal Neoplasms, including squamous cell carcinoma (SCC) and adenocarcinoma (ADC), is one of the most common digestive cancers and ranks sixth among all cancers in mortality (Zhang, 2013). It is estimated there are about 15,690 people dying from EN among 16,940 newly diagnosed EN cases in 2017 in the United States (Siegel et al., 2017). Some miRNAs have been confirmed to be closely related to EN in previous studies. For instance, one of latest reports suggested that various miRNAs (miR-144, miR-451, miR-98, miR-10b, and miR-363) were involved in EN by regulating their target genes (Du and Zhang, 2017). In our case study of EN, there were total top 10 and 45 out of the top 50 potential EN-related miRNAs confirmed in dbDEMC and miR2Database (see **Table 4**).

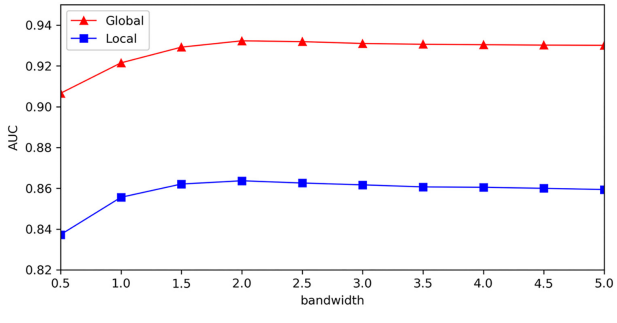


FIGURE 6 | Sensitivity analysis for bandwidth of Gaussian kernel.

Kidney Neoplasms, also known as renal cancer, accounts for about 3% of all adult neoplasms and its incidence rate is also increasing (Gonzalez-Satue et al., 2015). About 87% KN cases in adults were Renal cell carcinoma (RCC) that is the most common malignant epithelial tumor (Mytsyk et al., 2014). Some miRNA-KN associations have been revealed by experimental studies. For instance, a recent study indicated a clear correlation between higher expression of miR-21 and an aggravation in KN, which showed that miR-21 was useful in monitoring KN (Zaman et al., 2012). In addition, multiple miRNAs, including miR-215, miR-200c, miR-192, miR-194, and miR-141 were found downregulated in KN (Senanayake et al., 2012). After implementing EKRRMDA to predict potential KN-related miRNAs, we obtained that 8 miRNAs in top 10 predictions and 43 miRNAs in top 50 predictions were verified by dbDEMC (Yang et al., 2010) and miR2Disease (Jiang et al., 2009) (see **Supplementary Table S2**).

Lymphoma, a group of blood cell tumors, develops from lymphocytes and includes two main types, namely, Hodgkin

TABLE 4 | Prediction of the top 50 predicted miRNAs associated with EN.

miRNA	Evidence	miRNA	Evidence
hsa-mir-125b	dbdemc	hsa-mir-29b	dbdemc
hsa-mir-1	dbdemc	hsa-mir-9	dbdemc
hsa-mir-17	dbdemc	hsa-mir-195	dbdemc
hsa-mir-200b	dbdemc	hsa-let-7g	dbdemc
hsa-mir-221	dbdemc	hsa-mir-7	dbdemc
hsa-mir-16	dbdemc	hsa-mir-24	dbdemc
hsa-let-7e	dbdemc	hsa-mir-106a	dbdemc
hsa-mir-222	dbdemc	hsa-mir-106b	dbdemc
hsa-mir-133b	dbdemc	hsa-mir-199b	dbdemc
hsa-mir-30a	dbdemc	hsa-mir-122	unconfirmed
hsa-mir-18a	dbdemc	hsa-mir-124	dbdemc
hsa-mir-142	dbdemc	hsa-mir-30c	dbdemc
hsa-mir-29a	dbdemc	hsa-mir-429	dbdemc
hsa-mir-182	dbdemc	hsa-mir-132	dbdemc
hsa-mir-19b	dbdemc	hsa-mir-15b	dbdemc
hsa-mir-10b	dbdemc	hsa-mir-335	dbdemc
hsa-let-7f	unconfirmed	hsa-mir-224	dbdemc
hsa-let-7d	dbdemc	hsa-mir-127	dbdemc
hsa-mir-146b	dbdemc	hsa-mir-93	dbdemc
hsa-mir-181a	dbdemc	hsa-mir-497	dbdemc
hsa-mir-181b	dbdemc	hsa-mir-137	dbdemc
hsa-mir-125a	dbdemc	hsa-mir-18b	dbdemc
hsa-mir-107	dbdemc; miR2Disease	hsa-mir-204	unconfirmed
hsa-let-7i	dbdemc	hsa-mir-103a	unconfirmed
hsa-mir-218	unconfirmed	hsa-mir-23b	dbdemc

The first column records top 1–25 related miRNAs. The third column records the top 26–50 related miRNAs.

Lymphoma and non-Hodgkin Lymphoma (NHL) (Leich et al., 2011). According to research, about 90 percent of the Lymphoma cases are NHL (Alizadeh et al., 2000). There are plenty of miRNAs confirmed to be connected with Lymphoma. For example, miR-155 (contained in the BIC gene) is strongly up-regulated in Burkitt Lymphoma and several other types of Lymphomas (Metzler, 2004). And miRNA hsa-mir-19a exhibited an increased expression level compared with normal canine peripheral blood mononuclear cells (PBMC) and normal lymph nodes (LN) in canine B-cell Lymphomas (Uhl et al., 2011). After implementing EKRRMDA to predict potential lymphoma-related miRNAs we obtained that 8 miRNAs in top10 predictions and 43 miRNAs in top 50 miRNAs were verified by dbDEMC and miR2Disease (see **Supplementary Table S3**).

The second type of case study on Lung Neoplasms (LN) was implemented based on known associations in HMDD V2.0 database to illustrate the ability of EKRRMDA to predict miRNAs associated with the new disease. We hid all known miRNA-LN associations by changing their labels to "0" in adjacency matrix so that the LN could be treated as a new disease. We obtained a ranking list of miRNA-LN association scores and the top 50 potential miRNAs were shown in **Table 5**. Verification results showed that 49 miRNAs in top 50 predictions were confirmed by the dbDEMC, miR2Disease, and HMDD V2.0 databases. For example, a study in HMDD V2.0 indicated that expression of miRNA has-mir-21 (ranked first in the top 50 predictions), was more than two times in the squamous cell LN tissues compared with normal tissues (Gao et al., 2011).

Finally, to evaluate performance of EKRRMDA on different dataset, we implemented the third type of case study on Breast Neoplasms (BN) based on the known associations in HMDD V1.0 database that covers 1395 known miRNA-disease associations between 271 miRNAs and 137 diseases. Respectively, 10, 20, and 48 miRNAs in top 10, 20, 50 predictions were confirmed by dbDEMC, miR2Disease, and HMDD V2.0 (see **Table 6**). For example, has-let-7e, the miRNA ranked first in the top 50 predictions, was found to have close relationship with the development of BN in Chinese women (Jiang et al., 2013).

In addition, in order to further assess robustness of the model, we introduced random noise by randomly removing 20% known miRNA-disease associations in several case studies, i.e., we randomly changed 20% label "1" to "0" in adjacent matrix. To reduce the bias from random change, we repeated above experiment 10 times. We compared its average performance in top 10 and 50 predictions with our model in case studies. From the **Table 7**, we can observe that the number of confirmed miRNAs in top 10 and 50 predictions scarcely changed when random noise was introduced into case studies, which could show robustness of the model. To conclude, the case studies discussed above have demonstrated the outstanding prediction accuracy of EKRRMDA. In each case study, most of miRNAs in top 50 predictions were validated to be associate with the investigated

TABLE 5 | Prediction of the top 50 predicted miRNAs associated with LN as a new disease by removing all known associations containing LN in HMDD V2.0 database.

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbdemc;miR2Disease;HMDD V2.0	hsa-let-7e	miR2Disease;HMDD V2.0
hsa-mir-125b	miR2Disease;HMDD V2.0	hsa-mir-122	unconfirmed
hsa-mir-155	dbdemc;miR2Disease;HMDD V2.0	hsa-let-7f	miR2Disease;HMDD V2.0
hsa-mir-31	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-183	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-375	dbdemc;HMDD V2.0	hsa-mir-148a	dbdemc;HMDD V2.0
hsa-let-7a	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-222	dbdemc;HMDD V2.0
hsa-let-7b	miR2Disease;HMDD V2.0	hsa-mir-200a	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-34c	dbdemc;HMDD V2.0	hsa-mir-199a	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-7	miR2Disease;HMDD V2.0	hsa-mir-214	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-200b	dbdemc; miR2Disease;HMDD V2.0	hsa-mir-1	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-15b	dbdemc	hsa-mir-221	dbdemc;HMDD V2.0
hsa-mir-16	dbdemc;miR2Disease	hsa-mir-133a	dbdemc;HMDD V2.0
hsa-mir-34a	dbdemc;HMDD V2.0	hsa-mir-218	dbdemc;miR2Disease;HMDD V2.0
hsa-let-7g	dbdemc; miR2Disease; HMDD V2.0	hsa-mir-146a	dbdemc;miR2Disease;HMDD V2.0
hsa-let-7i	dbdemc;HMDD V2.0	hsa-mir-26a	dbdemc;miR2Disease;HMDD V2.0
hsa-let-7c	dbdemc; miR2Disease; HMDD V2.0	hsa-mir-205	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-196a	dbdemc;HMDD V2.0	hsa-mir-206	HMDD V2.0
hsa-mir-141	dbdemc;miR2Disease	hsa-mir-19a	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-34b	dbdemc;HMDD V2.0	hsa-mir-200c	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-100	dbdemc;HMDD V2.0	hsa-mir-133b	dbdemc;miR2Disease;HMDD V2.0
hsa-let-7d	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-29c	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-203	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-106b	dbdemc
hsa-mir-9	miR2Disease;HMDD V2.0	hsa-mir-93	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-145	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-15a	dbdemc
hsa-mir-101	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-499a	HMDD V2.0

The first column records top 1–25 related miRNAs. The third column records the top 26–50 related miRNAs.

TABLE 6 | Prediction of the top 50 predicted miRNAs associated with BN based on known associations in HMDD V1.0 database.

miRNA	Evidence	miRNA	Evidence
hsa-let-7e	dbdemc;HMDD V2.0	hsa-mir-195	dbdemc;miR2Disease;HMDD V2.0
hsa-let-7b	dbdemc;HMDD	hsa-mir-196b	dbdemc
hsa-let-7i	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-203	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-223	dbdemc;HMDD V2.0	hsa-mir-142	unconfirmed
hsa-let-7c	dbdemc;HMDD V2.0	hsa-mir-30e	unconfirmed
hsa-let-7g	dbdemc;HMDD V2.0	hsa-mir-32	dbdemc
hsa-mir-16	dbdemc;HMDD V2.0	hsa-mir-199b	dbdemc;HMDD V2.0
hsa-mir-126	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-99a	dbdemc
hsa-mir-92a	HMDD V2.0	hsa-mir-23b	dbdemc;HMDD V2.0
hsa-mir-92b	dbdemc	hsa-mir-30a	miR2Disease;HMDD V2.0
hsa-mir-106a	dbdemc	hsa-mir-335	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-101	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-532	dbdemc
hsa-mir-29c	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-107	dbdemc;HMDD V2.0
hsa-mir-191	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-224	dbdemc;HMDD V2.0
hsa-mir-373	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-98	dbdemc;miR2Disease
hsa-mir-99b	dbdemc	hsa-mir-27a	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-182	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-95	dbdemc
hsa-mir-181a	dbdemc;miR2Disease;HMDD V2.0	hsa-mir-128b	miR2Disease
hsa-mir-24	dbdemc;HMDD V2.0	hsa-mir-198	dbdemc
hsa-mir-100	dbdemc;HMDD V2.0	hsa-mir-31	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-15b	dbdemc	hsa-mir-491	dbdemc
hsa-mir-150	dbdemc	hsa-mir-193b	dbdemc;miR2Disease;HMDD V2.0
hsa-mir-18b	dbdemc;HMDD V2.0	hsa-mir-181d	dbdemc;miR2Disease
hsa-mir-372	dbdemc	hsa-mir-183	dbdemc;HMDD V2.0
hsa-mir-130a	dbdemc	hsa-mir-135a	dbdemc;HMDD V2.0

The first column records top 1–25 related miRNAs. The third column records the top 26–50 related miRNAs.

disease, and we would expect most of the remaining predictions to be verified in the future.

DISCUSSION

Considering that it costs much time and money to discover more potential miRNA-disease associations by traditional biological experiments, many computational models were developed to predict potential miRNA-disease associations, which could reduce cost and improve efficiency by preferentially verifying those promising associations. In this paper, we presented a machine-based prediction model named EKRRMDA. The novelty of the model was 2-fold. The first novelty was computational framework based on ensemble learning and feature dimensionality reduction. Since ensemble learning has been widely used to improve prediction accuracy, it was also worthwhile to design an ensemble learning model for prediction of potential miRNA-disease associations. In our prediction model, multiple base learnings were constructed based on random miRNA (disease) feature selection, each of which generated corresponding base classifier. However, our proposed ensemble learning model would increase computation complexity and inevitably brought some noise or redundancy, which motivated us employ feature dimensionality reduce techniques to address these issues. The second novelty was base classifier of the model. In this paper, we chose KRR as base classifier that always have been applied to drug-target association prediction and achieved excellent results (van Laarhoven and Marchiori, 2013), but to our knowledge, it has not been used for miRNA-disease association prediction. In model evaluation, Cross validations and case studies on EN, KN, Lymphoma, LN, and BN have shown the outstanding performance of EKRRMDA. We conclude that EKRRMDA would be a reliable computational model to predict disease-related miRNAs and could provide a substantial help in the prevention, diagnosis and treatment of human diseases.

TABLE 7 | The number of validated miRNAs among top 10 and top 50 predicted miRNAs in case studies between with all known miRNA-disease associations and with removing 20% associations.

Case study	Top 10 and all associations vs. removing 20% associations	Top 50 and all associations vs. removing 20% associations
The first type and EN	10 vs. 10	45 vs. 47
The first type and KN	8 vs. 8.7	43 vs. 41.6
The first type and Lymphoma	8 vs. 8.5	43 vs. 43.7
The second type and LN	10 vs. 10	49 vs. 48.9
The third type and BN	10 vs. 9.9	48 vs. 48

The prominent performance of the model could be attributed to the following points. Firstly, for the base classifier,

we established a bipartite local model by constructing two classifiers with KRR in two different spaces (the miRNA space and the disease space), which could solve the problem encountered by previous methods in figuring out a suitable way to merge miRNA and disease information. Secondly, multiple base classifiers were trained and integrated with ensemble learning strategy which generally bring more prediction accuracy than single classifier. Thirdly, we used the dimension reduction technique to eliminate noises, redundancy, or irrelevant information in the computation, which not only decreased the computational complexity, but also improved the prediction accuracy of the model.

However, the method had several limitations. First, the current known miRNA-disease associations were still inadequate for making much mre accuracy predictions, and with the increase of biological data in the future, the prediction performance of this method could be further improved. Second, similarity calculations for miRNAs and diseases had important impact on performance of model. We believe that integrating more biological information would contribute to obtaining more reliable similarity measures. Third, the choice of the parameter values remained to be further studied, such as parameter r used in random feature selection, truncation parameter k_m used in feature dimensionality reduction with TSVD. Especially, how to reasonably integrate results from different spaces would be a critical problem for future research.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

L-HP implemented the experiments, analyzed the result, and wrote the manuscript. L-QZ conceived the project, analyzed the result, and revised the manuscript. XC conceived the project, developed the prediction method, designed the experiments, analyzed the result, and revised the manuscript. XP analyzed the result and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

XC was supported by National Natural Science Foundation of China under Grant No. 61972399.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00040/full#supplementary-material

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511. doi: 10.1038/35000501

Bang, C., Fiedler, J., and Thum, T. (2012). Cardiovascular importance of the microRNA-23/27/24 family. *Microcirculation* 19, 208–214. doi: 10.1111/j.1549-8719.2011.00153.x

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.

Bruce, J. P., Hui, A. B. Y., Shi, W., Perezordonez, B., Weinreb, I., Xu, W., et al. (2015). Identification of a microRNA signature associated with risk of distant metastasis in nasopharyngeal carcinoma. *Oncotarget* 6, 4537–4550.

Calin, G. A., and Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nat. Rev. Cancer* 6, 857–866. doi: 10.1038/nrc1997

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing Multiple Parameters for Support Vector Machines. *Mach. Learn.* 46, 131–159. doi: 10.1023/a:1012450327387

Chen, H., and Li, J. (2017). "A Flexible and Robust Multi-Source Learning Algorithm for Drug Repositioning," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, (Boston, MA), 510–515.

Chen, X., Cheng, J. Y., and Yin, J. (2018a). Predicting microRNA-disease associations using bipartite local models and hubness-aware regression. *RNA Biol.* 15, 1192–1205. doi: 10.1080/15476286.2018.1517010

Chen, X., Guan, N. N., Sun, Y. Z., Li, J. Q., and Qu, J. (2018b). MicroRNA-small molecule association identification: from experimental results to computational models. *Brief. Bioinform.* bby098. doi: 10.1093/bib/bby098

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018c). EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.* 9, 3. doi: 10.1038/s41419-017-0003-x

Chen, X., Jiang, Z. C., Xie, D., Huang, D. S., Zhao, Q., Yan, G. Y., et al. (2017a). A novel computational model based on super-disease and miRNA for potential miRNA-disease association prediction. *Mol. Biosyst.* 13, 1202–1212. doi: 10.1039/c6mb00853d

Chen, X., Liu, M. X., and Yan, G. Y. (2012). RWRMDA: predicting novel human microRNA-disease associations. *Mol. Biosyst.* 8, 2792–2798. doi: 10.1039/c2mb25180a

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016a). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput. Biol.* 12:e1004975. doi: 10.1371/journal.pcbi.1004975

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018d). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265.

Chen, X., Wu, Q. F., and Yan, G. Y. (2017b). RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. *RNA Biol.* 14, 952–962. doi: 10.1080/15476286.2017.1312226

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018e). BNPMDA: Bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2019). MicroRNAs and complex diseases: from experimental results to computational models. *Briefings Bioinform.* 20, 515–539. doi: 10.1093/bib/bbx130

Chen, X., and Yan, G. Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.* 4:5501. doi: 10.1038/srep05501

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017c). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Briefings Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., Yan, C. C., Zhang, X., Li, Z., Deng, L., Zhang, Y., et al. (2015). RBMMMDA: predicting multiple types of disease-microRNA associations. *Sci. Rep.* 5:13877. doi: 10.1038/srep13877

Chen, X., Yan, C. C., Zhang, X., You, Z. H., Deng, L., Liu, Y., et al. (2016b). WBSMDA: within and between score for MiRNA-disease association prediction. *Sci. Rep.* 6:21106. doi: 10.1038/srep21106

Chen, X., Yan, C. C., Zhang, X., You, Z. H., Huang, Y. A., and Yan, G. Y. (2016c). HGIMDA: heterogeneous graph inference for miRNA-disease association prediction. *Oncotarget* 7, 65257–65269. doi: 10.18632/oncotarget.11251

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016d). Drug-target interaction prediction: databases, web servers and computational models. *Briefings Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Chen, X., Yin, J., Qu, J., and Huang, L. (2018f). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.*14:e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, X., Zhou, Z., and Zhao, Y. (2018g). ELLPMDA: Ensemble learning and link prediction for miRNA-disease association prediction. *RNA Biol.* 15, 807–818. doi: 10.1080/15476286.2018.1460016

Cristianini, N., Kandola, J., Elisseeff, A., and Shawe-Taylor, J. (2006). "On Kernel Target Alignment," in *Innovations in Machine Learning: Theory and Applications*. Berlin, Heidelberg, eds D. E. Holmes, and L. C. Jain, (Berlin: Springer), 205–256. doi: 10.1007/10985687_8

Du, J., and Zhang, L. (2017). Analysis of salivary microRNA expression profiles and identification of novel biomarkers in esophageal cancer. *Oncol. Lett.* 14, 1387–1394. doi: 10.3892/ol.2017.6328

Duan, K., Keerthi, S. S., and Poo, A. N. (2003). Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing* 51, 41–59. doi: 10.1016/S0925-2312(02)00601-X

Esquelakerscher, A., and Slack, F. J. (2006). Oncomirs - microRNAs with a role in cancer. *Nat. Rev. Cancer* 6:259. doi: 10.1038/nrc1840

Exterkate, P., Groenen, P. J. F., Heij, C., and van Dijk, D. (2016). Nonlinear forecasting with many predictors using kernel ridge regression. *Int. J. Forecasting* 32, 736–753. doi: 10.1016/j.ijforecast.2015.11.017

Ezzat, A., Wu, M., Li, X. L., and Kwoh, C. K. (2017). Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods* 129, 81–88. doi: 10.1016/j.ymeth.2017.05.016

Fauvel, M. (ed.) (2012). "Kernel matrix approximation for learning the kernel hyperparameters," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, (Munich).

Gao, W., Shen, H., Liu, L., Xu, J., Xu, J., and Shu, Y. (2011). MiR-21 overexpression in human primary squamous cell lung carcinoma is associated with poor patient prognosis. *J. Cancer Res. Clin. Oncol.* 137, 557–566. doi: 10.1007/s00432-010-0918-4

Gonzalez-Satue, C., Canas-Sole, L., Valverde-Vilamala, I., Tapia-Garcia, M., Pereira-Barrios, J. C., Areal-Calama, J., et al. (2015). Neoplasm in kidney graft. *Anal. Cause Tumour Treatm. Opt.* 35, 87–91. doi: 10.3265/Nefrologia.pre2014.Sep.12739

Gu, C., Liao, B., Li, X., and Li, K. (2016). Network consistency projection for human miRNA-disease associations inference. *Sci. Rep.* 6:36054. doi: 10.1038/srep36054

Jeong, H. C., Kim, E. K., Lee, J. H., Lee, J. M., Yoo, H. N., and Kim, J. K. (2011). Aberrant expression of let-7a miRNA in the blood of non-small cell lung cancer patients. *Mol. Med. Rep.* 4, 383–387. doi: 10.3892/mmr.2011.430

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37, D98–D104. doi: 10.1093/nar/gkn714

Jiang, Y., Qin, Z., Hu, Z., Guan, X., Wang, Y., He, Y., et al. (2013). Genetic variation in a hsa-let-7 binding site in RAD52 is associated with breast cancer susceptibility. *Carcinogenesis* 34, 689–693. doi: 10.1093/carcin/bgs373

Jiang, Q., Wang, G., Zhang, T., and Wang, Y. (2010). Predicting human microRNA-disease associations based on support vector machine. . *IEEE* 467–472. doi: 10.1109/BIBM.2010.5706611

Jopling, C. L., Yi, M., Lancaster, A. M., Lemon, S. M., and Sarnow, P. (2005). Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science*309, 1577–1581. doi: 10.1126/science.1113329

Latronico, M. V., Catalucci, D., and Condorelli, G. (2007). Emerging role of microRNAs in cardiovascular biology. *Circ. Res.* 101, 1225–1236. doi: 10.1161/circresaha.107.163147

Leich, E., Zamo, A., Horn, H., Haralambieva, E., Puppe, B., Gascoyne, R. D., et al. (2011). MicroRNA profiles of t(14;18)-negative follicular lymphoma support a late germinal center B-cell phenotype. *Blood* 118, 5550–5558. doi: 10.1182/blood-2011-06-361972

Li, J. Q., Rong, Z. H., Chen, X., Yan, G. Y., and You, Z. H. (2017). MCMDA: matrix completion for MiRNA-disease association prediction. *Oncotarget* 8, 21187–21199. doi: 10.18632/oncotarget.15061

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42, D1070. doi: 10.1093/nar/gkt1023

Liang, P., Lv, C., Jiang, B., Long, X., Zhang, P., Zhang, M., et al. (2012). MicroRNA profiling in denatured dermis of deep burn patients. *Burns* 38, 534–540. doi: 10.1016/j.burns.2011.10.014

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bull. Med. Library Assoc.* 88, 265–266.

Lynam—Lennon, N., Maher, S. G., and Reynolds, J. V. (2009). The roles of microRNA in cancer and apoptosis. *Biol. Rev.* 84, 55–71. doi: 10.1111/j.1469-185x.2008.00061.x

Meola, N., Gennarino, V. A., and Banfi, S. (2009). microRNAs and genetic diseases. *Pathogenetics* 2:7. doi: 10.1186/1755-8417-2-7

Metzler, M. (2004). High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma. *Genes Chromosomes Cancer* 39, 167–169. doi: 10.1002/gcc.10316

Milanese, J. S., Tibiche, C., Zou, J., Meng, Z., Nantel, A., Drouin, S., et al. (2019). Germline variants associated with leukocyte genes predict tumor recurrence in breast cancer patients. *NPJ Precis. Oncol.* 3:28. doi: 10.1038/s41698-019-0100-7

Mørk, S., Pletscherfrankild, S., Palleja, C. A., Gorodkin, J., and Jensen, L. J. (2014). Protein-driven inference of miRNA-disease associations. *Bioinformatics*30:392. doi: 10.1093/bioinformatics/btt677

Mytsyk, Y., Borys, Y., Komnatska, I., Dutka, I., and Shatynskamytsyk, I. (2014). Value of the diffusion-weighted MRI in the differential diagnostics of malignant and benign kidney neoplasms – our clinical experience. *Polish J. Radiol.* 79, 290–295. doi: 10.12659/PJR.890604

Natarajan, N., and Dhillon, I. S. (2014). Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*30, i60–i68. doi: 10.1093/bioinformatics/btu269

Pasquier, C., and Gardes, J. (2016). Prediction of miRNA-disease associations with a vector space model. *Sci. Rep.* 6:27036. doi: 10.1038/srep27036

Reddy, K. B. (2015). MicroRNA (miRNA) in cancer. *Cancer Cell Int.* 15, 1–6.

Rojano, E., Seoane, P., Ranea, J. A. G., and Perkins, J. R. (2019). Regulatory variants: from detection to predicting impact. *Briefings Bioinform.* 20, 1639–1654. doi: 10.1093/bib/bby039

Senanayake, U., Das, S., Vesely, P., Alzoughbi, W., Frohlich, L. F., Chowdhury, P., et al. (2012). miR-192, miR-194, miR-215, miR-200c and miR-141 are downregulated and their common target ACVR2B is strongly expressed in renal childhood neoplasms. *Carcinogenesis* 33, 1014–1021. doi: 10.1093/carcin/bgs126

Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., et al. (2013). Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. *Bmc Syst. Biol.* 7:101. doi: 10.1186/1752-0509-7-101

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387

Uhl, E., Krimer, P., Schliekelman, P., Tompkins, S. M., and Suter, S. (2011). Identification of altered MicroRNA expression in canine lymphoid cell lines and cases of B- and T-Cell lymphomas. *Genes Chromosomes Cancer* 50, 950–967. doi: 10.1002/gcc.20917

van Laarhoven, T., and Marchiori, E. (2013). Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One* 8:e66952. doi: 10.1371/journal.pone.0066952

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Vasudevan, S., Tong, Y., and Steitz, J. A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science* 318, 1931–1934. doi: 10.1126/science.1149460

Vovk, V. (2013). *Kernel Ridge Regression*. Berlin: Springer, 105–116.

Wang, C. C., Chen, X., Yin, J., and Qu, J. (2019). An integrated framework for the identification of potential miRNA-disease association based on novel negative samples extraction strategy. *RNA Biol.* 16, 257–269. doi: 10.1080/15476286.2019.1568820

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wilczynska, A., and Bushell, M. (2015). The complexity of miRNA-mediated repression. *Cell Death Differ.* 22, 22–33. doi: 10.1038/cdd.2014.112

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2017). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Xu, J., Li, C. X., Lv, J. Y., Li, Y. S., Xiao, Y., Shao, T. T., et al. (2011). Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. *Mol. Cancer Ther.* 10:1857. doi: 10.1158/1535-7163.MCT-11-0055

Xu, P. (1998). Truncated SVD methods for discrete linear ill-posed problems. *Geophys. J. Int.* 135, 505–514. doi: 10.1046/j.1365-246X.1998.00652.x

Xu, T., Zheng, X., Li, B., Jin, P., Qin, Z., and Wu, H. (2018). A comprehensive review of computational prediction of genome-wide features. *Briefings Bioinform.* doi: 10.1093/bib/bby110 [Epub ahead of print].

Xu, X., Li, J., Zou, J., Feng, X., Zhang, C., Zheng, R., et al. (2019). Association of germline variants in natural killer cells with tumor immune microenvironment subtypes, tumor-infiltrating lymphocytes, immunotherapy response, clinical outcomes, and cancer Risk. *JAMA Netw. Open* 2:e199292. doi: 10.1001/jamanetworkopen.2019.9292

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., et al. (2013). Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *Plos One* 8:e70204. doi: 10.1371/journal.pone.0070204

Xuan, P., Han, K., Guo, Y., Li, J., Li, X., Zhong, Y., et al. (2015). Prediction of potential disease-associated microRNAs based on random walk. *Bioinformatics* 31, 1805–1815. doi: 10.1093/bioinformatics/btv039

Yang, Y., Fu, X., Qu, W., Xiao, Y., and Shen, H.-B. (2018). MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA–disease association. *Bioinformatics* 34, 3547–3556. doi: 10.1093/bioinformatics/bty343

Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., et al. (2010). dbDEMC: a database of differentially expressed miRNAs in human cancers. *Bmc Genomics* 11(Suppl. 4):s5. doi: 10.1186/1471-2164-11-s4-s5

You, Z. H., Huang, Z. A., Zhu, Z., Yan, G. Y., Li, Z. W., Wen, Z., et al. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005455. doi: 10.1371/journal.pcbi.1005455

Yu, H., Chen, X., and Lu, L. (2017). Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm. *Sci. Rep.* 7:43792. doi: 10.1038/srep43792

Zaman, M. S., Shahryari, V., Deng, G., Thamminana, S., Saini, S., Majid, S., et al. (2012). Up-regulation of microRNA-21 correlates with lower kidney cancer survival. *PLoS One* 7:e31060. doi: 10.1371/journal.pone.0031060

Zhang, Y. (2013). Epidemiology of esophageal cancer. *World J. Gastroenterol.*19, 5598–5606. doi: 10.3748/wjg.v19.i34.5598

Zhao, Y., Chen, X., and Yin, J. (2019). Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics* 35, 4730–4738. doi: 10.1093/bioinformatics/btz297

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 © 2020 Peng, Zhou, Chen and Piao. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 07 February 2020
doi: 10.3389/fbioe.2020.00035

Dynamic Alternative Splicing During Mouse Preimplantation Embryo Development

Yongqiang Xing[1,2], Wuritu Yang[3], Guoqing Liu[1,2], Xiangjun Cui[1,2], Hu Meng[1,2], Hongyu Zhao[1,2], Xiujuan Zhao[1,2], Jun Li[1,2], Zhe Liu[1], Michael Q. Zhang[4] and Lu Cai[1,2]**

[1] School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou, China, [2] The Inner Mongolia Key Laboratory of Functional Genome Bioinformatics, Inner Mongolia University of Science and Technology, Baotou, China, [3] The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the Ministry of Education, Inner Mongolia University, Hohhot, China, [4] Department of Biological Sciences, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX, United States

The mechanism of alternative pre-mRNA splicing (AS) during preimplantation development is largely unknown. In order to capture the dynamic changes of AS occurring during embryogenesis, we carried out bioinformatics analysis based on scRNA-seq data over the time-course preimplantation development in mouse. We detected numerous previously-unreported differentially expressed genes at specific developmental stages and investigated the nature of AS at both minor and major zygotic genome activation (ZGA). The AS and differential AS atlas over preimplantation development were established. The differentially alternatively spliced genes (DASGs) are likely to be key splicing factors (SFs) during preimplantation development. We also demonstrated that there is a regulatory cascade of AS events in which some key SFs are regulated by differentially AS of their own gene transcripts. Moreover, 212 isoform switches (ISs) during preimplantation development were detected, which may be critical for decoding the mechanism of early embryogenesis. Importantly, we uncovered that zygotic AS activation (ZASA) is in conformity with ZGA and revealed that AS is coupled with transcription during preimplantation development. Our results may provide a deeper insight into the regulation of early embryogenesis.

Keywords: alternative splicing, gene expression, preimplantation development, zygotic gene activation, splicing factors

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Jiapei Yuan,
UT Southwestern Medical Center, United States
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Michael Q. Zhang
michael.zhang@utdallas.edu
Lu Cai
nmcailu@163.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *27 November 2019*
Accepted: *15 January 2020*
Published: *07 February 2020*

Citation:
Xing Y, Yang W, Liu G, Cui X, Meng H, Zhao H, Zhao X, Li J, Liu Z, Zhang MQ and Cai L (2020) Dynamic Alternative Splicing During Mouse Preimplantation Embryo Development. Front. Bioeng. Biotechnol. 8:35. doi: 10.3389/fbioe.2020.00035

INTRODUCTION

Decoding molecular mechanisms of totipotency and pluripotency is crucial to the understanding of reproductive biology and to regenerative medicine (Hamatani et al., 2004). Preimplantation process, which encompasses the period from fertilization to implantation, is a fundamental developmental stage that has been extensively studied in order to gain insight to totipotency and pluripotency (Yan et al., 2013; Petropoulos et al., 2016). With the development of single-cell RNA-seq (scRNA-seq) technology, the barrier of scarcity of preimplantation embryo materials has been overcome. The scRNA-seq is an unbiased and popular approach to investigate heterogeneous tissues and organs, especially for embryogenesis. To date, numerous scRNA-seq studies on mouse or human preimplantation embryos have identified a large number of genes and signaling

pathways involved in early stages of embryonic development (Hamatani et al., 2004, 2006; Yan et al., 2013; Petropoulos et al., 2016). However, the molecular regulatory mechanisms underlying preimplantation process remain incompletely understood, especially the effect of AS in this process.

AS is a ubiquitous and conserved regulatory mechanism of gene expression in which introns are removed and exons are joined in different combinations to create various alternative mRNA products (Zhang, 2002; Park et al., 2018). The distinct proteins produced from identical pre-mRNAs via AS may have different, even antagonistic functions (Park et al., 2018). AS greatly expands the diversity of transcriptome and proteome in higher eukaryotic organisms and plays an important role in numerous processes, such as cell differentiation, proliferation, apoptosis, organ development and the genesis of human disease, etc. (Kornblihtt et al., 2009; Kalsotra and Cooper, 2011; Singh and Cooper, 2012; Xiong et al., 2015; Scotti and Swanson, 2016). AS is also essential for mammalian early embryogenesis to generate a viable organism from a fertilized cell (Revil et al., 2010). Revil et al. (2010) studied splicing-sensitive exon microarray in embryonic 8–12 days mouse embryos and revealed that AS is frequent across early developmental stages and tissues. However, the detailed temporal and spatial patterns of AS during preimplantation development are poorly understood.

In mouse, pre- and early embryo development is a complex process that consists of sequential maturation events of the oocyte, fertilization (zygote) and embryo growth (2-cell, 4-cell, 8-cell, morula, and blastocyst) (Assou et al., 2011). Here, we utilized time-series scRNA-seq data consisting 21 single-cells from mouse seven consecutive stages of preimplantation development to dissect the dynamics of the gene expression and AS. A total of 4,952 genes were differentially expressed at the gene level (DEGs) in all the consecutive early developmental stages, of which 507 genes were also differentially alternatively spliced. The AS atlas was constructed for seven development stages and 1,170 differential AS events (DAS) in 836 genes were identified at the consecutive development stages. A regulatory cascade of AS that some splicing factors regulate AS by DEGs and DAS of their own gene transcript was found. A dataset of ISs during preimplantation development was established. Moreover, we uncovered that ZASA is in conformity with ZGA and revealed that AS is coupled with transcription during preimplantation in mouse. This study is expected to be helpful for elucidating the molecular and cellular mechanisms of preimplantation embryo development.

MATERIALS AND METHODS

Dataset

Fan et al. developed SUPeR-seq (single-cell universal poly(A)-independent RNA sequencing) method to sequence single cell complete transcriptome (poly(A)+ and poly(A)–) of mouse early embryos (Fan et al., 2015). We downloaded complete transcriptome data of 25 single cells generated from mouse occytes and preimplantation embryos. The embryos cover seven consecutive stages of preimplantation development: metaphase II oocyte, zygote, 2-cell, 4-cell, 8-cell, morula, and blastocyst. Then, we respectively dropped two poor-quality single-cells transcriptome data in occyte and zygote, RNA-seq data at every development stage of mouse preimplatation was composed of three single-cell sequenced on Illumina HiSeq 2,000 platform (**Table S1**). On average, every cell has 12.7 million in 101 bp paired-end reads.

FastQC v0.11.8 (Andrews, 2010) and Trimmomatic v.38 (Bolger et al., 2014) were used to perform QC (quality control) analysis for raw reads. FastQC analysis showed the adaptor was already cut before uploading GEO and quality of $3'$ end of reads is lower. We removed low quality reads (the average quality per base within 4-base wide window drops below 10, SLIDINGWINDOW:4:10). The reads containing poly(A)24/(T)24 sequences were trimmed off. The leading and trailing bases of a read were cut if quality is below 3 (LEADING:5, TRALING:5). All reads were outputted with read length of 91 bp (MINLEN:91, CROP:91). The average surviving rate and sequencing depth of paired-end reads after quality control is 79.1% and 10.0 million (**Table S1**).

Quantification of Transcript and Gene Expression

The gene annotation GTF file, nucleotide sequence FASTA file and transcript sequence FASTA file were downloaded from Gencode (vM10/GRCm38.p4). In this work, we only focused on coding gene. After filtering, the annotation GTF file composed of 22,021 coding genes was created.

The transcript qualification of different preimplantation development stages was carried out by combing Salmon v0.11.3 (Patro et al., 2017) and transcript sequence FASTA file. For indexing, because the read length is larger than 75 bp, we used the quasi mapping mode to build an auxiliary *k*-mer hash over *k*-mers of length 31 (-type quasi -k 31). Besides, the option to qualify duplicate transcripts ("-keepDuplicates") was turned on. For accurate quantification, the option to correct for the sequence specific bias ("-seqBias") was also turned on and all other parameters were on default settings. The TPM (Transcripts Per Kilobase Million) value of 86,623 transcripts corresponding all coding genes across all samples was calculated (**Table S2**).

To construct gene count matrix (22,021 × 21), TPM data of transcript generated by Salmon was processed using tximport version 1.10.1 R package (Soneson et al., 2015) with the default setting (**Table S3**).

The Identification of Differentially Expressed Genes

Gene expression analysis and cell type clustering were performed using Seurat v2.3.4 (Butler et al., 2018). Seurat is an R package designed for QC analysis, visualization, and exploration of single cell RNA-seq data. Seurat aims to enable users to identify and interpret sources of heterogeneity from single cell transcriptomic measurements, and to integrate diverse types of single cell data. By QC, only those genes that were expressed in at least 3 or more cells and cells that expressed more than 10,000 genes were retained. A 16,539 (genes) × 21 (samples)

Seurat object was created. After removing low-expressed genes, the "LogNormalize method" was used to normalize the gene expression. Next, the FindVariableGenes function was used to identify highly variable genes followed by scaling data (ScaleData) for downstream analysis. We clustered the cells using FindClusters function and visualized all cells by integrated tSNE. Finally, we used FindMarkers() function of Seurat to detect differentially expressed genes under every consecutive stages of preimplantation development. FindMarkers() function provides nine tests for differential expression which can be set with the test.use parameter. Here, test.use was set to DESeq2, which is based on a model using the negative binomial distribution (Love et al., 2014). The *avg_log$_e$ fold change* (*FC*) of gene abundances was calculated in each consecutive development stages. *P*-values were adjusted by the BH method for multiple testing correction (Benjamini and Hochberg, 1995). We selected adjust p value $\leq$ 0.05 and $\left|avg_\log_e FC\right| \leq 0.25$ as the threshold to judge the significance of differentially expressed genes.

The Identification and Quantification of AS Events

At present, there are many tools to detect and quantify AS events, such as SUPPA (Trincado et al., 2018), rMATs (Shen et al., 2014), MAJIQ (Vaquero-Garcia et al., 2016), etc. On the one hand, SUPPA is much faster than the other methods and achieves higher accuracy compared to other methods, especially at low sequencing depth and short read length (Trincado et al., 2018). On the other hand, this work only paid attention on AS events derived from pre-existing transcript annotations. Thus, AS analysis in this work was performed by SUPPA v2.3 (Trincado et al., 2018). SUPPA is a powerful and reliable tool to study splicing at the transcript isoform or at the local AS event level across multiple conditions. SUPPA was used to generate the AS events (e.g., A5SS, A3SS, SE, RI, MXE, AFE, ALE) from mouse annotation file. Then, AS event inclusion levels (*PSI*) from multiple developmental stages were quantified. Furthermore, SUPPA calculated the magnitude of splicing change (ΔPSI) and its significance across multiple development stages directly from TPM value of transcript involved in the event. For example, an exon skipping event across two development stages consists of an included transcript and a skipped transcript. Then, the included level *PSI* and splicing change ΔPSI can be defined as:

$$PSI = \frac{TPM_1}{TPM_1 + TPM_2}$$
$$\Delta PSI = \overline{PSI_1} - \overline{PSI_2}$$

where the TPM_1 and TPM_2 are the expression level of included transcript and skipped transcript, respectively. $\overline{PSI_1}$ and $\overline{PSI_2}$ are the mean of *PSI* of biological replicates for development stage 1 and development stage 2, respectively.

Criteria for judging DAS was that in contrast group (1) splicing change (ΔPSI) across two different developmental stages showed $\geq$ 0.1. (2) ΔPSI differs significantly with p value $\leq$ 0.05 (Calixto et al., 2018).

Identification of ISs

For the isoform switch analysis, we used the TSIS R package, which is a tool to detect significant transcript ISs in time-series data (Guo et al., 2017). ISs between any two consecutive development stages were identified using the default parameters in which (1) the probability of switch (i.e., the frequency of samples reversing their relative abundance at the switches) was set to >0.5; (2) the sum of the average differences of the two isoforms in both intervals before and after the switch point were set at ΔTPM >1; (3) the significance of the differences between the switched isoform abundances before and after the switch was set to $p < 0.05$; and (4) both intervals before and after switch must consist of at least 2 consecutive development stages to detect long lasting switches.

Gene Ontology and KEGG Enrichment Analysis

Gene Ontology and KEGG pathway enrichment analysis were performed using clusterProfiler package in R (http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) (Yu et al., 2012). The statistical significance threshold level for all GO enrichment and KEGG pathway analyses was *p.adjust* < 0.05.

Splicing Factor Analysis

A total of 446 mouse splicing factors were selected for analysis based on literature mining for previously described splicing functions (Han et al., 2013; Goldstein et al., 2017), and "RNA splicing" or "spliceosome"-associated Gene Ontology (GO) terms from MGI (Smith et al., 2018) (http://www.informatics.jax.org/marker) (**Table S4**).

RESULTS

The Global Outlook of DEGs and DAS During Preimplantation Development

To examine changes in gene expression and AS of mouse preimplantation embryos, we collected complete transcriptome data of 21 single cells from mouse occytes and preimplantation embryos (Fan et al., 2015). The embryos cover seven consecutive stages of preimplantation development: metaphase II oocyte, zygote, 2-cell, 4-cell, 8-cell, morula and blastocyst (see Materials and Methods). Every development stage includes three scRNA-seq replicates sequenced on Illumina HiSeq 2000 platform. On average, every cell has 12.7 million in 101 bp paired-end reads (**Table S1**). QC analysis of RNA-seq data was carried out using FastQC (Andrews, 2010) and Trimmomatic (Bolger et al., 2014). Consequently, the average clean paired-end reads in every cell is ∼10.0 million with length of 91 bp (**Table S1**).

In this study, we only focused on coding genes in mouse annotation GTF file. After filtering, 22,021 coding genes were considered in the downstream analysis. We first examined how many reads mapped to coding genes in every cell (**Figure 1A**). On average, every cell includes ∼8.3 million reads. Then, Salmon tool (Patro et al., 2017) and tximport R package (Soneson et al., 2015) were employed to quantify expression matrix of

transcripts and genes. Across all samples, we identified 18,272 protein-coding genes expressed in at least one sample. In total, 6,522 protein-coding genes were expressed in all samples. We selected the protein-coding genes that were expressed in at least 3 or more cells for downstream analysis. With this criterion, the gene count matrix was created, which includes 16,539 protein-coding genes along the rows and 21 samples along the columns. We observed that ∼82% protein-coding genes were expressed during preimplantation development. The mean of detected protein-coding genes is 12,496 across 21 cells and the number of detected protein-coding genes in every cell is higher than 10,000 (**Figure 1B**). The mean of detected protein-coding genes with TPM larger than 10 across 21 cells is 5,692 (**Table S1**). The number of protein-coding genes (1,3259) in the 2-cell stage is larger than other developmental stages. ZGA is the first major developmental event that occurs following fertilization (Schultz et al., 2018). After ZGA process, the genetic program governed by maternal transcripts/proteins should be switched to that dominated by transcripts/proteins from the newly formed zygotic genome (Kanka, 2003; Hamatani et al., 2004). During ZGA process of mouse embryos development, lots of zygotic genes are activated and maternal genes have not been degraded thoroughly. Given that 2-cell stage is major start of ZGA in mouse (Abe et al., 2018), the fact that the maximum number of protein-coding genes was observed in the 2-cell stage is reasonable.

The global gene expression profiles at different developmental stages should be distinguishable. Seurat is a widely used R package for scRNA-seq data analysis (Butler et al., 2018). Especially, Seurat was often used to identify cell identity. We applied Seurat to 16,539 (genes) × 21 (samples) count matrix in mouse embryos development. After removing low-expressed genes and normalizing the gene expression, 6,902 highly variable genes throughout each development stage were identified. In order to confirm the identity of every cell, we applied a graph-based clustering method on the most variable genes and identified seven clusters of cells as visualized by bi-dimensional t-distributed stochastic neighbor embedding (t-SNE) (**Figure 1C**). By comparing with the experimental source of cells, we observed that every cell can be clustered correctly as actual development stages. Thus, 21 cells selected as samples have a high reliability for downstream analysis.

Developmental process requires precise spatial-temporal regulation of gene expression and AS. Here, we used time-series scRNA-seq to examine the dynamics of DEGs at the gene level and DAS at transcript level. To analyze the time-series RNA-seq data at gene levels, we employed FindMarkers() function of Seurat package to detect DEGs between each two consecutive stages of the preimplantation development. If avg_log_eFC is positive values, it indicates that the gene is more expressed in the first group, and vice versa. Here, we used a more stringent criteria: a gene was regarded as differentially expressed if $|\text{avg_log}_e\text{FC}| \geq 0.25$ (≥1.3-FC) and adjusted p value ≤ 0.05. Under these criteria, a total of 4,952 genes were identified as differentially expressed throughout all successive preimplantation developmental stages (**Figure 1D**). Of these, 37.5% were consistently up-regulated, 35.5% were consistently down-regulated and 27.0% were up-regulated or down-regulated over different consecutive development stages. Moreover, we applied SUPPA tool on the transcript-level data generated by Salmon (Trincado et al., 2018) to identify genes that were DAS between consecutive preimplantation developmental stages. We recognized 836 DASGs, of which 507 are overlapped with DEGs and 329 are not. It indicates that 507 genes are simultaneously regulated at both transcriptional level and AS level, and 329 genes are only regulated by AS. As a typical example, a total of 2,233 DEGs and 250 DASGs between zygote and 2-cell were detected. Of DASGs, 92 are also DEGs. Furthermore, heatmap showing the expression levels of DEGs (**Figure S1**) and the inclusion levels of DAS events (**Figure S2**) across seven consecutive stages of preimplantation development suggested some DEGs and DAS events is stage-specific. So, the DEGs and DAS events were analyzed in detail.

Analysis of DEGs in Consecutive Developmental Stages of Preimplantation Embryo

ZGA is essential for replacing the degraded maternal transcripts with zygotic transcripts (Yan et al., 2013). In mouse embryos, major ZGA process reportedly occurs at the 2-cell and 4-cell stages (Abe et al., 2018). The greatest DEGs number between 2-cell and zygote compared with other consecutive stages during preimplantation development was detected (**Figure 1E**). It indicates that the transcriptome difference between these two stages is greatest. By functional enrichment analysis on DEGs, we confirmed some previous conclusions, such as the zygotic-specific transcription and translation machinery is established during ZGA (**Figures S3**, **S4** and **Table S5**) (Yan et al., 2013). Besides, some genes were also strongly enriched in splicing-associated processes, such as mRNA processing (gene number = 61, $p.adjust = 1.19 \times 10^{-12}$), RNA splicing (gene number = 50, $p.adjust = 3.44 \times 10^{-09}$), mRNA catabolic process (gene number = 38, $p.adjust = 2.62 \times 10^{-08}$), and alternative mRNA splicing (gene number = 10, $p.adjust = 0.047$), indicating that the biological process of mRNA splicing are activated in ZGA (**Table S5**). The significantly up-regulated gene *Dhx33* in 2-cell stage plays essential roles in mRNA translation, pre-mRNA splicing and ribosome biogenesis (Zhang et al., 2015) (**Figure 2A**). The pabpc1 protein that binds the poly(A) tail of mRNA involved in cytoplasmic regulatory processes of mRNA metabolism, such as pre-mRNA splicing. We found the expression level of gene *pabpc1* is significantly up-regulated in 2-cell (**Figure 2A**). This finding implied that AS may initiate in ZGA. It was known that mitochondrial metabolism contributes a major role in the supply of ATP during preimplantation embryo development (Wilding et al., 2009). The *Tomm20* gene is critical for synthesis of mitochondrial pre-proteins. The substantial amounts of ATP are consumed during ZGA. Thus, the expression level of *Tomm20* gene is elevated obviously during ZGA (**Figure 2A**). The most up-regulated gene in the 2-cell stage is *Tmem72* [Transmembrane protein 72-like, FC (2-cell/Zygote) = 403]. *Tmem72* encodes a transmembrane protein and the biological function of *Tmem72* is unknown. *Tmem72* is localized to the mitochondria in human clear cell renal cell

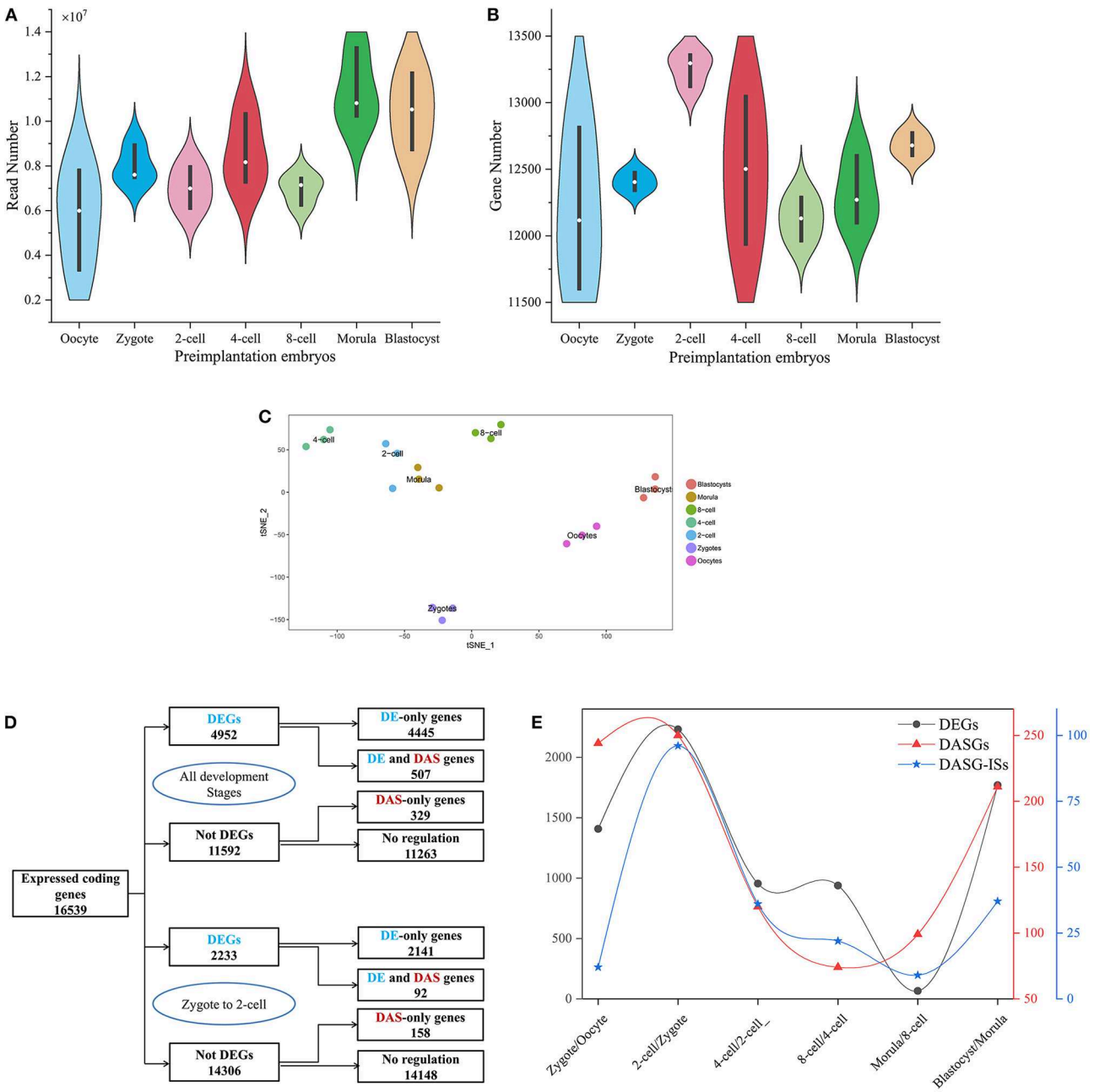


FIGURE 1 | A schematic description of dataset, DEGs and DAS. **(A,B)** is the frequency distribution of reads mapped to protein-coding genes and protein-coding genes captured at every preimplantation development stage. The white circle denotes median of all values. Black rectangle denotes interquartile range (quartile to third quartile). **(C)** Cell type assignment using the most variable genes across all preimplatation development stages following t-SNE-based visualization of 21 cells. Cells that marked with same color were clustered at the same developmental stage. **(D)** The number of DEG and DASGs at all preimplantation development stages. The blue denotes DEGs and the red denotes DASGs. **(E)** The number of DEGs, DASGs and DASG-ISs for every consecutive stages of preimplantation development. DASGs represent genes in which DAS was identified. DAS-ISs represent isoform switches identified from DASGs.

carcinoma and is associated with metastasis (Wrzesinski et al., 2015). Further research is required to investigate the reason for the up-regulation of *Tmem72* in 2-cell and 4-cell stage compared to other development stages and its functional role in the mouse preimplantation development (**Figure 2A**). We also analyzed the expression profiles of occyte-specific genes including *Oas1e*, *Aspm*, *Rgs2*, *Fbxw28*, etc. (**Figure S5**).

In mouse preimplantation development, zygotic genes are activated until 4-cell stage. A total of 1,770 up-regulated genes and 1,428 down-regulated genes between 4-cell and zygote

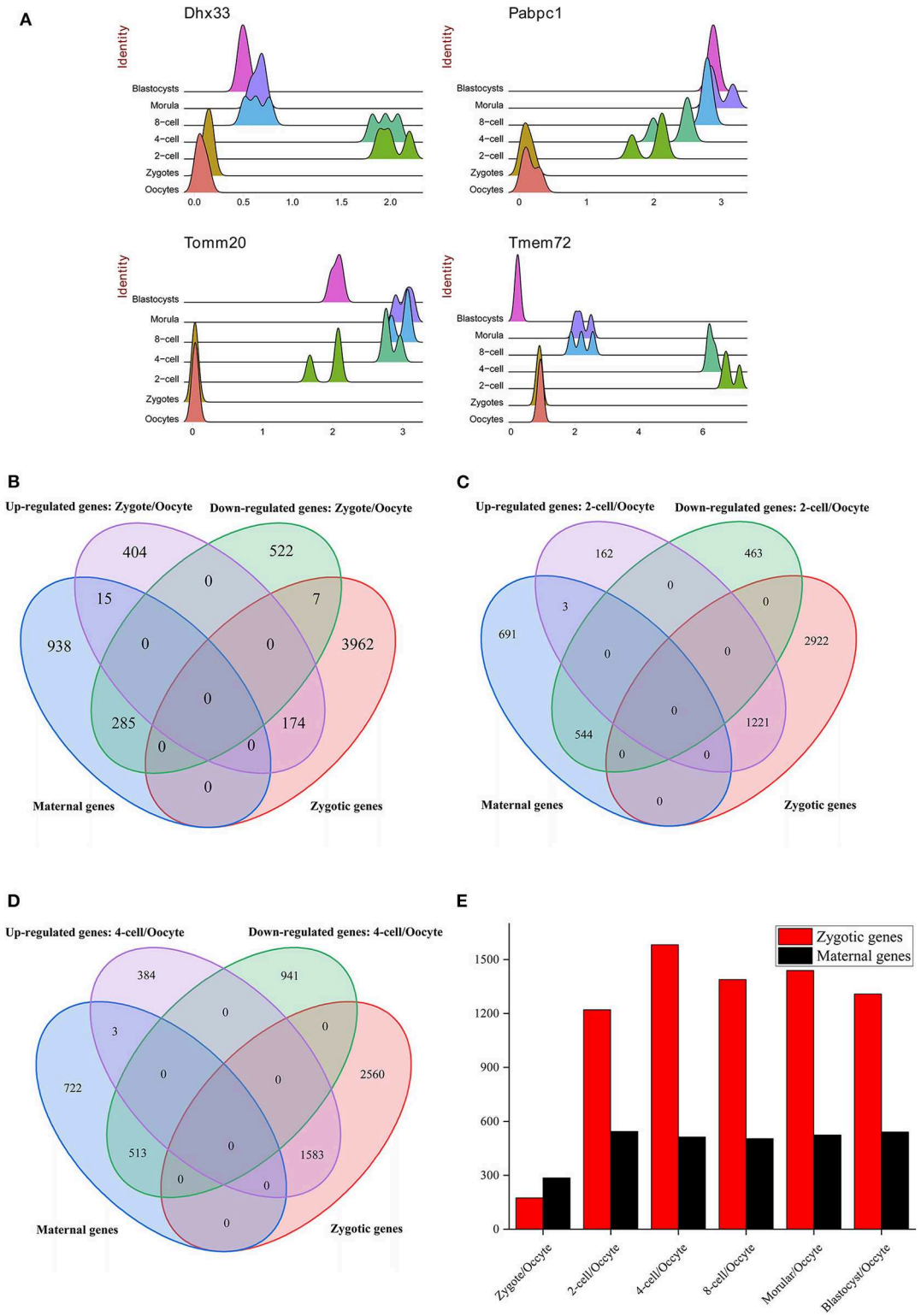


FIGURE 2 | The distribution of DE genes. **(A)** The expression atlas of significantly up-regulated genes in different preimplantation development stages. The gene expression level was normalized by Seurat (see Materials and Methods). **(B–D)** The intersection between DE genes (**B**: zygote/oocyte, **C**: 2-cell/oocyte, **D**: 4-cell/oocyte) and maternal/zygotic genes. The maternal and zygotic gene sets were derived from Fan et al. (2015). **(E)** The distribution of DE genes overlapped with maternal and zygotic genes. Zygotic genes marked with red color denotes the up-regulated genes overlapped with zygotic genes from Fan et al. (2015). Maternal genes marked with black color denotes the down-regulated genes overlapped with maternal genes from Fan et al. (2015).

stages were identified. GO enrichment analysis results for these DEGs were similar with DEGs between 2-cell and zygotic stages (**Figure S6**, **Table S6**). Besides, the DEGs number between 8-cell and morular is smallest compared with other consecutive stages of preimplantation development, suggesting that the transcriptomes of these two stages of embryos are similar.

By using SUPeR-seq, Fan et al. (2015) identified 1,238 annotated maternal genes and 4,143 annotated zygotic genes. In order to illuminate ZGA process, DEGs before and after ZGA were compared with maternal and zygotic genes. Obviously, compared with oocyte, up-regulated and down-regulated genes in zygote, 2-cell and 4-cell are overwhelmingly overlapped with zygotic and maternal genes, respectively (**Figures 2B–D**). By comparing the overlapped genes under different development stages, it was suggested that ZGA initiates during one-cell stage, bursts during 2-cell stage and hit the peak during 4-cell stage (**Figure 2E**). This conclusion is consistent with Abe et al. (2018). They concluded that ZGA in mouse initiates at the mid-one-cell stage (minor ZGA) and is dramatically activated after 2-cell stage (major ZGA). If minor ZGA was inhibited transiently, most of embryos were arrested at the 2-cell stage. Thus, minor ZGA is crucial for the maternal-to-zygotic transition (Abe et al., 2018).

DAS Profiles in Consecutive Developmental Stages of Preimplantation Embryo

It has been well-known that pre-mRNA splicing can occur co-transcriptionally on nascent transcripts. However, isoforms abundance generated by AS may be masked by gene-level measurement. Here, we posed two questions: (i) when is pre-mRNA splicing activated? (ii) is AS activation coupled with ZGA? To answer these questions, we systematically examined the dynamics of AS during mouse preimplantation embryonic development.

SUPPA is a robust tool to study the local AS event level across multiple conditions. We employed it to generate seven simple AS events [alternative 5′ splice site (A5SS), alternative 3′ splice site (A3SS), skipping exon (SE), retained intron (RI), mutually exclusive exons (MXE), alternative first exon (AFE), alternative last exon (ALE)] from annotation file and quantified the AS event inclusion levels (*PSI*) in 7 preimplantation developmental stages. If the *PSI* value of AS event is in the range of 0–1 in every replicate sample of every stage, this event was identified as true AS event in this stage. The number of AS events in every developmental stage corresponding to coding gene was listed in **Table 1**. In mouse annotation file, a total of 58,597 AS events involved with 11,462 coding genes were identified. The ratio of coding genes occurring AS is 52.05%, which is remarkably lower than that in human annotation file (76.67%). The highest proportion of AS pattern is AFE and SE, which account for 39.3 and 23.8% of all AS events, respectively. In seven preimplantation developmental stages, the average number of AS event is 24,802 implicated with 6,877 coding genes, which is distinctly decreasing than that in annotation file (11,462). Similarly, AFE and SE are prevalent in all preimplantation developmental stages. The RI events are relatively sparse. Due to the structural complexity of MXE and limit of SUPPA, the accuracy and number of identifying MXE is poorer. Thus, we didn't take account of MXE pattern in sequence conservation analysis of AS. Isoforms generated by these AS events either encoded different protein variants or regulated the protein concentration via nonsense-mediated decay (NMD) mechanism.

TABLE 1 | The number of AS events for different developmental stages.

	A5SS	A3SS	SE	RI	MXE	AFE	ALE	Sum
GTF	6,335	7,089	13,945	3,059	1,203	23,054	3,912	58,597/11,462*
Oocyte	2,965	3,506	7,701	1,491	394	6,910	1,129	24,096/6,741*
Zygote	3,098	3,684	7,876	1,545	410	7,221	1,174	25,008/6,875*
2-cell	3,276	3,905	8,099	1,698	409	7,387	1,211	25,985/7,147*
4-cell	3,089	3,636	7,426	1,662	360	6,574	1,111	23,858/6,659*
8-cell	3,074	3,663	7,196	1,722	326	6,003	1,004	22,988/6,624*
Morula	3,269	3,835	7,486	1,806	368	6,619	1,101	24,484/6,780*
Blastocyst	3,599	4,160	8,217	1,973	413	7,613	1,224	27,199/7,311*

**The number after "/" in Sum column denotes the gene number involved with AS.*

AS can give rise to distinct protein products. If the length of alternative region of AS event is ($3n$, $n = 1, 2, 3, \ldots$) bp, this AS will conserve the reading frame and only add some new amino acids. Thus, the 3D-structure and function of protein products translated from AS isoforms are similar. If the length of alternative region of AS event is ($3n + 1$) bp or ($3n + 2$) bp, this AS will shift the reading frame and change all amino acids after splice site. Thus, the function of protein products translated from AS isoforms is more variable (Roy and Penny, 2007; Kovacs et al., 2010). Here, we systematically analyzed the ability of conserved reading frame (CRF) of AS events during preimplantation developmental stages (**Table 2**). For the type of RI and A5SS, the average percentage of AS events with length of alternative region equals ($3n$) bp is close to that with length of alternative region equals ($3n + 1$) bp or ($3n + 2$) bp. It means that CRF ability of RI and A5SS is moderate, but the ability of alterative reading frame (ARF) is strong. If RI is widespread, the proteome will become disorder. Under evolutionary selection pressure, the number of RI pattern becomes rare in mammal transcriptome. It has been revealed that the translation of mRNA derived from ARF was often suppressed by a premature termination codon (PTC) that results in NMD of the mRNA product (McGlincy and Smith, 2008). This mechanism imposed restriction on the protein-coding ability of RI and explain that why the function of numerous RI is unclear. On the contrary, the average percentage of AS events with length of alternative region equals ($3n$) bp in type of SE and A3SS is ∼48%, which is significantly higher than that with length of alternative region equals ($3n + 1$) bp or ($3n + 2$) bp (Mann–Whitney U test: $p < 0.01$). This finding suggested that SE and A3SS have strong CRF ability and moderate ARF ability. Thus, SE is universal across mammal genome. Besides, there isn't remarkably difference of the CRF and ARF ability between different developmental stages.

Among the 5,453 genes with two or more AS events occurred at specific stage, 1,979 were in occyte, 2,068 in zygote, 2,288 in 2-cell, 2,074 in 4-cell, 1,783 in 8-cell, 2,059 in morular and 2,229 in blastocyst. The 2-cell stage has the highest number of AS genes,

TABLE 2 | The length characteristic of alternative region of AS.

	A5SS			A3SS			SE			RI		
	%3 = 0	%3 = 1	%3 = 2	%3 = 0	%3 = 1	%3 = 2	%3 = 0	%3 = 1	%3 = 2	%3 = 0	%3 = 1	%3 = 2
Oocyte	0.3838	0.3295	0.2867	0.5000	0.2496	0.2504	0.4951	0.2497	0.2552	0.3387	0.3508	0.3105
Zygote	0.3838	0.3254	0.2908	0.4891	0.2638	0.2470	0.4892	0.2541	0.2567	0.3366	0.3476	0.3159
2-cell	0.3846	0.3291	0.2863	0.4891	0.2671	0.2438	0.4843	0.2532	0.2625	0.3345	0.3321	0.3333
4-cell	0.3784	0.3305	0.2910	0.4860	0.2690	0.2451	0.4760	0.2586	0.2654	0.3273	0.3454	0.3273
8-cell	0.3709	0.3390	0.2902	0.4767	0.2722	0.2512	0.4722	0.2639	0.2639	0.3275	0.3444	0.3281
Morula	0.3741	0.3334	0.2924	0.4746	0.2751	0.2503	0.4698	0.2634	0.2668	0.3300	0.3439	0.3261
Blastocyst	0.3682	0.3331	0.2987	0.4748	0.2779	0.2474	0.4737	0.2613	0.2651	0.3300	0.3431	0.3269
DAS	0.3743	0.3073	0.3184	0.5120	0.2771	0.2108	0.4825	0.2558	0.2616	**0.3800**	0.2200	0.4000

%3 = 0, %3 = 1 and %3 = 2 denote that the length of alternative region of AS equals (3n) bp, (3n + 1) bp, and (3n + 2) bp, respectively. The value in every cell denotes the percentage.

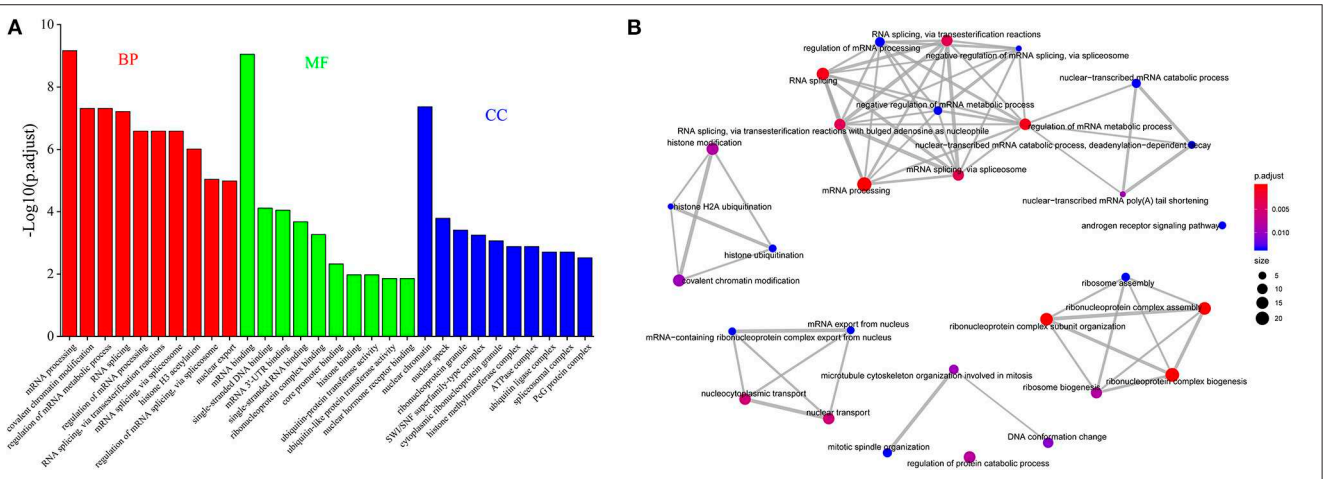


FIGURE 3 | The functional analysis of genes occurring DAS. **(A)** GO enrichment analysis of a conserved gene dataset occurring AS during preimplantation development. **(B)** The network of the enriched GO-BP terms for DASGs between zygote and 2-cell stage.

implicating that the transcriptome profile in 2-cell stage is more complicated. The transcript diversity before zygote was mainly originated from maternal transcripts. After zygote, the zygotic transcripts begin to synthesize. Thus, we observed an elevated AS number during zygote and 2-cell stages. Therefore, it can be proposed that AS may be activated after zygote, especially at 2-cell stage. This process is referred to as ZASA. Obviously, the time point of ZASA is in conformity with ZGA. It was known that the embryonic stem cells in blastocyst stage will be rapidly differentiated into endoderm, mesoderm and ectoderm lineages (Feng et al., 2012). A lot of regulatory proteins need to regulate this highly sophisticated differentiation process. More AS in blastocyst could provide an important source of protein diversity in this stage. Hence, the elevated AS number was observed in blastocyst (**Table 1**). These data suggested that the profile of AS is dynamic at different stages during preimplantation stages. Besides, we founded that the vast majority of AS events in preimplantation developmental stages is biased toward either high (>80%) or low (<20%) inclusion ratio (**Figure S7**), which is consistent with the previous study by Busch and Hertel (2013).

Furthermore, we observed that 213 genes were expressed with multiple AS events (≥2) within every cell at all of the seven developmental stages and of which 12.68% is overlapped with splicing factor. The GO enrichment analysis showed that many genes were enriched on pre-mRNA splicing regulation (**Figure 3A**). We extrapolated that a conserved gene set regulating AS during preimplantation development might be found.

Moreover, by combining the transcript-level data, diffSplice module of SUPPA was utilized to identify DAS and DASGs between different preimplantation development stages (see Materials and Methods). A total of 6,546 DASs derived from 5,610 DASGs were identified for the contrast groups of 7 preimplantation development stages. After deleting duplicates between different contrast groups, 2,269 DASs embedded in 1,060 DASGs were listed. For the seven consecutive development stages, 1,170 DAS derived from 998 DASGs were identified. After deleting duplicates between different contrast groups, 1,060 DASs embedded in 836 DASGs were listed (**Table 3**). It was shown that the number of DAS and DASGs from zygote to 2-cell stages is the greatest (**Figure 1E**). This peak point is also coincided with ZGA. Besides, we elaborated the distribution of DAS pattern in consecutive development stages (**Table S7** and **Figure S8**). It was indicated SE and AFE are the most widespread DAS. However,

TABLE 3 | The number of DAS and DASGs.

	Oocyte	Zygote	2-cell	4-cell	8-cell	Morula	Blastocyst
Oocyte	0	**281/244**	380/311	426/350	392/326	479/382	565/449
Zygote		0	**304/250**	365/291	342/283	382/318	461/376
2-cell			0	**138/120**	218/192	247/208	422/344
4-cell				0	**85/74**	126/113	310/259
8-cell					0	**115/99**	261/210
Morula						0	**247/211**
Blastocyst							0

Cutoff: p-value ≤ 0.05 and dPSI ≥ 0.1. Bold digits is corresponding to the seven consecutive development stages.

the SE percentage of DAS between occyte and zygote is the highest, implying that DAS generated by SE may be important for the formation of zygote. It must be emphasized that **Table 2** showed that the ability of CRF of RI events in DAS was elevated. It implied that RI may play a great role in regulatory mechanism associated with DAS events.

In addition, we found that many DASGs are significantly enriched in GO-BP terms of splicing regulation, stem cell population maintenance, ribosome assembly, histone modification, etc. Especially between zygote and 2-cell stage, a total of 83 GO-BP terms involved 122 DASGs were significantly enriched, of which the majority terms were associated with splicing regulation. For dissecting the function of enriched terms, we interwoven the top 30 most significantly enriched terms into a network with edges connecting overlapping gene sets (**Figure 3B**). The larger the mutually overlapping gene sets were, the more likely the terms to be clustered together. It was indicated that five functional modules were identified, of which a module was involved with histone modification, and all of the other modules were closely related to pre-mRNA splicing process. It is well-known that histone modification is a key marker of exon definition and AS regulation (Luco et al., 2011; Zhou et al., 2014). It may imply that DASGs could play important roles for AS regulation during preimplantation development.

Also, each stage-specific DEGs and DASGs were identified (**Table S8**). Obviously, the number of stage-specific DEGs and DASGs between 2-cell and zygotes are greatest. The KEGG pathway enrichment showed that the majority of the pathways involved in stage-specific DEGs were significantly enriched in RNA transport, spliceosome, mRNA surveillance pathway, oocyte meiosis, cell cycle, and disease, etc. The stage-specific DASGs were mainly enriched in the pathway of mRNA surveillance and hormone signaling.

AS of Splicing Factors Associated With Pre-embryonic Development

RNA-binding proteins (RBPs) play critical roles in post-transcriptional gene regulation (PTGR), such as regulation of AS, mRNA stabilization, mRNA location, polyadenylation and translation. Gerstberger et al. (2014) manually curated 1,542 human RBPs that interact with all known classes of RNAs, described their families and evolutionary conservation across species, and analyzed their expression across tissues and their potential roles in developmental processes. The mechanism of AS involves *cis*-acting RNA elements, *trans*-acting proteins, epigenetic factors, etc. Most of these trans-acting proteins are RBPs, especially SFs (Carazo et al., 2018).

In this study, a total of 446 mouse splicing factors were selected for analysis based on literature mining for previously described splicing functions (Han et al., 2013; Goldstein et al., 2017), and "RNA splicing" or "splicesome"-associated Gene Ontology (GO) terms from MGI. Venn diagram displayed ∼77% (342) SFs is included in human RBPs (**Figure S9**). We observed that about 50% (207/446) SFs are differentially expressed during preimplantation development (**Figure 4A**). Furthermore, 39 SFs have intersection with 836 DASGs, which means these SFs undergo self-AS in the regulatory process of mRNA processing (**Figures 4B,C**, **Table S4**). Moreover, 17 DASGs that belong to SF between zygote and 2-cell were detected. As compared to other five consecutive stage of preimplantation development, this number is the greatest, suggesting that more SFs function during ZASA especially from zygote to 2-cell stage.

SFs are pivotal factors for all AS regulation. For clarifying the specific of these SFs on preimplantation development, we performed hierarchical clustering based on Pearson Correlation coefficient of gene expression level between different developmental stages (**Figure 5**). Cells that clustered together were at the same developmental stages in all cases, with the exception that a morula-stage cell was interchanged with a blastocyst-stage cell. Furthermore, the developmental time series was also approximately captured from oocytes to blastocysts, as neighboring stages clustered together in the analysis as to be expected, similar to what has been previously reported by Yan et al. (2013). Only gene expression information of 39 SFs was employed, but the clustering result was good enough. This phenomenon suggested these SFs expression is specific for preimplantation development and is crucial for normal preimplantation development.

SRSF3 (serine/arginine-rich splicing factor 3, alias: Srp20) is the smallest member of the SR proteins (serine-arginine-rich family of nuclear phosphoproteins) family of splicing factors. In UniProt Database, SRSF3 has two transcript isoforms (P84104–1 and P84104–2), and P84104–1 is the dominant isoform. P84104–2 is produced at very low levels due to a premature stop codon in the mRNA, leading to NMD. Interaction with YTHDC1, a RNA-binding protein that recognizes and binds N6-methyladenosine (m6A)-containing RNAs, promotes recruitment of SRSF3 to its mRNA-binding elements adjacent to m6A sites, leading to exon-inclusion during AS. It was revealed that SRSF3 is essential for mouse development. If SRSF3 was knocked out, blastocyst formation was prevented and caused death of preimplantation embryos at the morular stage (Jumaa et al., 1999). SRSF3 is also essential for later developmental decisions, such as those in B-cell development. In consistent with this conclusion, we observed that gene expression level of SRSF3 is remarkable elevated in morular and blastocyst stages compared to in oocytes and early stages of embryonic development. Besides, SRSF3 has the lowest expression level in 2-cell stage. It was revealed the majority maternal RNAs of SRSF3 are degraded in 2-cell stage (**Figure 6**).

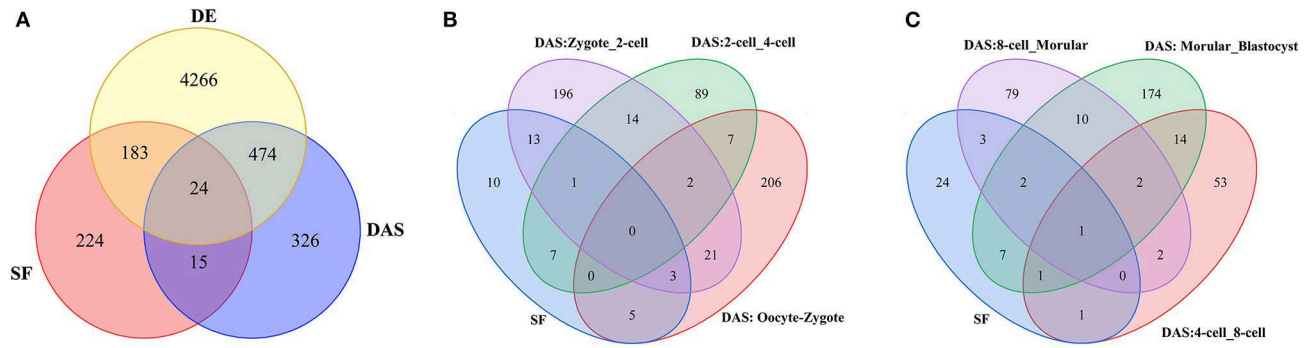


FIGURE 4 | (A) Venn diagram of SF, DE and DASGs across seven consecutive stages of preimplantation development. SF denotes 446 splicing factors. DE denotes 4,947 DE genes in consecutive development stage. For example, DAS denotes 836 genes undergoing DAS in consecutive development stage. **(B,C)** Number of SFs at every consecutive stage of preimplantation development. SF represents 39 SFs undergoing DAS across seven consecutive stages of preimplantation development. For example, DAS: Occyte_Zygote denotes these genes undergoing DAS from occyte to zygote stages.

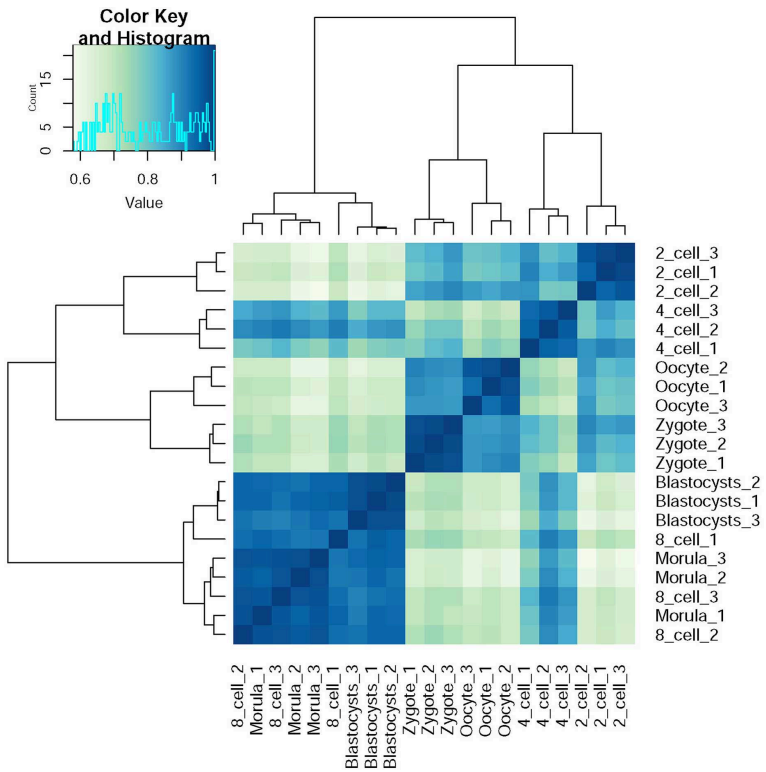


FIGURE 5 | The hierarchical clustering of preimplantation development stages based on specific SFs.

The gene expression level of SRSF3 escalates after this stage. It was indicated that majority transcripts of SRSF3 are synthesized by zygotic activation.

Furthermore, we also identified an exon skipping AS event in SRSF3, of which isoforms are translated P84104–1 and P84104–2. The detailed AS profile of SRSF3 can be viewed in **Figure 6**. The P84104–1 skips an exon (chr17:29039454–29039909) and the transcript expression level is higher, particularly in morular and blastocysts. Because the degradation of maternal RNAs, exon inclusion level *PSI* in 2-cell stage is significantly lower than other developmental stages. After 2-cell stage, transcript expression level of P84104–1 is compensated by ZGA. The isoform P84104–1 is dominant in morular and blastocyst stage. The blastocyst is the first developmental stage with known differentiated cell lineages, suggesting that P84104–1 isoform of SRSF3 is essential for initiating this early genetic programme. This result also implied that the AS of SRSF3 is popular in pre-embryonic development stages (Sen et al., 2013). The relative concentration of RNA-binding activator and repressor of splicing machinery is an important regulator of splice-site recognition (Wang

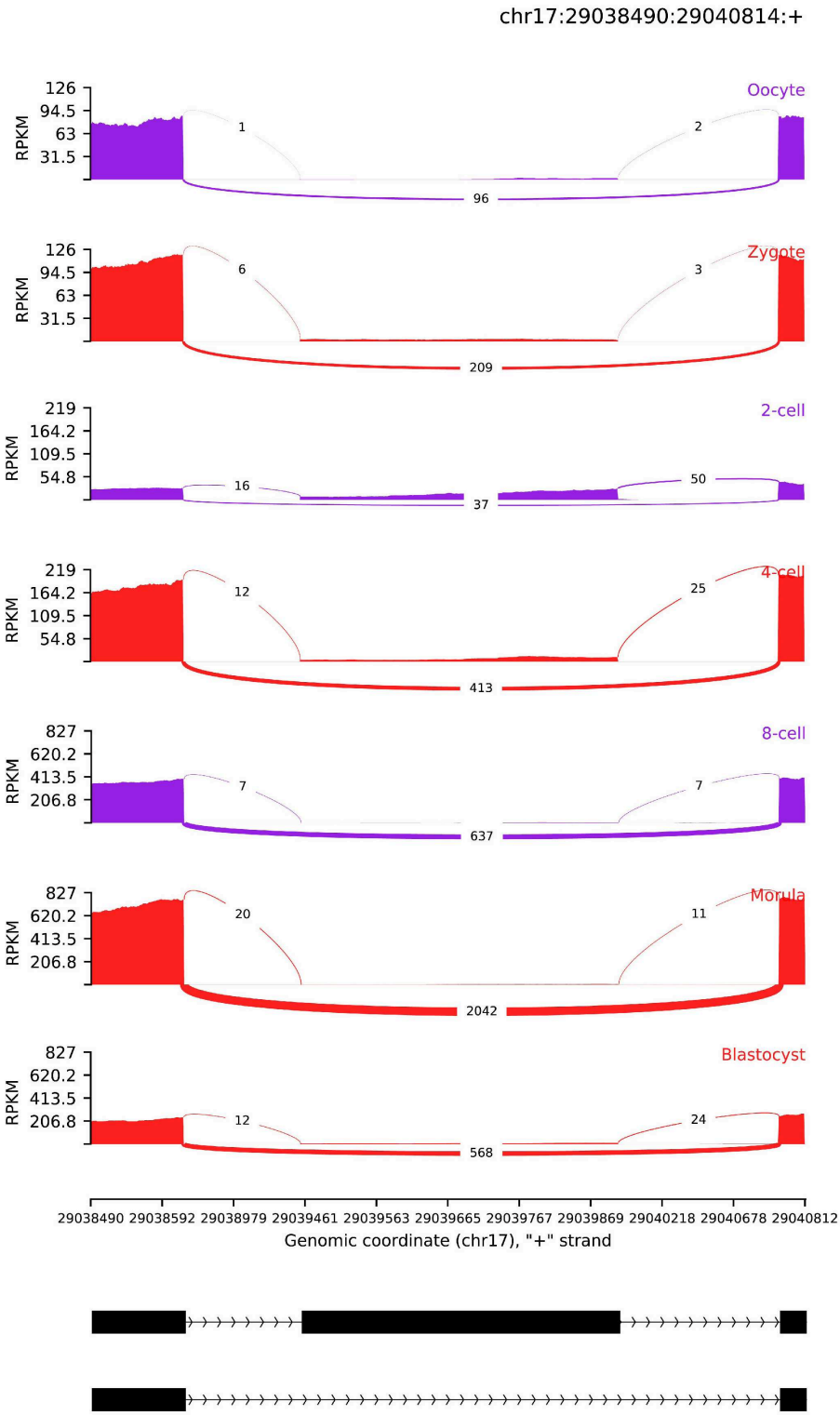


FIGURE 6 | AS read coverage of SRSF3. The x-axis and y-axis denotes genomic coordinate and transcript expression level (RPKM), respectively. The black rectangle and line represent exon and intron, respectively.

et al., 2015). SR proteins and hnRNPs (heterogeneous nuclear ribonucleoproteins) are RNA-binding activator and repressors, respectively. As a SR protein, the concentration of SRSF3 can be modulated by self-splicing. Then, splice site recognition of other genes which have the potential binding site of SRSF3 can be regulated by SRSF3.

The Identification of ISs in Consecutive Stages of Preimplantation Development

Over 1,000 DAS events embedded into 836 genes were identified during seven consecutive development stages. Every DAS gene included more than one isoform. We used Time-Series Isoform Switch (TSIS) program to detect ISs, where the expression level of different isoforms is reversed during preimplantation development (Guo et al., 2017). As input file of TSIS, the abundance (TPM) of 3,096 transcripts involved with all DAS was extracted from transcript expression matrix. A total of 212 significant ($p < 0.05$) ISs that embraced two transcript isoforms were identified in 836 unique DASGs. TSIS determines the two time points between which a significant isoform switch occurs, and consistent with the DE and DAS, the majority (62.26%) occurred between 2 cell/zygote and 4 cell/2-cell (**Figure 1E** and **Table S9**). Thus, in response to ZGA, there are crest of ISs between 2 cell/zygote.

Supt6 (alias: *Spt6* or *Supt6h*) is a transcription elongation factor which binds histone H3 and plays a critical role in the regulation of transcription elongation and mRNA processing. It produces two different transcript isoforms, of which *Supt6*-201 can be translated into protein with 1,726 residues, and protein product of *Supt6*-202 has not been detected yet (Hubbard et al., 2002). The transcript abundance can be modulated by SE of exon 6 in *Supt6*-202. *Supt6* showed a significant IS between 4-cell and 8-cell (**Figure 7A**). It suggested protein factor translated from *Supt6*-201 plays a crucial role in ZGA, and *Supt6*-202 may play a role in post-implantation development. During ZGA, *Supt6* can promote activation of transcriptional elongation via *Tat*, and enhance the transcription elongation by RNA polymerase II (RNAPII). *Supt6* can also recruit mRNA export factors (*Alyref*/*Thoc4, Exosc10*) and histone-lysine N-methyltransferase (*Setd2*) to assist mRNA splicing, mRNA export and elongation/splicing-coupled H3K36 methylation by forming *Supt6*:IWS1:CTD complex. Xu et al. (2019) revealed *Setd2* plays a vital role in establishing the maternal epigenome and exerts important impacts for preimplantation. The expression profile of *Setd2* is similar with *Supt6* (**Figure S10**), which demonstrated *Setd2* and *Supt6* may locate in the same regulated network. *Setd2* generates 10 transcripts, which were respectively annotated as protein-coding, nonsense mediated decay or no-protein isoforms. We observed transcript abundance of *Setd2*-201, *Setd2*-204, and *Setd2*-210 is dominant and the trend of these transcripts is identical with *Setd2*. The *Setd2*-201 and *Setd2*-210 was already labeled as protein-coding isoforms. However, the protein product from *Setd*2-204 has not been found so far. It implied that *Setd2*-204 may exert important function in previously unknown pathways during preimplantation development.

Adar enzyme can catalyze the hydrolytic deamination of adenosine to inosine (A-to-I) in double-stranded RNA (dsRNA). It may participate in biological regulation in a number of ways that include mRNA translation, pre-mRNA splicing, RNA stability, genetic stability and RNA structure-dependent activities, and so on (The UniProt Consortium, 2019). *Adar* modulates trans-acting factors involved in the AS machinery by affecting splicing regulatory elements (SREs) within exon (Solomon et al., 2013). Here, three transcript isoforms with TPM = 0 in 21 cells were dropped, and expression profiles of 6 transcript isoforms were analyzed (**Figure 7B**). Qiu et al. (2016) constructed A-to-I RNA editome during early human embryogenesis and demonstrated *Adar* expression and A-to-I RNA editing level remained relatively stable until 4-cell stage, but dramatically decreased at 8-cell stage, continually decreased at morula stage. Similar to human embryogenesis, in mouse embryogenesis, we demonstrated *Adar* expression level was stable and remarkably elevated until 2-cell stage, but sharply decreased at 4-cell stage, continually decreased until blastocysts stage. It was deduced that A-to-I RNA editing level is also parallel with *Adar* expression level in mouse embryogenesis. García-López et al. (2013) has revealed that A-to-I editing in microRNAs in mouse preimplantation embryos is mediated by *Adar*. We speculated A-to-I RNA editing is dynamically changed during preimplantation development in a stage-specific fashion and plays a vital role in activating zygotic genes. Furthermore, a clearly IS was identified between zygote and 2-cell. Expression level of *Adar*-204 sharply increased and that of *Adar*-201 and *Adar*-205 significantly decreased from zygote to 2-cell stage. It indicated that (1) on the condition that *Adar* expression level is relatively constant, abundance of transcript isoforms is variable during preimplantation development; (2) transcript isoforms executing dominant regulating role is different during different preimplantation development stages. Besides, the lincRNA (*Adar*-206) was expressed during ZGA. As non-coding RNA, *Adar*-206 may execute special regulatory role during ZGA.

AS is a common form of post-transcriptional regulation in metazoan. Concomitantly, it has been estimated that over one third of the AS events also create aberrant transcript isoforms that trigger NMD pathway (Bao et al., 2016). As a RNA surveillance mechanism, NMD machinery eliminates aberrant transcript harboring PTC (premature termination codon) signal and plays an essential role in safeguarding the transcriptomic fidelity in the cell. The NMD machinery includes three core factors: *Upf1*, *Upf2*, and *Upf3*, in addition to *Smg1-7*, which are highly conserved in eukaryotes (Schweingruber et al., 2013). In recent years, some studies demonstrated that *Upf2*-dependent NMD pathway performs an essential role in Spermatogenesis, tissue development, disease (Thoren et al., 2010; Nguyen et al., 2014; Bao et al., 2016). To explore whether the NMD pathway plays a role in mouse embryogenesis, we plotted expression profiles of *Upf2* (**Figure 7C**). It was showed that *Upf2* expression level is fluctuant and reached peak at zygote stage. We can extrapolate NMD pathway is critical for preimplantation development, especially for fertilization. By analyzing IS, we observed that *Upf2*-202 isoform is dominant during fertilization and blastocyst formation, and *Upf2*-201 isoform is more prevalent during ZGA. Besides, as lincRNA, the expression level of *Upf2*-203 is very lower.

Cnot6 is a subunit of the CCR4-NOT core transcriptional regulation complex, which is one of the major cellular mRNA deadenylases. It is linked to various cellular processes including transcription and translation regulation, mRNA degradation, miRNA-mediated repression, cell proliferation, cell survival and

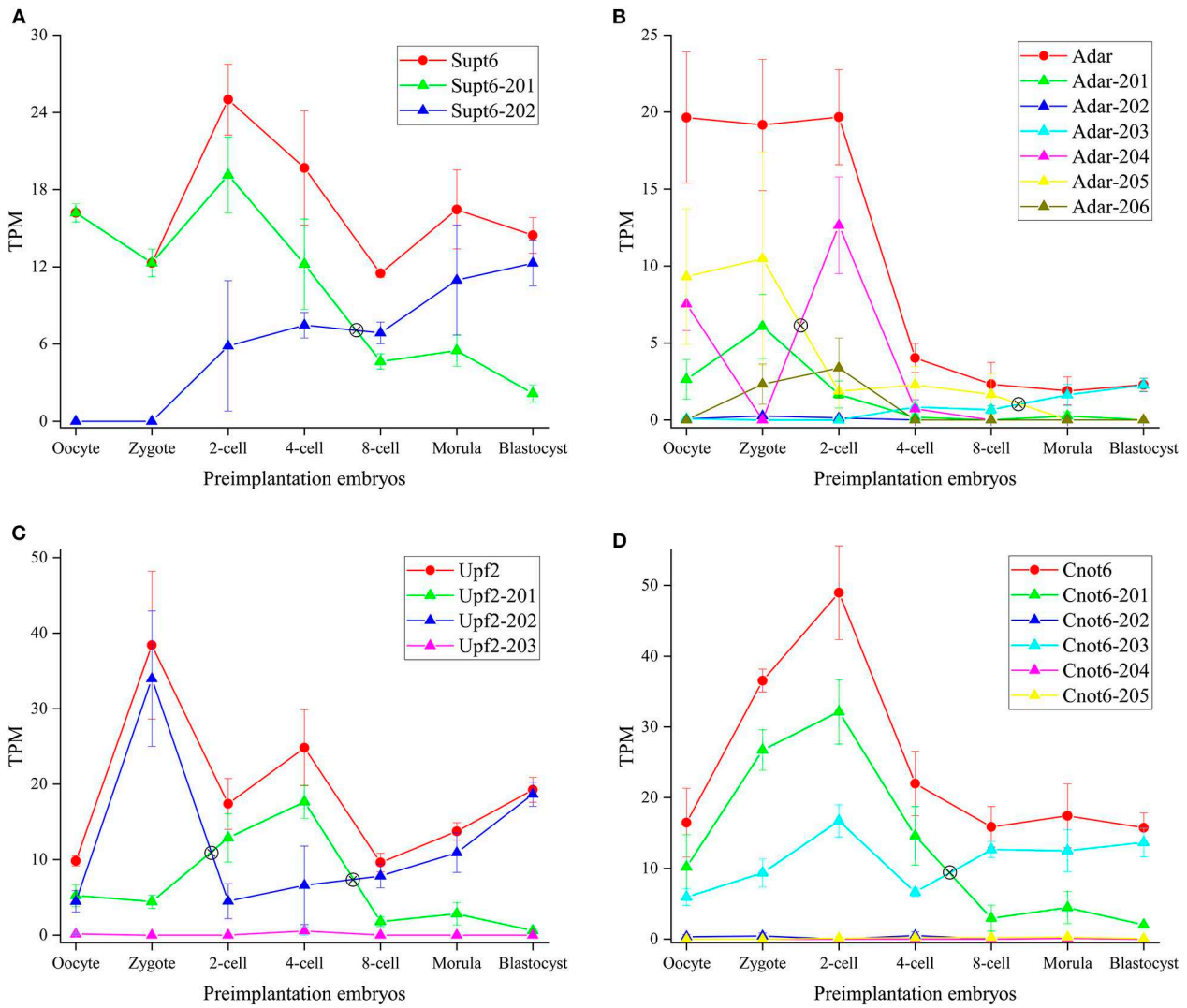


FIGURE 7 | Expression profiles of DASGs. **(A–D)** Represent the gene and transcript expression profiles of *Supt6, Adar, Upf2* and *Cnot6*, respectively. The y-axis denotes TPM (Transcripts Per Kilobase Million). The symbol ⊗ denotes switch point. The red line denotes gene expression level and other color lines denote transcript expression level.

cellular senescence. This gene has 5 transcripts, of which *Cnot6*-201and *Cnot6*-203 are translated into proteins and *Cnot6*-202, *Cnot6*-204, and *Cnot6*-205 are labeled as lincRNA. Previous work revealed that *Cnot1* and *Cnot3* are critical for deadenylation of maternal mRNA during mouse early embryogenesis (Ma et al., 2015; Liu et al., 2016). The expression profile showed *Cnot6*-201 and *Cnot6*-203 is overwhelming expressed product during preimplantation development (**Figure 7D**). Before 8-cell stage, *Cnot6*-201 is main transcript product. On the contrary, *Cnot6*-203 isoform is dominant after 8-cell stage. It suggested *Cnot6*-201 executes main role during ZGA, and *Cnot6*-203 plays important role in development of inner cell mass and blastocyst formation.

Moreover, we also plotted expression profiles of *Msh4* and *Luc7l*. *Msh4* is involved in meiotic recombination and segregation of homologous chromosomes at meiosis (**Figure S11**). A differential SE event and IS were identified between zygote and 2-cell. Obviously, these transcript isoforms is maternal-derived and its expression decreases with preimplantation development. The *Luc7l* encodes a putative RNA-binding protein similar to the yeast *Luc7p* subunit of the U1 snRNP splicing complex that is normally required for 5′ splice site selection. The expression of *Luc7l* and its 9 transcript isoforms exhibit oscillated patterns with preimplantation development, suggesting that *Luc7l* is likely to play a role during ZGA (**Figure S12**).

DISCUSSION

During preimplantation development from an occyte, cells progressively develop toward to the blastocyst as zygotic genome

is activated. It has been extensively studied that gene expression level is spatial-temporally dynamic during early embryonic development (Fan et al., 2015; Schultz et al., 2018). Recently, some evidence indicated that AS could correlate closely with preimplantation development, suggesting a key role for splicing in regulating early embryonic development (Revil et al., 2010; Yan et al., 2013). However, previous results about the diversity and function of AS in early embryonic development were mainly based on a few isolated examples. In this study, we carried out genome-wide comprehensive analysis on seven preimplantation developmental stages to capture the dynamic changes of gene expression and AS during early stages of embryonic development.

The accurate identification and quantification of transcripts and genes paves the way of downstream omics analysis. Here, the performance of different quantifying strategies was compared (**Table S10**). As alignment-free transcript quantification, the Salmon outperforms HISAT2 belonged to alignment-based transcript quantification. Since our goal was to measure the abundance of the known coding-gene isoforms, we selected Salmon to perform transcript and gene quantification in 21 scRNA-seq datasets. Identifying the set of DEGs across different developmental stages is an important goal in this study. DESeq2 for analyzing count-based NGS data can accurately detect DEGs in bulk RNA-seq data. We observed that the number of DEGs identified by DESeq2 was greater than that by Seurat (**Table S10**). However, the comparison of different quantifying strategy between 2-cell and zygote showed 95.83% DEGs identified by Salmon + Seurat had been included in those identified by Salmon + DEseq2 (**Figure S13**). We postulated that if DESeq2 were applied directly to scRNA-seq data, the false-positive rate would be relatively high. In contrast, the result derived from Seurat would be more accurate (Freytag et al., 2018). Thus, we employed Seurat to detect DEGs during preimplantation development.

In the time-course analysis of preimplantation embryo, 4,952 DEGs and 836 DASGs were respectively detected in consecutive seven developmental stages. Over 10% DEGs were differentially alternatively spliced. It suggested that the crosstalk between transcription and AS regulation might occur during preimplantation development. In concert with major ZGA in mouse preimplantation embryo (Abe et al., 2018), DEGs between 2-cell and zygote achieved the maximum, especially up-regulated genes. Functional enrichment analysis of DEGs revealed that with the initiation of transcription and translation, splicing machinery may also be assembled during ZGA. It was well-known that co-transcriptional splicing is ubiquitous for long mammalian genes (Luco et al., 2011). The fact that transcription and splicing machinery is simultaneously established during ZGA may indicate co-transcriptional splicing maybe universal in preimplantation development. Based on differentially expressed genes during zygote to 2-cell stages, Zeng and Schultz (2005) employed Ingenuity Pathway Analysis (IPA) to identify 25 regulatory networks implicated with ZGA. The most remarkable network is composed of 35 genes and centered on *Myc*. By filtering the DEGs between zygote and 2-cell stages, we identified 34 up-regulated genes embedded this network. Out of the 34 genes, 21 genes are above the threshold (adjust p value ≤ 0.05 and $\left|\log_e FC\right| \geq 0.25$). Most genes in this network belong to ribosomal genes. This is consistent with protein synthesis and ribosome biogenesis being two major biological themes that emerge from zygote to 2-cell embryos. Besides, we constructed the top DEGs dataset between 2-cell and zygote stages including 134 top up-regulated and 152 top down-regulated genes, and deciphered the biological function. This dataset was dependable and provided a guideline for decoding the ZGA mechanism from experiment. We also detected that the number of DEGs between 4-cell and zygote stages was larger than that between 2-cell and zygote stages. It confirmed zygotic genes were activated until 4-cell stage during mouse preimplantation development. By characterizing the frequency distribution of maternal and zygotic genes, the conclusion that ZGA includes minor ZGA and major ZGA was verified.

On average, 24,802 AS events, which involved in 6,877 multi-exon protein-coding genes, were identified in every preimplantation developmental stage. The gene number occurring AS is remarkably lower than that in annotation file. This result can be mainly caused by the lower transcript complexity of preimplantation embryos, the limitation of sequencing depth, the imperfect annotation file and the defective of tools. We investigated the CRF and ARF of AS patterns and found the CRF ability of SE was significantly stronger than that of RI. It is well-known that AFE can change the gene expression level by modulating promoter activity. Thus, the percentage of the AFE and SE are dominant and this tendency is conserved in seven developmental stages. By counting the gene number occurring AS events, we found the AS profile is dynamic at different preimplantation developmental stages and the gene number with AS in 2-cell is higher than other stages. It was concluded that the time point of ZASA was coincided with ZGA and AS was activated around ZGA. Besides, a conserved gene set composed of 213 genes was constructed, which is expressed in every cell of all stages with multiple AS events and regulates AS during preimplantation development.

By identifying and analyzing DAS and DASG, we found the number of DAS and DASGs from zygote to 2-cell stages was the greatest. This result once again demonstrated that ZASA may be coupled with ZGA. During ZGA, a mass of regulated proteins are recruited to regulate gene activation. DAS of pre-mRNA can provide more diversely regulated proteins, which ensure that ZGA is executed successfully (Hamatani et al., 2004; Revil et al., 2010; Park et al., 2018). The functional enrichment analysis demonstrated that many DASGs may play important roles in splicing regulation. For DASGs between zygote and 2-cell stage, 5 functional modules closely related to pre-mRNA splicing process were hunted. It can be inferred that DASGs may be key regulator of AS during preimplantation development. This result also verified that AS may be activated with ZGA from the perspective of potential biological function and pathway.

As trans-acting proteins, SFs execute critical roles in AS. Over 50% SFs are differentially expressed and 39 SFs are differentially spliced during mouse preimplantation development. Especially from zygote to 2-cell, 17 SFs were differentially spliced. This

finding showed SFs preform more function during ZASA especially from zygote to 2-cell stage. Furthermore, only using 39 SFs spliced differentially, almost all of samples can be clustered correctly. It demonstrated expression profiles of SFs are specific for different preimplantation development stages. To take SRSF3 as an example, we elaborated the dynamic changes of gene and transcript isoforms coverage during preimplantation development. Gene expression differences and AS of SFs affect the splicing modulation of a large number of targeted AS events, suggesting the existence of a regulatory cascade that SFs may regulate AS by DE and AS of their own gene transcripts during preimplantation development.

Expression level of transcript isoforms always is hidden by gene expression. Although the gene expression is relatively constant, the dominant transcript isoform is variable during time-series in early embryonic development. We identified 212 ISs in 836 DASGs where the expression level of different isoforms is reversed during preimplantation development. It must be emphasized that the crest of ISs number occurs between 2 cell/zygote. This result indicated the role of ISs during ZGA and once again confirmed that the ZASA and ZGA are synchronous. We characterized the expression profiles of gene and their transcript isoforms during 7 developmental stages and predicted the regulatory function of every transcript isoforms. *Supt*6 performs regulation function in transcription elongation and mRNA processing. We unveiled *Supt6*-201 and *Supt6*-202 may play pivotal roles in ZGA and post-implantation development, respectively. Furthermore, we investigated the crosstalk between *Supt*6 and *Setd*2, and showed *Setd2*-204 may exert important function in previously unknown pathways during preimplantation development. This provided a new insight to decoding the *Setd*2. By charactering expression profile of *Adar* and their transcripts, we proposed that A-to-I RNA editing level is dynamic during preimplantation development and play a vital role in activating zygotic genes. Moreover, it was uncovered that the lincRNA (*Adar*-206) exerts special regulatory role during ZGA. After analyzing IS feature of *Upf2* implicated with NMD pathway, the dominant isoform is identified at every developmental stage. This will facilitate researchers to clarify the NMD mechanism. As a major cellular mRNA deadenylases, the *Cnot6* expression level is significantly increased during ZGA. We can infer from the remarkable IS that *Cnot6*-201 performs key role during ZGA, and *Cnot6*-203 may play vital role in development of inner cell mass and blastocyst formation. In summary, unraveling regulatory role of DASGs during embryogenesis from transcript abundance profiles provided a new way for decoding the mystery of preimplantation development.

Overall, the dynamic atlas of DE, AS, and DAS over preimplantation development was established and was comprehensively analyzed. It was inferred that splicing factors could auto-regulate AS by self-DE and self-AS during preimplantation development. Over 200 ISs which may play crucial roles during early embryogenesis were identified. Importantly, we uncovered that ZASA is coincided with ZGA and verified that AS is coupled with transcription during preimplantation development in mouse. This study provided valuable resource and specific functional predictions for further targeted experimental validations to elucidating the regulated mechanisms of embryogenesis and early embryotic development.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YX analyzed the data, wrote, reviewed, and edited the manuscript. WY, HM, and ZL constructed the dataset. GL contributed to discussing and reviewing the manuscript. XC constructed the regulatory network. HZ, XZ, and JL discussed the results. LC and MZ supervised the study.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (61662055, 61671256, 31660322, 61861036), the Natural Science Foundation of Inner Mongolia (2018MS03024, 2019MS03081), the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT-20-B05), and Inner Mongolia University of Science and Technology Innovation Fund for Excellent Young Scholars (2019YQL01).

ACKNOWLEDGMENTS

We would like to acknowledge Yong Chen, Peng Xie, Yang Yang for many helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00035/full#supplementary-material

Table S1 | The basic information of dataset.

Table S2 | Transcript expression matrix of coding genes.

Table S3 | Count Matrix of coding genes.

Table S4 | Splicing factors.

Table S5 | GO enrichment of DE genes between 2-cell and Zygote.

Table S6 | GO enrichment of DE genes between 4-cell and Zygote.

Table S7 | The distribution of DAS patterns in consecutive development stages.

Table S8 | The number of stage-specific DEGs and DASGs.

Table S9 | Isoform switches (ISs) identified in 836 unique DASGs.

Table S10 | Comparison for different detecting methods of DE genes.

Figure S1 | The heatmap showing the expression levels (*TPM* value) of DEGs across seven consecutive stages of preimplantation development. The row names are the name of DEGs.

Figure S2 | The heatmap showing the inclusion levels (*PSI* value) of DAS events across seven consecutive stages of preimplantation development. The row names are the name of DAS events.

Figure S3 | GO-MF and GO-CC enrichment analysis of up-regulated and down-regulated genes between 2-cell and zygote stages. The dotplots only showed the top 15 most significant terms.

Figure S4 | The gene expression atlas of significantly down-regulated genes in different preimplantation development stages. The gene expression level was normalized by Seurat (see Materials and Methods).

Figure S5 | Functional enrichment analysis of the top DE genes between 2-cell and zygote stages.

Figure S6 | GO-BP, GO-MF and GO-CC enrichment analysis of up-regulated and down-regulated genes between 4-cell and zygote stages. The dotplots only showed the top 15 most significant terms.

Figure S7 | The inclusion ratio distribution of AS events in zygote.

Figure S8 | The distribution of DAS pattern between consecutive development stages.

Figure S9 | Venn diagram between SFs of mouse and human RBPs.

Figure S10 | Expression profile of *Sted2*. The *y*-axis denotes TPM (Transcripts Per Kilobase Million). The symbol ⊗ denotes switch point. The red line denotes gene expression level and other color lines denote transcript expression level.

Figure S11 | Expression profile of *Msh4*. The *y*-axis denotes TPM (Transcripts Per Kilobase Million). The symbol ⊗ denotes switch point. The red line denotes gene expression level and other color lines denote transcript expression level.

Figure S12 | Expression profile of *Luc7l*. The *y*-axis denotes TPM (Transcripts Per Kilobase Million). The symbol ⊗ denotes switch point. The red line denotes gene expression level and other color lines denote transcript expression level.

Figure S13 | Comparison of different quantifying strategy between 2-cell and zygote. Transcript quantification methods include alignment-free transcript quantification-Salmon and alignment-based transcript quantification-Hisat2. Identifying methods of DE genes include DEseq2 and Seurat.

REFERENCES

Abe, K. I., Funaya, S., Tsukioka, D., Kawamura, M., Suzuki, Y., and Suzuki, M. G. (2018). Minor zygotic gene activation is essential for mouse preimplantation development. *Proc. Natl. Acad. Sci. U.S.A.* 115, E6780–E6788. doi: 10.1073/pnas.1804309115

Andrews, S. (2010). FASTQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Assou, S., Boumela, I., Haouzi, D., Anahory, T., Dechaud, H., and De Vos, J. (2011). Dynamic changes in gene expression during human early embryo development: from fundamental aspects to clinical applications. *Hum. Reprod. Update* 17, 272–290. doi: 10.1093/humupd/dmq036

Bao, J., Vitting-Seerup, K., Waage, J., Tang, C., Ge, Y., and Porse, B. T. (2016). UPF2-dependent nonsense-mediated mRNA decay pathway is essential for spermatogenesis by selectively eliminating longer 3′UTR transcripts. *PLoS Genet.* 12:e1005863. doi: 10.1371/journal.pgen.1005863

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. A Stat. Soc.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Busch, A., and Hertel, K. J. (2013). HEXEvent: a database of human exon splicing events. *Nucleic Acids Res.* 41, D118–D124. doi: 10.1093/nar/gks969

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

Calixto, C. P. G., Guo, W., James, A. B., Tzioutziou, N. A., Entizne, J. C., and Panter, P. E. (2018). Rapid and dynamic alternative splicing impacts the *Arabidopsis* cold response transcriptome. *Plant Cell* 30, 1424–1444. doi: 10.1105/tpc.18.00177

Carazo, F., Romero, J. P., and Rubio, A. (2018). Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors. *Brief Bioinform.* 20, 1358–1375. doi: 10.1093/bib/bby005

Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., and Tang, F. (2015). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* 16, 148. doi: 10.1186/s13059-015-0706-1

Feng, X., Zhang, J., Smuga-Otto, K., Tian, S., Yu, J., and Stewart, R. (2012). Protein kinase C mediated extraembryonic endoderm differentiation of human embryonic stem cells. *Stem Cells* 30, 461–470. doi: 10.1002/stem.1018

Freytag, S., Tian, L., Lönnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Res.* 7:1297. doi: 10.12688/f1000research.15809.2

García-López, J., Hourcade J. de D., and Del Mazo, J. (2013). Reprogramming of microRNAs by adenosine-to-inosine editing and the selective elimination of edited microRNA precursors in mouse oocytes and preimplantation embryos. *Nucleic Acids Res.* 41, 5483–5493. doi: 10.1093/nar/gkt247

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. *Nat. Rev. Genet.* 15, 829–845. doi: 10.1038/nrg3813

Goldstein, O., Meyer, K., Greenshpan, Y., Bujanover, N., Feigin, M., and Ner-Gaon, H. (2017). Mapping whole transcriptome splicing in mouse hematopoietic stem cells. *Stem Cell Rep.* 8, 163–176. doi: 10.1016/j.stemcr.2016.12.002

Guo, W., Calixto, C. P. G., Brown, J. W. S., and Zhang, R. (2017). TSIS: an R package to infer alternative splicing isoform switches for time series data. *Bioinformatics* 33, 3308–3310. doi: 10.1093/bioinformatics/btx411

Hamatani, T., Carter, M. G., Sharov, A. A., and Ko, M. S. (2004). Dynamics of global gene expression changes during mouse preimplantation development. *Dev. Cell* 6, 117–131. doi: 10.1016/S1534-5807(03)00373-3

Hamatani, T., Ko, M. Sh., Yamada, M., Kuji, N., Mizusawa, Y. SHOJI, M., et al. (2006). Global gene expression profiling of preimplantation embryos. *Hum. Cell* 19, 98–117. doi: 10.1111/j.1749-0774.2006.00018.x

Han, H., Irimia, M., Ross, P. J., Sung, H. K., Alipanahi, B., David, L., et al. (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* 498, 241–245. doi: 10.1038/nature12270

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., and Clark, L. (2002). The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41. doi: 10.1093/nar/30.1.38

Jumaa, H., Wei, G., and Nielsen, P. J. (1999). Blastocyst formation is blocked in mouse embryos lacking the splicing factor SRp20. *Curr. Biol.* 9, 899–902. doi: 10.1016/S0960-9822(99)80394-7

Kalsotra, A., and Cooper, T. A. (2011). Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* 12, 715–729. doi: 10.1038/nrg3052

Kanka, J. (2003). Gene expression and chromatin structure in the pre-implantation embryo. *Theriogenology* 59, 3–19. doi: 10.1016/S0093-691X(02)01267-0

Kornblihtt, A. R., Schor, I. E., Allo, M., and Blencowe, B. J. (2009). When chromatin meets splicing. *Nat. Struct. Mol. Biol.* 16, 902–903. doi: 10.1038/nsmb0909-902

Kovacs, E., Tompa, P., Liliom, K., and Kalmar, L. (2010). Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5429–5434. doi: 10.1073/pnas.0907841107

Liu, Y., Lu, X., Shi, J., Yu, X., Zhang, X., and Zhu, K. (2016). BTG4 is a key regulator for maternal mRNA clearance during mouse early embryogenesis. *J. Mol. Cell. Biol.* 8, 366–368. doi: 10.1093/jmcb/mjw023

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

Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. *Cell* 144, 16–26. doi: 10.1016/j.cell.2010.11.056

Ma, J., Fukuda, Y., and Schultz, R. M. (2015). Mobilization of dormant Cnot7 mRNA promotes deadenylation of maternal transcripts during mouse oocyte maturation. *Biol. Reprod.* 93:48. doi: 10.1095/biolreprod.115.130344

McGlincy, N. J., and Smith, C. W. (2008). Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem. Sci.* 33, 385–393. doi: 10.1016/j.tibs.2008.06.001

Nguyen, L. S., Wilkinson, M. F., and Gecz, J. (2014). Nonsense-mediated mRNA decay: inter-individual variability and human disease. *Neurosci. Biobehav. Rev.* 46, 175–186. doi: 10.1016/j.neubiorev.2013.10.016

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.* 102, 11–26. doi: 10.1016/j.ajhg.2017.11.002

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., et al. (2016). Single-Cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 165, 1012–1026. doi: 10.1016/j.cell.2016.03.023

Qiu, S., Li, W., Xiong, H., Liu, D., Bai, Y., Wu, K., et al. (2016). Single-cell RNA sequencing reveals dynamic changes in A-to-I RNA editome during early human embryogenesis. *BMC Genomics* 17:766. doi: 10.1186/s12864-016-3115-2

Revil, T., Gaffney, D., Dias, C., Majewski, J., and Jerome-Majewska, L. A. (2010). Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics* 11:399. doi: 10.1186/1471-2164-11-399

Roy, S. W., and Penny, D. (2007). Intron length distributions and gene prediction. *Nucleic Acids Res.* 35, 4737–4742. doi: 10.1093/nar/gkm281

Schultz, R. M., Stein, P., and Svoboda, P. (2018). The oocyte-to-embryo transition in mouse: past, present, and future. *Biol. Reprod.* 99, 160–174. doi: 10.1093/biolre/ioy013

Schweingruber, C., Rufener, S. C., Zünd, D., Yamashita, A., and Mühlemann, O. (2013). Nonsense-mediated mRNA decay–mechanisms of substrate mRNA recognition and degradation in mammalian cells. *Biochim. Biophys. Acta.* 1829, 612–623. doi: 10.1016/j.bbagrm.2013.02.005

Scotti, M. M., and Swanson, M. S. (2016). RNA mis-splicing in disease. *Nat. Rev. Genet.* 17, 19–32. doi: 10.1038/nrg.2015.3

Sen, S., Jumaa, H., and Webster, N. J. (2013). Splicing factor SRSF3 is crucial for hepatocyte differentiation and metabolic function. *Nat. Commun.* 4:1336. doi: 10.1038/ncomms2342

Shen, S., Park, J. W., Lu, Z., Lin, L., Henry, M. D., Wu, Y. N., et al. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5593–E5601. doi: 10.1073/pnas.1419161111

Singh, R. K., and Cooper, T. A. (2012). Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.* 18, 472–482. doi: 10.1016/j.molmed.2012.06.006

Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E., Bult, C. J., and Mouse Genome Database Group. (2018). Mouse genome database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res.* 46, D836–D842. doi: 10.1093/nar/gkx1006

Solomon, O., Oren, S., Safran, M., Deshet-Unger, N., Akiva, P., and Jacob-Hirsch, J. (2013). Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR). *RNA* 19, 591–604. doi: 10.1261/rna.038042.112

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* 4:1521. doi: 10.12688/f1000research.7563.1

The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Thoren, L. A., Nørgaard, G. A., Weischenfeldt, J., Waage, J., Jakobsen, J. S., and Damgaard, I. (2010). UPF2 is a critical regulator of liver development, function and regeneration. *PLoS ONE* 5:e11650. doi: 10.1371/journal.pone.0011650

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., and Elliott, D. J. (2018). SUPPA2: fast, accurate, and uncertainty aware differential splicing analysis across multiple conditions. *Genome Biol.* 19:40. doi: 10.1186/s13059-018-1417-1

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., González-Vallinas, J., Lahens, N. F., Hogenesch, J. B., et al. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* 5:e11752. doi: 10.7554/eLife.11752

Wang, Y., Liu, J., Huang, B. O., Xu, Y. M., Li, J., and Huang, L. F. (2015). Mechanism of alternative splicing and its regulation. *Biomed. Rep.* 3, 152–158. doi: 10.3892/br.2014.407

Wilding, M., Coppola, G., Dale, B., and Di Matteo, L. (2009). Mitochondria and human preimplantation embryo development. *Reproduction* 137, 619–624. doi: 10.1530/REP-08-0444

Wrzesinski, T., Szelag, M., Cieślikowski, W. A., Ida, A., Giles, R., Zodro, E., et al. (2015). Expression of pre-selected TMEMs with predicted ER localization as potential classifiers of ccRCC tumors. *BMC Cancer* 15:518. doi: 10.1186/s12885-015-1530-4

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., et al. (2015). RNA splicing. the human splicing code reveals new insights into the genetic determinants of disease. *Science* 347:1254806. doi: 10.1126/science.1254806

Xu, Q., Xiang, Y., Wang, Q., Wang, L., Brind'Amour, J., and Bogutz, A. B. (2019). SETD2 regulates the maternal epigenome, genomic imprinting and embryonic development. *Nat. Genet.* 51, 844-856. doi: 10.1038/s41588-019-0398-7

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131–1139. doi: 10.1038/nsmb.2660

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

Zeng, F., and Schultz, R. M. (2005). RNA transcript profiling during zygotic gene activation in the preimplantation mouse embryo. *Dev. Biol.* 283, 40–57. doi: 10.1016/j.ydbio.2005.03.038

Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* 3, 698–709. doi: 10.1038/nrg890

Zhang, Y., You, J., Wang, X., and Weber, J. (2015). The DHX33 RNA helicase promotes mRNA translation initiation. *Mol. Cell. Biol.* 35, 2918–2931. doi: 10.1128/MCB.00315-15

Zhou, H. L., Luo, G., Wise, J. A., and Lou, H. (2014). Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Res.* 42, 701–713. doi: 10.1093/nar/gkt875

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 © 2020 Xing, Yang, Liu, Cui, Meng, Zhao, Zhao, Li, Liu, Zhang 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 Genetics

METHODS
published: 07 February 2020
doi: 10.3389/fgene.2019.01371

Phylogenetic Tree Inference: A Top-Down Approach to Track Tumor Evolution

Pin Wu[1,2], *Linjun Hou*[1,2], *Yingdong Zhang*[3] *and Liye Zhang*[1]*

[1] School of Life Science and Technology, ShanghaiTech University, Shanghai, China, [2] University of Chinese Academy of Sciences, Beijing, China, [3] Library and Information Center, ShanghaiTech University, Shanghai, China

Recently, an increasing number of studies sequence multiple biopsies of primary tumors, and even paired metastatic tumors to understand heterogeneity and the evolutionary trajectory of cancer progression. Although several algorithms are available to infer the phylogeny, most tools rely on accurate measurements of mutation allele frequencies from deep sequencing, which is often hard to achieve for clinical samples (especially FFPE samples). In this study, we present a novel and easy-to-use method, PTI (Phylogenetic Tree Inference), which use an iterative top-down approach to infer the phylogenetic tree structure of multiple tumor biopsies from same patient using just the presence or absence of somatic mutations without their allele frequencies. Therefore PTI can be used in a wide range of cases even when allele frequency data is not available. Comparison with existing state-of-the-art methods, such as LICHeE, Treeomics, and BAMSE, shows that PTI achieves similar or slightly better performance within a short run time. Moreover, this method is generally applicable to infer phylogeny for any other data sets (such as epigenetics) with a similar zero and one feature-by-sample matrix.

Keywords: phylogenetics, tumor evolution, multi-region sequencing, lineage tracing, allele frequency

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Xiaoqi Zheng,
Shanghai Normal University, China
Zhenguo Lin,
Saint Louis University, United States

***Correspondence:**
Liye Zhang
zhangly@shanghaitech.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *29 September 2019*
Accepted: *16 December 2019*
Published: *07 February 2020*

Citation:
Wu P, Hou L, Zhang Y and Zhang L (2020) Phylogenetic Tree Inference: A Top-Down Approach to Track Tumor Evolution. Front. Genet. 10:1371. doi: 10.3389/fgene.2019.01371

INTRODUCTION

Cancer is an evolutionary process that is shaped by selection pressure and the accumulation of somatic mutations, resulting in a high level of heterogeneity within and between tumor samples (Marusyk et al., 2012; Yates and Campbell, 2012). Such heterogeneity in genomes can be used to distinguish tumor subclonal populations and track the evolutionary trajectory of cancer progression. Metastasis is normally considered as the last step during cancer progression and is still the major cause of cancer death but poorly understood mechanistically. A number of studies sequenced multiple biopsies of primary and metastatic tumors to elucidate the order of mutation accumulation and the origin of distal metastasis (Gundem et al., 2015; Yates et al., 2017; Ferronika et al., 2019). A better understanding of metastasis process may eventually lead to novel diagnosis and treatment strategies.

A number of computational methods are available to infer the genotypes of tumor cell populations. However, most existing methods infer the phylogeny of cancer evolution based on somatic mutations variant allele frequencies (VAFs) from DNA deep sequencing data (Jiao et al., 2014; Malikic et al., 2015; Yates et al., 2015; Nieboer et al., 2018). Several traditional phylogenetic inference methods

utilize multiple sequence alignments, neighbor joining with Pearson correlation distances, maximum parsimony algorithm or maximum likelihood algorithm based on variant presence patterns across samples (Kim et al., 2015; Lu et al., 2016; Zhao et al., 2016; Choi et al., 2017; Naxerova et al., 2017; Zhai et al., 2017). Most of these methods are computationally intensive and require a long running time. In 2015, LICHeE was developed to construct multi-sample tumor phylogenetic trees and tumor subclonal decomposition from accurate VAFs of somatic single nucleotide variants (SSNVs) obtained by deep sequencing (Popic et al., 2015). LICHeE first groups subsets of somatic mutations that have similar presence-absence patterns as well as similar VAFs across multiple tumor samples. Then, it constructs a constrained network to infer the relationships among clusters of somatic mutations and identify tumor phylogenetic trees. Several other methods adopt similar principles but different methodological frameworks, such as Treeomics and BAMSE (Reiter et al., 2017; Toosi et al., 2019). Treeomics was developed to reconstruct the phylogeny of metastases and map subclones to their anatomic locations. It uses total reads and variant reads of SSNVs from multiple related normal and tumor samples of individual cancer patients as input files. Then it uses a Bayesian inference model to identify evolutionarily compatible mutation patterns and then infer evolutionary trees. Another probabilistic method, named BAMSE infers subclonal history and lineage tree reconstruction of heterogeneous tumor samples using somatic mutations read counts as input. The posterior probability of tree is inferred by a Bayesian model that integrates prior belief about the number of subclones, the composition of the tumor, and the process of subclonal evolution. However, users have to decide the number of subclones, which is normally difficult to estimate. There are two major issues common for these methods. Most importantly, it is often difficult to obtain accurate allele frequency from clinical samples, such as formalin-fixed, paraffin-embedded (FFPE) samples (Astolfi et al., 2015). Results of these methods are also sensitive to several key parameters, and yet there is no easy way for users to decide on these parameters.

Here, we propose PTI (Phylogenetic Tree Inference), a novel method which use an iterative top-down approach to infer the phylogenetic tree structure of multiple tumor biopsies from same patient using somatic mutations without the needs of accurate allele frequencies. In addition, PTI has only one parameter to set, and we also provide clear instructions on how to set this parameter.

METHODS

PTI is a method designed to use an iterative top-down approach to infer the rooted phylogenetic tree among multiple samples of the same patient. In this section, we provide an overview of our approach (**Figure 1**). First, PTI identifies shared mutations for all samples and defines the number of shared mutations as the length of the root trunk. Then, PTI uses an iterative top-down approach to find the optimal branch split until all samples reach the leaf nodes. PTI also annotates the mutations of known driver genes on the tree structure, which facilitates an intuitive understanding of the key mutation events during cancer progression.

Identify and Remove the Shared Mutations From All Samples

Assume we obtained multiple samples s, $s \in \{1,2,\ldots,n\}$ from same patient. Using somatic mutation caller, such as Mutect2 or VarScan2 (Koboldt et al., 2012; Cibulskis et al., 2013), we can detect the mutation allele frequencies in each sample. Define these somatic mutations by $r_1,r_2\ldots,r_m$. We then build a binary matrix M with rows labeled $r_1,r_2\ldots,r_m$, and columns labeled $s_1,s_2\ldots,s_n$, such that $M_{ij} = 1$ if and only if the VAF of somatic mutation r_i in sample s_j is greater or equal to a given threshold. The more high-confidence mutations are used to construct a phylogenetic tree, the more accurate the structure of the tree. However, when the number of somatic mutations within a patient is too large, an optional filtering step of somatic mutations based on allele frequency (default is 0.1) can be implemented by PTI. Matrix M is the input of PTI. We calculate the intersection of mutations in all samples of the same patient and define the intersection as the length of the root trunk. After removing the shared mutations from all samples, the next step is to find the optimal branch split for the filtered data set M_{filter}.

Find the Optimal Branch Split

Genomic variations in cancer cells gradually accumulate over the course of carcinogenesis and cancer development (Gerlinger et al., 2012; Sato et al., 2016). So despite the complexity in cancer evolution, more than two ways split is rarely observed on the evolutionary tree in existing studies (Hong et al., 2015; Schwarz et al., 2015; Brown et al., 2017). Therefore, PTI will iterate through all possible two-way branch splits, $S = \{(1, n-1), (2, n-2), \ldots(|\frac{n}{2}|, n-|\frac{n}{2}|)\}$ to infer the optimal branch split. Notably, our method indeed is able to detect more than two ways split at any given evolutionary node (**Supplementary Note 1**).

For each possible branch split S_t, $t \in \{1, 2, \ldots, |\frac{n}{2}|\}$, C_n^t combinations are included. Let θ be an object of the numbers of shared mutations measured on all possible branch splits. For each possible branch split S_t and combination c, $c \in \{1, 2, \ldots, C_n^t\}$ the corresponding element θ_{tc}, represents the number of shared mutations in the larger group. If n is even, there is no larger group of the possible branch split $(\frac{n}{2}, n-\frac{n}{2})$. Then in this case, $\theta_{\frac{n}{2}c}$ presenting the smaller number of shared mutations in two equal size groups.

In order to determine which possible branch split is the best, we define ∂ to be an vector of ratio measured on all possible branch splits which is calculated via the Equation (1), where $\theta_{\mathrm{t\ max}}$ represents the maximum value and $\theta_{\mathrm{t\ sec_max}}$ represents the secondary maximum value in θ_t:

$$\partial_t = \frac{\theta_{\mathrm{t\,max}}}{\theta_{\mathrm{t\,sec_max}}} \quad (1)$$

If the optimal split occurs in S_t, then the ratio between the best combination and second best combination should be much larger compared with non-optimal splits (see **Supplementary Note 2** for details description of rationale of using this ratio).

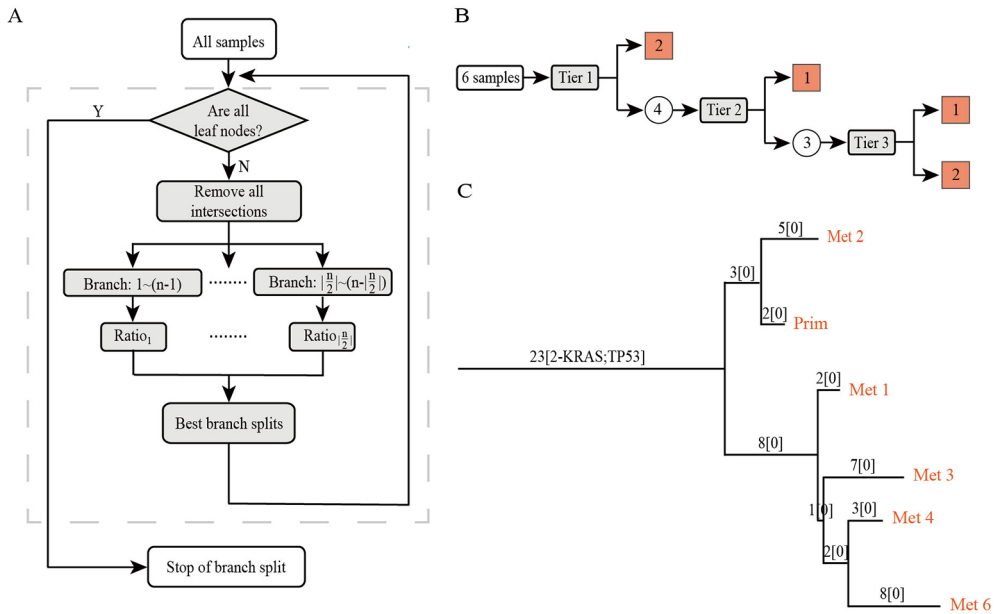


FIGURE 1 | The overview of PTI. **(A)** The workflow of PTI. **(B)** Taking the patient with six samples as an example, the specific process of PTI inferring the phylogenetic tree is shown in detail. The square grid indicates that the leaf node has been reached, and the circle indicates that the remaining samples still have to be iterated to find the best branch split. **(C)** An example of a phylogenetic tree inferred by PTI. The phylogenetic tree is a rooted tree whose leaf nodes are samples. The annotation information on the branches includes the length of the branches, which is equal to the number of shared mutations, and the annotation of driver genes.

Then, the samples that reach the leaf node after the optimal split will be removed from the data set M_{filter}. This method will iterate over the rest of the samples until all samples are split into leaf nodes. For a patient, there may be more than one tree structures with an equal ∂ value. To determine the optimal tree structure, an aggregated mutation count (W_T) is calculated for each tree structure using mutations on all trunks that contain two or more leaf nodes. The tree with the largest scores will be the optimal tree. Let i, $i \in \{0,1,\ldots,k\}$, represents all trunk levels in each tree structure. In trunk level i, there are N_i trunks so that we let j, $j \in \{1, \ldots, N_i\}$, represent all trunks in each trunk level. We also define ω_{ij} as the length of trunk j in trunk level i and define χ_{ij} which represents the number of leaf nodes involved in trunk j in trunk level i. Then, the weight score W_T of tree structure T will be calculated by:

$$W_T = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \omega_{ij} \chi_{ij} \tag{2}$$

without root trunk level ($i = 0$) because the tree structures of same patient have same root trunk.

Annotation of Driver Mutations on Phylogenetic Tree

It is well known that there are more passenger mutations than driver mutations in cancer genome. Understanding the time of occurrence and the distribution of driver mutations in different samples is important to understand the evolution of tumor progression. Therefore, our method also annotates the putative 299 driver genes on the branches of the tree for downstream analysis (Bailey et al., 2018). It should be noted that in the tree structure, there may be more than one tree branch with annotation information of the same driver gene. This may be caused by the same mutation or different mutations of the same driver gene, which can be answered by an auxiliary information file corresponding to the tree structure file.

As this method assume there is a major single clone for each sample due to multiple biopsies, we do notice that in rare cases this method will output multiple solutions instead of one optimal solution when some samples are consisted of more than one major clones (**Supplementary Note 3**).

RESULTS

Results on High-Grade Serous Ovarian Cancer

To evaluate our method, we compared the performance of PTI with two state-of-the-arts methods, LICHeE and Treeomics on high-grade serous ovarian cancer (HGSC) data set which were obtained from European Genome-Phenome Archive (accession EGAS00001000547) (Bashashati et al., 2013). PTI used all mutations with AF >= 0.01 while LICHeE and Treeomics were run with the parameter defined in their published paper. Then we compared the results of these three methods with the results given in the original literature based on mutations and copy

number alterations. In order to evaluate the similarity of two tree structures, we defined a tree structure similarity scoring system. The similarity score represents the proportion of the identical paths in the tree topology and ranges from zero to one (**Supplementary Note 4**). PTI showed slightly better performance compared with LICHeE and Treeomics. Only PTI correctly predicted identical structure in case 4, where sample j and f-i are grouped into one branch (**Table 1** and **Figure 2A**). 4 of 6 results predicted by PTI showed the identical structure which similarity score is equal to 1. None of the three methods showed highly consistent structures in Case 1 and 5 as original results (**Table 1** and **Figure 2B**). In Case 1, results from PTI and LICHeE were highly consistent. For Case 5, when PTI use all somatic mutations with AF >= 0.01, there are only minor differences between the tree structures inferred by PTI and Treeomics. However, if PTI use all single nucleotide somatic mutations obtained from original paper including 8 tumor samples of Case 5, in contrast to the results in the original literature which suggested an early divergence of sample c, PTI first separated the sample h to achieve the most common mutations (n = 4) in the remaining samples (**Supplementary Note 5**). Careful analysis of the somatic mutation data set revealed that if the sample c is diverged first, there is only one shared somatic mutation in the remaining samples.

TABLE 1 | The comparison of PTI, LICHeE, and Treeomics based on HGSC data set.

Patient_ID	PTI (AF >= 0.01)	LICHeE	Treeomics
Case 1	0.14	0.15	0.43
Case 2	1	1	1
Case 3	1	1	1
Case 4	1	0.57	0.83
Case 5	0.27	0.23	0.34
Case 6	1	1	1

Results on Clear Cell Renal Carcinomas Data Set

We performed a separate comparison between PTI, LICHeE, and BAMSE on clear cell renal carcinomas (ccRCC) data set from eight individuals which were obtained from European Genome-Phenome Archive (accession EGAS00001000667) (Gerlinger et al., 2014). Since LICHeE only used the variant allele frequency of somatic single nucleotide variants to reconstruct

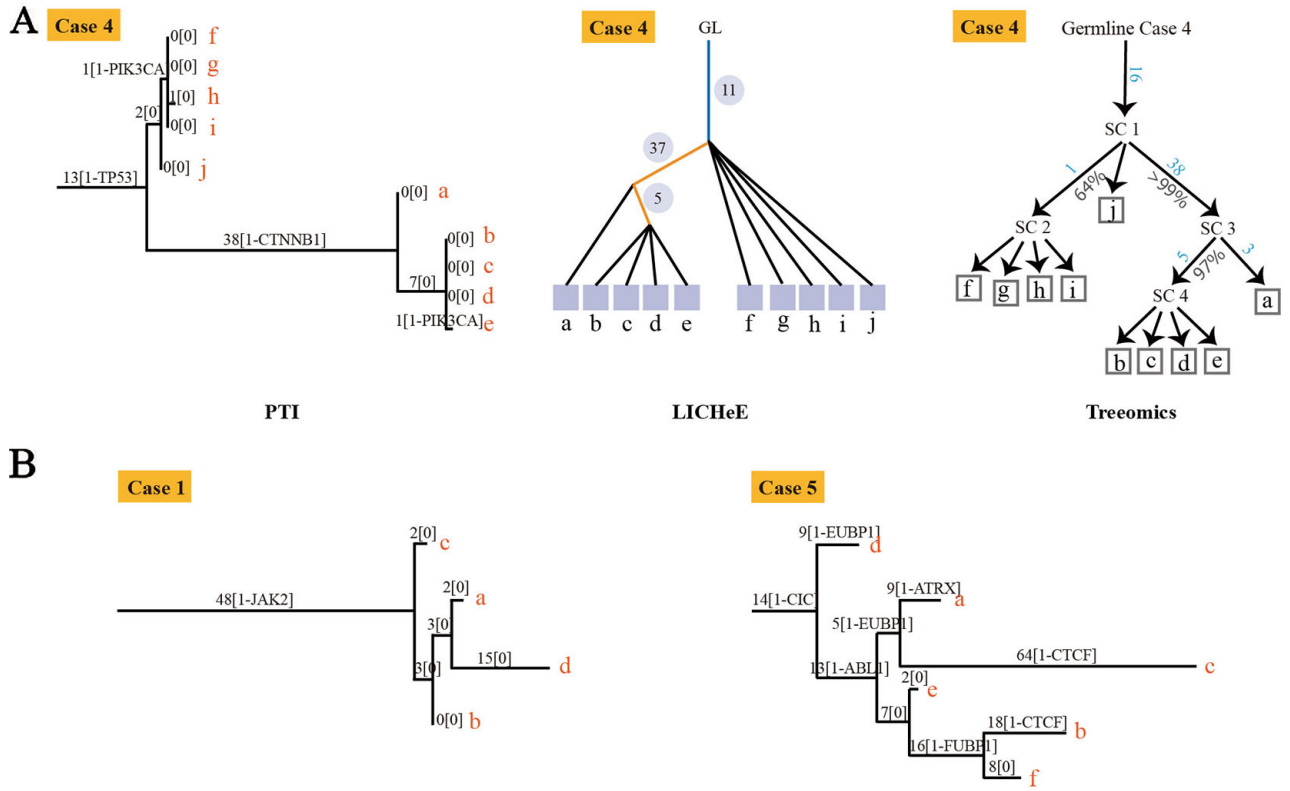


FIGURE 2 | The tree structures inferred by PTI for HGSC Cases 1, 4 and 5. **(A)** The tree structure of Case 4 inferred by PTI, LICHeE, and Treeomics respectively. In the LICHeE result: yellow line, internal branch; light blue line, trunk; black line, contribution link; the numbers inside the circle, SSNVs; the light purple square, tumor region. In the Treeomics result: numbers in blue correspond to the acquired variants in the branches. Percentages (gray) denote bootstrap values (1,000 samples). SC, subclone. **(B)** For Case 1, PTI first separates the sample c to achieve the best branch split, which is same as the result of LICHeE. For Case 5, the tree structure inferred by PTI and Treeomics has high similarities using six samples and eight samples (**Supplementary Note 5**).

TABLE 2 | The comparison of PTI and other methods on ccRCC data set.

Patient_ID	PTI (AF >= 0.01)	LICHeE	BAMSE
EV003	0.67	0.48	/
EV005	1	0.70	0.24
EV006	0.89	0.17	0.1
EV007	0.86	0.85	0.33
RMH002	0.55	1	0.55
RMH004	0.5	0.57	0.5
RMH008	1	1	0.83
RK26	0.65	0.54	/

the phylogenetic process, PTI took the same set of SNVs with AF >= 0.01 as the input data set. Since this data set lacked information about total reads and variant reads of mutation (Treeomics needs such information), so we only compared results of PTI, LICHeE, BAMSE with those from original literature that used VAF-based clustering, variant presence pattern and maximum parsimony algorithm (Gerlinger et al., 2014). It should be noted that the trees inferred by BAMSE were obtained from Toosi et al. (2019). The comparison shows that PTI and LICHeE performed similarly in terms of accuracy and speed in the ccRCC data set while all tree structures inferred by BAMSE had single-branch or multi-branch differences (**Table 2**, **Figure S7**).

Results on Breast Cancer Data Set

PTI also was benchmarked with LICHeE and Treeomics on breast cancer data set and then compared these results with the results from original literature based on both somatic mutations and copy number alterations. Breast cancer data set were obtained from European Genome-Phenome Archive (accession EGAS00001000760) (Brown et al., 2017). The running time of PTI, as well as the other two methods, was short, just within seconds (**Table S5**). PTI and Treeomics both showed higher accuracy compared with LICHeE. In results predicted by PTI, 6 out of 8 patients showed identical structures, while the other two patients P1 and P2 showed highly similar tree structures as the results in Brown et al. (2017), which may be caused by more than one subclonal population in one biopsy (**Figure 3**). For example, in patient P1, sample M1 (Metastatic tumor sample) includes A-clone and B-clone, sample M4 (Metastatic tumor sample) contains only A-clone, and samples M3 and P contain B-clone. Therefore, the sample M1 is grouped together with the sample M4 or with the samples P-M3, which is determined by

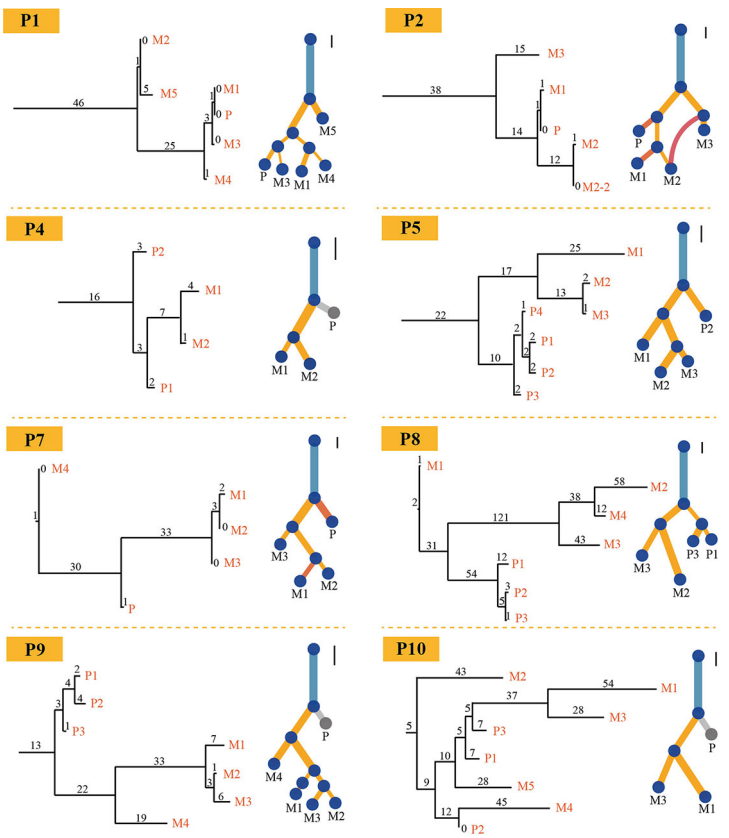


FIGURE 3 | Comparison of trees for eight breast cancer patients. Eight phylogenetic tree structures without annotation information about the driver gene inferred by PTI using all somatic mutations with >=1500X coverage and >=3% VAF (on the left) are compared with the trees (on the right) published in Brown et al. (2017) for each patient on breast cancer data set. As for PTI results, the number of shared mutations is proportional to the length of branch and labeled above each branch. Also, the scale bars on the upper right corner of the results given in original published paper represent 10 SNVs and provide an indication of the original length of the trees.

the proportion of somatic mutations involved in A-clone and B-clone in sample P (**Supplementary Note 6**). Treeomics also showed good performance, and 6 out of 8 results were identical. However, in the LICHeE results, 5 out of 8 patients showed single-branch or multi-branch differences in tree structures. We also tested these two methods on different AF cutoffs. The accuracy of the tree structures of the three methods was slightly improved, but PTI still showed better performance (**Table S6**). Moreover, we also demonstrated that PTI performed robustly in the low coverage data set by applying it to the HISEQ data set for 8 breast cancer patients (**Table S7**).

Results on 13 Cancer Types Data Set

We also ran PTI on a real data set including somatic single nucleotide mutations from 40 patients of 13 cancer types where allele frequency data is not available. This data set were obtained from BioStudies database (accession S-EPMC4776530) (Zhao et al., 2016). We applied PTI to infer the tree structure and then compared our results with the results that based on the multiple sequence alignments, maximum likelihood algorithm, the maximal parsimony algorithm, and the Bayesian inference criteria implemented in original study. And then, based on similarity score, we categorized the comparison results into four groups: similarity score = 1, similarity score$\in$ [0.5,1), similarity score $\in$ [0.2,0.5) and similarity score $\in$[0,0.2), representing various degrees of tree structure similarity. The comparison shows that 92.5% of our results have the same or similar tree structure (similarity score greater than 0.2) as the results in Zhao et al. (2016) (**Figure 4**, **Figure S8**), which again suggest that our method can be applicable in a wide range of applications.

DISCUSSION

As PTI assumes that there is one major clone in each biopsy, when there is more than one major subclone, PTI will assign the sample to the subclone with a higher mutation count by not their relative cellular abundances of two subclones (**Figure 3** and **Table S4**). This may lead to some discrepancies in the tree structures compared with other methods. But this case is rarely observed in the studies for multi-region sequencing, we only observe one case in all cases we tested.

In this study, we present PTI, a novel and easy-to-use method to infer the phylogenetic tree of tumor progression using just somatic mutations without the need for deep sequencing to obtain high-confident allele frequency measurement. Our comparison to other existing methods, such as LICHeE, Treeomics, BAMSE, and other traditional methods, shows that PTI achieves similar or slightly better performance within a short run time, normally less than a minute. This feature is important for studying clinical samples that are difficult to obtain accurate allele frequency information, such as formalin-fixed, paraffin-embedded (FFPE) samples. Moreover, the input file for PTI is a similar zero and one feature-by-sample matrix so that this method is generally applicable to infer phylogeny for any other data sets that can be converted into this format (such as epigenetics). In fact, this method is also well suited for single cell data sets to evaluate the similarity between single cells and construct their phylogenies.

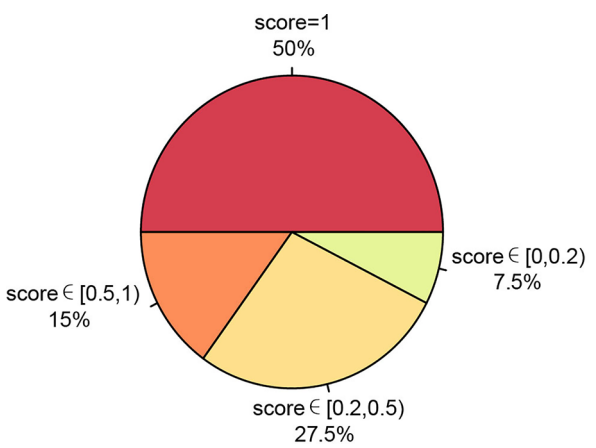


FIGURE 4 | Summary of our method's performance on 13 cancer types data set without allele frequency. The consistency between PTI and original results based on similarity scores were represented as a pie chart.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The European Genome-phenome Archive (EGAS00001000547, EGAS00001000667, EGAS00001000760) and Biostudies (S-EPMC4776530).

AUTHOR CONTRIBUTIONS

PW and LZ designed the algorithm and interpreted the results. PW implemented the algorithm and conducted the analyses. LH assisted in obtaining and preprocessing the data sets. YZ assisted in setting up the software environment for performance comparison.

FUNDING

This work is funded by the National Key Research and Development Program of China (2018YFC1004602), National Natural Science Foundation of China (NSF 31871332) and a startup fund to L.Z. from ShanghaiTech University.

ACKNOWLEDGMENTS

We would like to thank Xiuqi Pan for data repeatability testing. We would like to thank the support from the HPC platform of ShanghaiTech University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01371/full#supplementary-material

REFERENCES

Astolfi, A., Urbini, M., Indio, V., Nannini, M., Genovese, C. G., Santini, D., et al. (2015). Whole exome sequencing (WES) on formalin-fixed, paraffin-embedded (FFPE) tumor tissue in gastrointestinal stromal tumors (GIST). *BMC Genomics* 16, 892. doi: 10.1186/s12864-015-1982-6

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385 e318. doi: 10.1016/j.cell.2018.02.060

Bashashati, A., Ha, G., Tone, A., Ding, J., Prentice, L. M., Roth, A., et al. (2013). Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J. Pathol.* 231, 21–34. doi: 10.1002/path.4230

Brown, D., Smeets, D., Szekely, B., Larsimont, D., Szasz, A. M., Adnet, P. Y., et al. (2017). Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nat. Commun.* 8, 14944. doi: 10.1038/ncomms14944

Choi, Y. J., Rhee, J. K., Hur, S. Y., Kim, M. S., Lee, S. H., Chung, Y. J., et al. (2017). Intraindividual genomic heterogeneity of high-grade serous carcinoma of the ovary and clinical utility of ascitic cancer cells for mutation profiling. *J. Pathol.* 241, 57–66. doi: 10.1002/path.4819

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219. doi: 10.1038/nbt.2514

Ferronika, P., Hof, J., Kats-Ugurlu, G., Sijmons, R. H., Terpstra, M. M., De Lange, K., et al. (2019). Comprehensive profiling of primary and metastatic ccRCC reveals a high homology of the metastases to a subregion of the primary tumour. *Cancers* 11, 812. doi: 10.3390/cancers11060812

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New Engl. J. Med.* 366, 883–892. doi: 10.1056/NEJMoa1113205

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., et al. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* 46, 225–233. doi: 10.1038/ng.2891

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M. C., Papaemmanuil, E., et al. (2015). The evolutionary history of lethal metastatic prostate cancer. *Nature* 520, 353–357. doi: 10.1038/nature14347

Hong, M. K. H., Macintyre, G., Wedge, D. C., Van Loo, P., Patel, K., Lunke, S., et al. (2015). Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat. Commun.* 6, 6605. doi: 10.1038/ncomms7605

Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinf.* 15, 35. doi: 10.1186/1471-2105-15-35

Kim, T. M., Jung, S. H., An, C. H., Lee, S. H., Baek, I. P., Kim, M. S., et al. (2015). Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clin. Cancer Res.* 21, 4461–4472. doi: 10.1158/1078-0432.CCR-14-2413

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., Mclellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Lu, Y. W., Zhang, H. F., Liang, R., Xie, Z. R., Luo, H. Y., Zeng, Y. J., et al. (2016). Colorectal cancer genetic heterogeneity delineated by multi-region sequencing. *PloS One* 11, e0152673. doi: 10.1371/journal.pone.0152673

Malikic, S., Mcpherson, A. W., Donmez, N., and Sahinalp, C. S. (2015). Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* 31, 1349–1356. doi: 10.1093/bioinformatics/btv003

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* 12, 323–334. doi: 10.1038/nrc3261

Naxerova, K., Reiter, J. G., Brachtel, E., Lennerz, J. K., Van De Wetering, M., Rowan, A., et al. (2017). Origins of lymphatic and distant metastases in human colorectal cancer. *Science* 357, 55–60. doi: 10.1126/science.aai8515

Nieboer, M. M., Dorssers, L. C. J., Straver, R., Looijenga, L. H. J., and De Ridder, J. (2018). TargetClone: a multi-sample approach for reconstructing subclonal evolution of tumors. *PloS One* 13, e0208002. doi: 10.1371/journal.pone.0208002

Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R. B., and Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.* 16, 91. doi: 10.1186/s13059-015-0647-8

Reiter, J. G., Makohon-Moore, A. P., Gerold, J. M., Bozic, I., Chatterjee, K., Iacobuzio-Donahue, C. A., et al. (2017). Reconstructing metastatic seeding patterns of human cancers. *Nat. Commun.* 8, 14114. doi: 10.1038/ncomms14114

Sato, F., Saji, S., and Toi, M. (2016). Genomic tumor evolution of breast cancer. *Breast Cancer* 23, 4–11. doi: 10.1007/s12282-015-0617-8

Schwarz, R. F., Ng, C. K. Y., Cooke, S. L., Newman, S., Temple, J., Piskorz, A. M., et al. (2015). Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PloS Med.* 12, e1001789. doi: 10.1371/journal.pmed.1001789

Toosi, H., Moeini, A., and Hajirasouliha, I. (2019). BAMSE: Bayesian model selection for tumor phylogeny inference among multiple samples. *BMC Bioinf.* 20, 282. doi: 10.1186/s12859-019-2824-3

Yates, L. R., and Campbell, P. J. (2012). Evolution of the cancer genome. *Nat. Rev. Genet.* 13, 795–806. doi: 10.1038/nrg3317

Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., et al. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* 21, 751–759. doi: 10.1038/nm.3886

Yates, L. R., Knappskog, S., Wedge, D., Farmery, J. H. R., Gonzalez, S., Martincorena, I., et al. (2017). Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* 32, 169–184 e167. doi: 10.1016/j.ccell.2017.07.005

Zhai, W., Lim, T. K., Zhang, T., Phang, S. T., Tiang, Z., Guan, P., et al. (2017). The spatial organization of intra-tumour heterogeneity and evolutionary trajectories of metastases in hepatocellular carcinoma. *Nat. Commun.* 8, 4565. doi: 10.1038/ncomms14565

Zhao, Z. M., Zhao, B., Bai, Y., Iamarino, A., Gaffney, S. G., Schlessinger, J., et al. (2016). Early and multiple origins of metastatic lineages within primary tumors. *Proc. Natl. Acad. Sci. U. S. A.* 113, 2140–2145. doi: 10.1073/pnas.1525677113

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 © 2020 Wu, Hou, Zhang 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: 14 February 2020
doi: 10.3389/fcell.2020.00075

Nc2Eye: A Curated ncRNAomics Knowledgebase for Bridging Basic and Clinical Research in Eye Diseases

Yan Zhang†, Zhengbo Xue†, Fangjie Guo†, Fulong Yu, Liangde Xu and Hao Chen**

School of Biomedical Engineering, School of Ophthalmology and Optometry and Eye Hospital, Wenzhou Medical University, Wenzhou, China

Eye diseases (EDs) represent a group of disorders affecting the visual system, most of which can lead to visual impairment and blindness. Accumulating evidence reveals that non-coding RNAs (ncRNAs) are closely associated with a wide variety of EDs. However, abundant associations between ncRNAs and EDs are scattered across the published literature, obstructing a global view of ncRNA-ED associations. A public resource of high-quality manually curated ncRNAomics knowledge associated with EDs remains unavailable. To address this gap, we thus developed Nc2Eye (http://nc2eye.bio-data.cn/), which is the first knowledgebase dedicated to providing a comprehensive ncRNAomics resource for bridging basic and clinical research in EDs. Through a comprehensive review of more than 2400 published papers, Nc2Eye catalogs 7088 manually curated ncRNA-ED associations involving 4363 ncRNAs across eight species. We also provide detailed descriptions and annotation information for each ncRNA-disease association such as ncRNA categories, experimental methods, expression pattern and related clinical drugs. To further expand the pathogenic ncRNAs, we also collected more than 90 high-throughput EDs-related transcriptome datasets. Furthermore, a user-friendly interface was constructed for convenient and flexible data browsing, querying, and retrieving. We believe that Nc2Eye is a timely and valuable knowledgebase for significantly improving and useful for discovery of new diagnostic and therapeutic biomarkers.

Keywords: eye diseases, non-coding RNAs, knowledgebase, epigenomics, website

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Xing Chen,
China University of Mining and Technology, China
Hongwei Wang,
Sun Yat-sen University, China

****Correspondence:***
Liangde Xu
xuld@eye.ac.cn
Hao Chen
chenhao@mail.eye.ac.cn

†These authors share first authorship

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

Received: *16 December 2019*
Accepted: *29 January 2020*
Published: *14 February 2020*

Citation:
Zhang Y, Xue Z, Guo F, Yu F, Xu L and Chen H (2020) Nc2Eye: A Curated ncRNAomics Knowledgebase for Bridging Basic and Clinical Research in Eye Diseases. Front. Cell Dev. Biol. 8:75. doi: 10.3389/fcell.2020.00075

INTRODUCTION

Eye diseases (EDs) represent a group of disorders affecting the visual system. It was estimated that 217 million people (3.1% of the world population) suffered from moderate to severe visual disturbance in the world in 2015, while 36 million were blind (Bourne et al., 2017). There are more than 150 kinds of EDs recorded in medical subject headings (MeSH) (Bhattacharya et al., 2011), most of which can lead to visual impairment and blindness, such as cataract, age-related macular degeneration (AMD), diabetic retinopathy (DR), retinoblastoma (RB), and degenerative myopia, etc. The pathogenesis of EDs at the molecular level remains poorly understood.

Non-coding RNAs (ncRNAs), such as microRNA (miRNA), long non-codingRNA (lncRNA) and circlar RNA (circRNA), are a large category of functional RNA molecules involved in regulating many biological processes (Esteller, 2011). Numerous studies demonstrated that ncRNAs play important roles in various EDs (Li et al., 2016; Rassi et al., 2017; Zhang et al., 2017). For example, the lncRNA *PVT1* is up-regulated in RB tissues and silencing *PVT1* can suppress the tumor growth (Wu et al., 2019). Overexpression of circRNA *cPWWP2A* was reported to alleviate retinal vascular dysfunction in DR by inhibiting *miR-579* activity (Liu et al., 2019). *miR-27a* is up-regulated in uveal melanoma cell lines (Venza et al., 2014) and Sun et al. (2009) found that genistein inhibited uveal melanoma cell proliferation in part through decreasing *miR-27a* expression. These studies suggested that ncRNAs not only contribute to the understanding of the molecular mechanisms of EDs, but also have important implications for the development of new therapeutic targets.

With the progression of disease-related ncRNA study, some disease-associated ncRNA resources were constructed, such as LncRNADisease (Chen et al., 2013), miR2Disease (Jiang et al., 2009), and MNDR (Cui et al., 2018). These databases contained a wide variety of human diseases and there were only a few ncRNAs involved in EDs. Subsequently, several ophthalmology-related databases were built recently. For example, KmeyeDB (Kawamura et al., 2010) and RetinoGenetics (Ran et al., 2014) provided gene mutations in EDs. miRNeye (Karali et al., 2010) focused on the differentially expressed miRNAs in ocular tissues. iSyTE (Kakrana et al., 2018) integrated all publicly available lens gene expression data. However, there is still no an ED-specialized database that provides the comprehensive resource on diverse types of ncRNAs across various EDs.

To fill the gap, we constructed Nc2Eye, a manually curated database, to provide experimentally validated ncRNA-ED associations. The current version of Nc2Eye contains 7088 associations between 104 EDs and 4363 ncRNAs in eight species, through a comprehensive review of more than 2400 published papers. To investigate the underlying molecular mechanism of EDs, Nc2Eye gathered more than 90 high-throughput transcriptome datasets in EDs, and normal controls. In addition, Nc2Eye provided a user-friendly interface for convenient and flexible data browsing, querying, retrieving and submitting. Nc2Eye is a timely and valuable resource to significantly improve our understanding of ncRNA dysfunction in EDs.

DATA COLLECTION AND DATABASE CONTENT

Experimentally Supported ncRNA-ED Associations

To ensure the accuracy and reliability in the data collection process, we manually curated all Nc2Eye entries by the following steps. First, we searched the PubMed database (Sayers et al., 2019) with the MeSH term "EDs" and a list of keywords about ncRNAs, including "ncRNA," "non-coding RNA," "non-coding," "lncRNA," "long non-coding RNA," "miRNA," "microRNA," "siRNA," "shRNA," "snoRNA," "piRNA," "circular RNA," "circRNA," "miR-," "and "let-." Considering that some papers have not been assigned MeSH term, we chose the eye disease names from MeSH term as keywords. Then, we searched with the keyword combination: each eye disease and ncRNA category name as complementary. After that, we selected published papers about the ED-related ncRNAs and extracted experimentally supported ncRNA-ED associations manually. All selected literature was reviewed by at least two researchers. In this step, we retrieved the ncRNA name, ncRNA category (e.g., miRNA, lncRNA, and circRNA), disease name, species, tissues/cell line, methods (e.g., microarray, qPCR, and Western blot), expression pattern (e.g., up-regulated, down-regulated), functional description and reference (PubMed ID, year of publication, title). In addition, we recorded whether the ncRNA was related to drugs according to the paper. Finally, we standardized the eye disease names refer to the MeSH term and searched the gene ID, Ensembl ID, and synonyms of each lncRNA from Ensembl (Zerbino et al., 2018) and NCBI Gene database (Brown et al., 2015).

After comprehensively reviewing more than 2400 published papers, a total of 7088 associations between 104 EDs and 4363 ncRNAs in eight species were manually collected. There were 132 of these ncRNAs were reported to be related with drugs in 14 EDs. Distribution of ncRNA-associated entries in each species and ncRNA category is listed in **Table 1**.

High-Throughput Transcriptome Datasets in EDs

A great number of studies based on high-throughput experiments have emerged in recent years (Liu X. et al., 2018;

TABLE 1 | Statistics for the ncRNA-ED entries in the Nc2Eye database.

Species	miRNA	lncRNA	circRNA	siRNA	shRNA	piRNA	snoRNA	snRNA	Total
Homo sapiens	3044	968	479	274	48	18	–	2	4833
Mus musculus	1016	165	2	137	41	–	4	1	1366
Rattus norvegicus	623	9	2	75	24	–	–	–	733
Canis familiaris	47	–	–	–	2	–	–	–	49
Macaca mulatta	17	5	–	14	3	–	–	–	39
Bos taurus	8	–	–	17	1	–	–	–	26
Oryctolagus cuniculus	1	–	–	21	–	–	–	–	22
Danio rerio	20	–	–	–	–	–	–	–	20
Total	4776	1147	483	538	119	18	4	3	7088

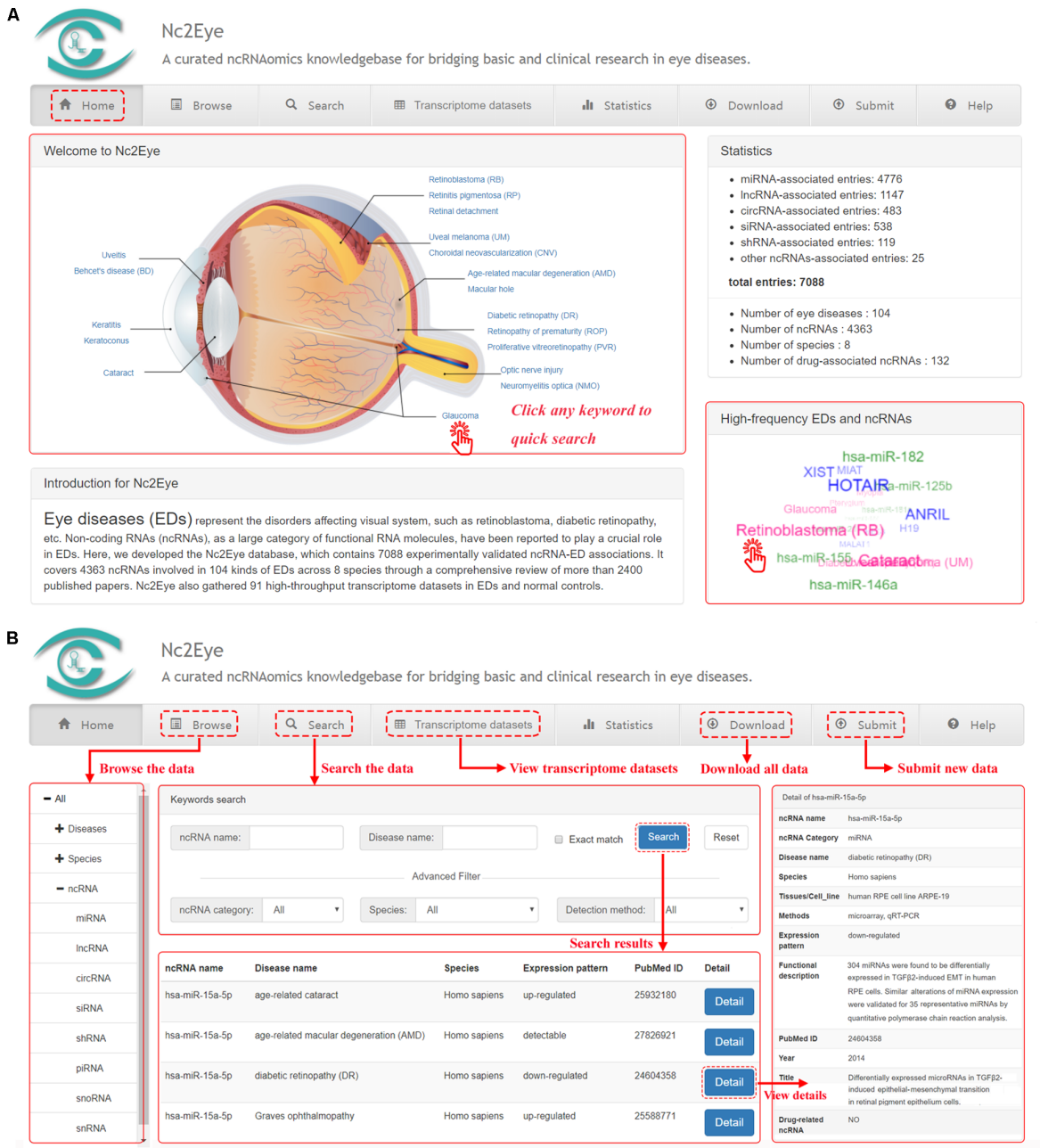


FIGURE 1 | A schematic workflow of Nc2Eye. **(A)** The 'Home' page allow to quick research for ncRNA-ED associations. **(B)** The 'Browse' and 'Search' pages allow the users to browse and search ncRNA-ED associations. The 'Transcriptome datasets' page shows public high-throughput transcriptome datasets. Users can download all ncRNA-ED association data in the 'Download' page and submit new ncRNA-ED association in the 'Submit' page.

Liu Y. et al., 2018; Xu et al., 2018). Mining and collecting high-throughput transcriptome datasets will help to investigate the underlying molecular mechanism of EDs. Nc2Eye gathered 91 microarray and next-generation sequencing datasets in EDs versus normal controls, most of which come from gene expression omnibus (GEO) database. Each entry contains "diseases name," "species," "sample type," "case/control description," "sample size," "data type," "accession ID," "platform," and "Pubmed ID." Users can conveniently browse and select transcriptome datasets according to their research requirements and click the GSE ID to jump to the GEO website for further exploration. GEO has developed several tools for data visualization and analysis which enable researchers to easily analyze this data and do not require to download any file (Clough and Barrett, 2016).

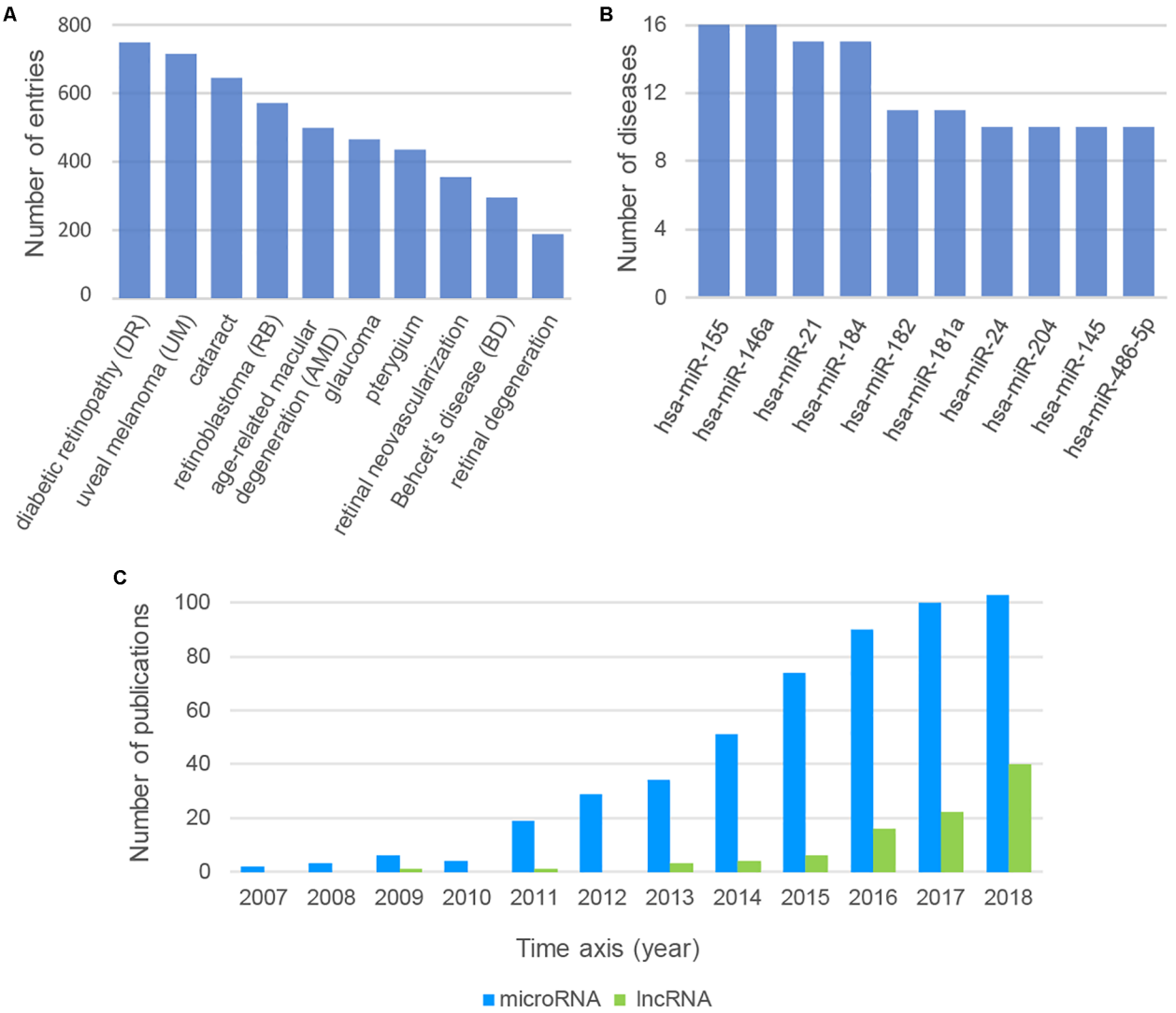


FIGURE 2 | (A) The top10 diseases with the most ncRNA associations. **(B)** The top10 ncRNAs with the most disease relationships. **(C)** The number of ncRNA-ED publications each year.

Finally, all data in Nc2Eye were organized using MySQL (version 5.7.26). The web interface was built in PHP and the web services were deployed using Apache (version 2.4.6). The Nc2Eye database is freely available at http://nc2eye.bio-data.cn/.

USER INTERFACE

Nc2Eye constructed a user-friendly web interface for convenient and flexible data browsing, querying, retrieving and submitting (**Figure 1**). In the "Home" page, users can click keywords in the eye anatomy and "high-frequency EDs and ncRNAs' to quickly research. In the "Browse" page, users can glance through Nc2Eye by clicking on a specific ED name, species or a class of ncRNA, and a list of corresponding entries will be displayed. Nc2Eye provides a "Search" page that enables users to search by ncRNA name, ED name or both. Defaulting to fuzzy search, and Nc2Eye also supports exact search. Moreover, users can use the "Advanced Filter" to filter the results by ncRNA category, species and detection method. In the "Transcriptome datasets" page, 91 expression profiles including EDs versus normal controls are listed and can be filtered by inputting keywords from any column. Nc2Eye also provides a submission interface that allows researchers to submit new ncRNA-ED association data which is not documented. At last, all ncRNA-ED association data in Nc2Eye can be downloaded in the "Download" page and a user tutorial of the website is available in the "Help" page.

DISCUSSION AND CONCLUSION

Eye diseases are a common group of disorders in the visual system, causing visual impairment and blindness. Emerging evidence demonstrated that the dysregulation of ncRNAs played

critical roles in various EDs. We developed Nc2Eye, a curated knowledgebase of experimentally validated ncRNAs associated with EDs, which provides a global landscape of ncRNAs in EDs.

We could find some important messages behind the large and complex data resources by analyzing the data from Nc2Eye. Top10 diseases with the most ncRNA associations were listed (**Figure 2A**). Diabetic retinopathy (DR) ranked first, which is associated with 583 ncRNAs, including 32 drug-related ncRNAs. This result suggested that we can develop more drugs to treat DR by targeting these ncRNAs. The top10 ncRNAs with the most disease relationships were shown in **Figure 2B**. The *hsa-miR-155* has the most connection with diseases and it is associated with 16 diseases, such as DR and RB, indicating the importance of this miRNA in ophthalmopathy. In addition, we tracked the number of ncRNA-ED publications each year (**Figure 2C**) and found the research about lncRNA is increasing year by year which implies lncRNAs will become a hot topic in ophthalmopathy.

CONCLUSION

In conclusion, Nc2Eye, an eye disorders-specialized database, provides global insights into ncRNA functions in various EDs. We believe that Nc2Eye will be beneficial to researchers to dissect the underlying mechanism of ophthalmopathy. In the future, a confidence score system will be developed to estimate the reliability of a specific ncRNA-ED association according to experimental evidence. In addition, we plan to update the database every 3 months to extend newly ED-related relationships and make it more powerful.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://nc2eye.bio-data.cn/download.html.

AUTHOR CONTRIBUTIONS

HC and LX conceived and designed the experiments. YZ, ZX, and FY collected and analyzed the data. FG designed the website. YZ and LX wrote and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (61903277), the Scientific Research Foundation for Talents of Wenzhou Medical University, and the Affiliated Eye Hospital for Ophthalmology and Optometry (KYQD20190103 and KYQD20190102).

REFERENCES

Bhattacharya, S., Ha-Thuc, V., and Srinivasan, P. (2011). MeSH: a window into full text for document summarization. *Bioinformatics* 27, i120–i128. doi: 10.1093/bioinformatics/btr223

Bourne, R. R. A., Flaxman, S. R., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., et al. (2017). Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *Lancet Glob. Health* 5:e0888-97. doi: 10.1016/S2214-109X(17)30293-0

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., et al. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 43, D36–D42. doi: 10.1093/nar/gku1055

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986. doi: 10.1093/nar/gks1099

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9-5

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). MNDR v2.0: an updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res.* 46, D371–D374. doi: 10.1093/nar/gkx1025

Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–874. doi: 10.1038/nrg3074

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37, D98–D104. doi: 10.1093/nar/gkn714

Kakrana, A., Yang, A., Anand, D., Djordjevic, D., Ramachandruni, D., Singh, A., et al. (2018). iSyTE 2.0: a database for expression-based gene discovery in the eye. *Nucleic Acids Res.* 46, D875–D885. doi: 10.1093/nar/gkx837

Karali, M., Peluso, I., Gennarino, V. A., Bilio, M., Verde, R., Lago, G., et al. (2010). miRNeye: a microRNA expression atlas of the mouse eye. *BMC Genom.* 11:715. doi: 10.1186/1471-2164-11-715

Kawamura, T., Ohtsubo, M., Mitsuyama, S., Ohno-Nakamura, S., Shimizu, N., and Minoshima, S. (2010). KMeyeDB: a graphical database of mutations in genes that cause eye diseases. *Hum. Mutat.* 31, 667–674. doi: 10.1002/humu.21240

Li, F., Wen, X., Zhang, H., and Fan, X. (2016). Novel insights into the role of long noncoding RNA in ocular diseases. *Int. J. Mol. Sci.* 17:478. doi: 10.3390/ijms17040478

Liu, C., Ge, H. M., Liu, B. H., Dong, R., Shan, K., Chen, X., et al. (2019). Targeting pericyte-endothelial cell crosstalk by circular RNA-cPWWP2A inhibition aggravates diabetes-induced microvascular dysfunction. *Proc. Natl. Acad. Sci. U.S.A.* 116, 7455–7464. doi: 10.1073/pnas.1814874116

Liu, X., Liu, B., Zhou, M., Fan, F., Yu, M., Gao, C., et al. (2018). Circular RNA HIPK3 regulates human lens epithelial cells proliferation and apoptosis by targeting the miR-193a/CRYAA axis. *Biochem. Biophys. Res. Commun.* 503, 2277–2285. doi: 10.1016/j.bbrc.2018.06.149

Liu, Y., Chen, Y., Wang, Y., Zhang, X., Gao, K., Chen, S., et al. (2018). microRNA profiling in glaucoma eyes with varying degrees of optic neuropathy by using next-generation sequencing. *Invest Ophthalmol. Vis. Sci.* 59, 2955–2966. doi: 10.1167/iovs.17-23599

Ran, X., Cai, W. J., Huang, X. F., Liu, Q., Lu, F., Qu, J., et al. (2014). 'RetinoGenetics': a comprehensive mutation database for genes related to inherited retinal degeneration. *Database* 2014:bau047. doi: 10.1093/database/bau047

Rassi, D. M., De Paiva, C. S., Dias, L. C., Modulo, C. M., Adriano, L., Fantucci, M. Z., et al. (2017). Review: MicroRNAS in ocular surface and dry eye diseases. *Ocul. Surf.* 15, 660–669. doi: 10.1016/j.jtos.2017.05.007

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., et al. (2019). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 47, D23–D28. doi: 10.1093/nar/gky1069

Sun, Q., Cong, R., Yan, H., Gu, H., Zeng, Y., Liu, N., et al. (2009). Genistein inhibits growth of human uveal melanoma cells and affects microRNA-27a and target gene expression. *Oncol. Rep.* 22, 563–567. doi: 10.3892/or-00000472

Venza, M., Dell'Aversana, C., Visalli, M., Altucci, L., Teti, D., and Venza, I. (2014). Identification of microRNA expression patterns in cutaneous and uveal melanoma cell lines. *Tumori* 100, e4–e7. doi: 10.1700/1430.15828

Wu, X. Z., Cui, H. P., Lv, H. J., and Feng, L. (2019). Knockdown of lncRNA PVT1 inhibits retinoblastoma progression by sponging miR-488-3p. *Biomed. Pharmacother.* 112:108627. doi: 10.1016/j.biopha.2019.108627

Xu, B., Ma, R., Ren, H., and Qian, J. (2018). Genome-wide analysis of uveal melanoma metastasis-associated lncrnas and their functional network. *DNA Cell. Biol.* 37, 99–108. doi: 10.1089/dna.2017.4015

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761. doi: 10.1093/nar/gkx1098

Zhang, S. J., Chen, X., Li, C. P., Li, X. M., Liu, C., Liu, B. H., et al. (2017). Identification and characterization of circular RNAs as a New Class of putative biomarkers in diabetes retinopathy. *Invest. Ophthalmol. Vis. Sci.* 58, 6500–6509. doi: 10.1167/iovs.17-22698

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 © 2020 Zhang, Xue, Guo, Yu, Xu 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 19 February 2020
doi: 10.3389/fbioe.2020.00101

The Integrated Transcriptome Bioinformatics Analysis Identifies Key Genes and Cellular Components for Spinal Cord Injury-Related Neuropathic Pain

Runzhi Huang[1,2†], Tong Meng[1,2†], Rui Zhu[1,2], Lijuan Zhao[1,2], Dianwen Song[3], Huabin Yin[3], Zongqiang Huang[4], Liming Cheng[1,2]* and Jie Zhang[1,2,5]**

[1] Division of Spine, Department of Orthopedics, Tongji Hospital Affiliated to Tongji University School of Medicine, Shanghai, China, [2] Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Orthopaedic Department of Tongji Hospital, School of Medicine, Tongji University, Shanghai, China, [3] Department of Orthopedics, Shanghai General Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China, [4] Department of Orthopedics, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [5] Department of Prevention, Tongji University School of Medicine, Tongji University, Shanghai, China

Background: Spinal cord injury (SCI) is one of the most devastating diseases with a high incidence rate around the world. SCI-related neuropathic pain (NeP) is a common complication, whereas its pathomechanism is still unclear. The purpose of this study is to identify key genes and cellular components for SCI-related NeP by an integrated transcriptome bioinformatics analysis.

Methods: The gene expression profile of 25 peripheral blood samples from chronic phase SCI patients (E-GEOD-69901) and 337 normal peripheral blood samples were downloaded from ArrayExpress and Genotype-Tissue Expression Portal (GTEx), respectively. A total of 3,368 normal peripheral blood mononuclear cells (PBMC) were download from Sequence Read Archive (SRA713577). Non-parametric tests were used to evaluate the association between all of differential expression genes (DEGs) and SCI-related NeP. CellPhoneDB algorithm was performed to identify the ligand–receptor interactions and their cellular localization among single PBMCs. Transcription factor (TF) enrichment analysis and Gene Set Variation Analysis (GSVA) were used to identify the potential upstream regulatory TFs and downstream signaling pathways, respectively. Co-expression analysis among significantly enriched TFs, key cellular communication genes and differentially expressed signaling pathways were performed to identify key genes and cellular components for SCI-related NeP.

Results: A total of 2,314 genes were identified as DEGs between the experimental and the control group. Five proteins (ADRB2, LGALS9, PECAM1, HAVCR2, LRP1) were identified in the overlap of proteins in the significant ligand-receptor interactions of PBMCs and protein-protein interaction (PPI) network based on the DEGs. Only HAVCR2

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Michael Poidinger,
Royal Children's Hospital, Australia
Sainan Li,
Harvard Medical School,
United States

***Correspondence:**
Zongqiang Huang
gzhuangzq@163.com
Liming Cheng
mengtong@medmail.com.cn;
limingcheng@tongji.edu.cn
Jie Zhang
jiezhang@tongji.edu.cn

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

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *18 November 2019*
Accepted: *03 February 2020*
Published: *19 February 2020*

Citation:
Huang R, Meng T, Zhu R, Zhao L, Song D, Yin H, Huang Z, Cheng L and Zhang J (2020) The Integrated Transcriptome Bioinformatics Analysis Identifies Key Genes and Cellular Components for Spinal Cord Injury-Related Neuropathic Pain. Front. Bioeng. Biotechnol. 8:101. doi: 10.3389/fbioe.2020.00101

was significantly associated with NeP ($P = 0.005$). Besides, the co-expression analysis revealed that TF YY1 had significantly co-expression pattern with cellular communication receptor HAVCR2 ($R = -0.54$, $P < 0.001$) in NK cells while HAVCR2 was also co-expressed with mTOR signaling pathway ($R = 0.57$, $P < 0.001$). The results of RT-qPCR and external dataset validation supported the signaling axis with the most significant co-expression patterns.

Conclusion: In peripheral blood of chronic SCI, HAVCR2 might act as a key receptor on the surface of NK cells and interact with ligand LGALS9 secreted by $CD14^+$ monocytes, inhibiting NK cells through mTOR signaling pathway and ultimately predicting the occurrence of SCI-related NeP. This hypothetical signaling axis may provide prognostic biomarkers and therapeutic targets for SCI-related NeP.

Keywords: spinal cord injury, neuropathic pain, peripheral blood, single-cell sequencing, cellular communication

INTRODUCTION

Spinal cord injury (SCI) refers to the damage to the spinal cord due to trauma, disease or degeneration (Cheshire et al., 1996; Brienza et al., 2018). According to the National Spinal Cord Injury Statistical Center, there are about 12,000 new cases of SCI each year in the United States (Natinal Spinal Cord Injury Statistical Center [NSCISC], 2014). The global SCI incidence is 40 to 80 new cases per million population per year (New et al., 2014). It not only induces locomotor deficits or even complete paralysis physically, but also generates despairing psychological stress (Budh and Osteraker, 2007). Therapeutically, there are no effective treatment strategies for SCI-induced neurological deficits, leading to a high disability rate and adding heavy burdens to the individual family and the whole society. In order to bring tangible benefits to patients with SCI, there is a pressing need to explore the pathologic mechanisms which may provide candidate targets for treatment (Azarhomayoun et al., 2018; Musubire et al., 2019).

Generally, SCI is categorized into three phases: the acute phase (0–15 days), the sub-acute phase (3–5 months) and the chronic phase (6–12 months) (Pouw et al., 2011). Although the functional status of the chronic phase may be considered clinically similar, regardless of the level of injury, new types of pathologies at both micro and macro level occurs involving a variety of aberrant molecules and cellular components, especially immune cells (Metz et al., 2000; Chen et al., 2013; Van Niekerk et al., 2016). The most common pathological features during the chronic phase are the formation of the glial scar resulting from persistent glial activation and neuronal hyperactivity associated with reactive astrocytes, microglias, and infiltrating macrophages (Takeura et al., 2019). In addition, all of these pathological features are related to neuropathic pain (NeP). Thus, we suppose that identifying the mechanism of NeP and predicting its occurrence.

Neuropathic pain is reported to occur in 40–50% of SCI patients and typically develops within the first year following SCI as the chronic presentation (Siddall et al., 2003; Werhagen et al., 2004). Currently, its treatment is difficult and the efficacy of the recommended treatment options are modest (Finnerup et al., 2015). Although the pathological mechanism of NeP is still unclear, it may be associated with the dynamic process of nerve regeneration and immune response (Yune et al., 2007; Lee et al., 2008; Pinzon et al., 2008; Takeura et al., 2019). In addition, no factor has been identified to predict the occurrence of SCI-related NeP. With regard to human specimens, peripheral blood from patients with SCI is the more accessible and minimally invasive than injured spinal cord in clinical practice.

Thus, an integrated transcriptome bioinformatics analysis based on bulk RNA sequence and single-cell RNA sequence was performed to identify differentially expressed genes and cellular communications, key ligand-receptor interactions and their cellular localization in peripheral blood of patients with SCI. In addition, potential upstream transcription factors (TFs) and downstream signaling pathways of key cellular communication genes were also identified to draw a signaling axis, which might provide candidate predictors and therapeutic targets for SCI and SCI-related NeP.

MATERIALS AND METHODS

Data Collection

This study was approved by the Ethics Committee of Tongji Hospital affiliated to Tongji University School of Medicine. The gene expression profile of 25 peripheral blood samples from chronic phase SCI patients (E-GEOD-69901) (Platform: Affy Primeview Gene Expression Array) were downloaded from ArrayExpress[1] as the experimental group. Because we wanted to identity DEGs between peripheral blood of normal people and

Abbreviations: CIBERSORT, cell type identification by estimating relative subsets of RNA transcripts; DEG, differential expression gene; FDR, false discovery rate; GO, gene ontology; GSVA, Gene Set Variation Analysis; GTEx, Genotype-Tissue Expression Portal; KEGG, Kyoto Encyclopedia of Genes and Genomes; NeP, neuropathic pain; PBMC, peripheral blood mononuclear cells; PCA, principal component analysis; PPI, protein-protein interaction; RPKM, Reads Per Kilobase per Million; SCI, spinal cord injury; SRA, Sequence Read Archive; TF, transcription factor; Tim-3, T-Cell Immunoglobulin Mucin Family Member 3; t-SNE, t-distributed Stochastic Neighbor Embedding; TSO, template switch oligo.

[1]https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-69901/

patients with SCI, but E-GEOD-69901 did not have the data of normal people. Therefore, we did not use the control set that was published with the E-GEOD-69901. The control group including 337 normal peripheral blood samples was downloaded from the Genotype-Tissue Expression Portal[2] (GTEx) (Consortium, 2015). A total of 3,368 normal peripheral blood mononuclear cells (PBMC) were download from Sequence Read Archive[3] (SRA713577) (Freytag et al., 2018). The single cell set was one 10X genomics object of 3,368 cells from the same one person. In order to ensure the repeatability of the experiment, we used the RData file including matrixes of Reads Per Kilobase per Million mapped reads (RPKM) and raw counts from the PanglaoDB (Franzen et al., 2019). Besides, we have carried out an external validation. Two Affy Primeview dataset (GSE82152 and E-MTAB-5151) including normal peripheral blood samples were used as the control group for differential expression analysis. We did not use these published data in the initial study because of the small sample size.

Differential Gene Expression Analysis

First of all, non-peripheral-blood specific expression genes (no expression was detected in both control group and experimental group) were filtered. The limma package was used to find differential expression genes (DEGs) after normalization between two batches of data (Ritchie et al., 2015). Limma algorithm was originally developed for the analysis of microarray data, and its protocol for RNA-seq analysis also normalized the data using voom algorithm to process it into data similar to microarray for analysis (Ritchie et al., 2015). Thus, we used the GTEx dataset as the control group for a larger sample size. And the data of GTEx were normalized by voom algorithm and the batch effect of data were eliminated by the function named normalizeBetweenArrays. The standard of DEGs was an absolute log2 fold change greater than 2 and false discovery rate (FDR) P-value < 0.05.

Functional Enrichment Analysis and Construction of Protein-Protein Interaction Network

To further explore the function of the DEGs above, the functional enrichment analysis was performed using the clusterProfiler including gene ontology (GO) term and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Yu et al., 2012). String database was used to construct a protein-protein interaction (PPI) network based on the DEGs and the names of all the interacting proteins and the protein-coding genes were extracted from the network (Szklarczyk et al., 2019). Besides, non-parametric tests were used to assess the association between all DEGs and NeP.

Processing of Single-Cell RNA-Seq Data

Raw reads in the sra file were first separated into pair-ended reads fastq files, which were trimmed to remove polyA tail sequence and the template switch oligo (TSO) sequence. Then, the clean reads were aligned to the hg38 human transcriptome (UCSC) and quantified by the Cell Ranger Single Cell Software Suite 3.3.1[4].

For all 3,368 normal sequenced single PBMCs, cells with either fewer than 100,000 transcripts or fewer than 1,500 genes were filtered out. Besides, only the genes expressed in at least three single cells and their expression levels greater than 1 were considered for downstream analysis. Additionally, the Seurat method was applied to downstream analysis (Butler et al., 2018).

First, "vst" selection method was used to find variable genes, which were the input features for initial principal component analysis (PCA) (Butler et al., 2018). Then, the jackstraw analysis was performed to select the principal components (PCs) with P-values < 0.05 (Chung and Storey, 2015). Significant PCs were incorporated into further t-distributed Stochastic Neighbor Embedding (t-SNE) to identify different cell clusters with DEGs (resolution = 0.50). The standard of DEGs was an absolute log2 fold change greater than 0.50 and FDR value <0.05. Only the genes with an absolute log fold change greater than 0.5 and FDR P-value < 0.05 were selected as DEGs. The distribution and expression of top 10 DEGs were displayed by feature plots and heat maps, respectively. Additionally, scMatch (Hou et al., 2019), singleR (Aran et al., 2019), and CellMarker (Zhang et al., 2019) database were used as references for defining each cluster.

Identification of the Significant Cellular Communication Among PBMCs

CellPhoneDB (Vento-Tormo et al., 2018; Efremova et al., 2019), a repository of ligand-receptor complexes and a statistical tool to predict the cell-type specificity of cell-cell communication via these molecular interactions, was performed to identify the ligand-receptor interactions and their cell localization among single PBMCs. The names of all the interacting proteins and the protein-coding genes were extracted from the network. Then, the Venn plot was used to illustrate the intersection of CellPhoneDB results and the PPI network.

Validation by CIBESORT Algorithm

Cell type identification by estimating relative subsets of RNA transcripts (CIBERSORT) algorithm was used to characterize the cell composition from the complex tissues according to their gene expression profiles (Newman et al., 2015). Then the fraction of 22 types of immune cell was estimated in peripheral blood based on the normalized gene expression profiles. Samples with a CIBERSORT output of $P < 0.05$ were considered to be eligible for further analysis. Then, the Wilcoxon rank-sum test was applied to identify the immune cells with the fractions between peripheral blood samples from SCI patients and normal control samples. It should be pointed out that the CIBERSORT algorithm could correct for differences between different platforms and batches of data in the initial version (Newman et al., 2015). Additionally, the next generation CIBERSORT algorithm (CIBERSORTx) (Newman et al., 2019), including enable batch correction and disable quantile normalization algorithm, were applied to confirm the reliability of the results.

[2]https://commonfund.nih.gov/GTEx/

[3]https://www.ncbi.nlm.nih.gov/sra/SRX4149408[accn]

[4]http://10xgenomics.com/

Identification of the Potential Upstream Transcription Factors of Key Cellular Communication Genes

The DAVID database UCSC TFBS function module was used for TF enrichment analysis to identify target TFs among the DEGs. Differentially expressed TFs with enrichment analysis FDR < 0.05 were defined as significantly enriched TFs.

GSVA and Co-expression Analysis

Gene Set Variation Analysis (GSVA) pathway analysis was performed to evaluate the expression levels of 185 KEGG pathways. The limma method was also used to find differentially expressed pathways between peripheral blood samples from SCI patients and normal control samples (Ritchie et al., 2015). The FDR *P*-value < 0.05, the log (fold-change) >0.5 or <-0.5 was defined a downregulated or upregulated pathway, respectively.

Eventually, a co-expression Pearson correlation analysis among significantly enriched TFs, key cellular communication genes and differentially expressed KEGG pathways.

Reverse Transcription Quantitative Real-Time PCR (RT-qPCR) Assays and External Dataset Validation

Total RNA was isolated from was extracted from human whole blood of 16 patients with fractures complicated with SCI, 16 patients with fractures but no SCI and 8 normal adults, using QIAamp RNA Blood Mini Kits (Qiagen, catalog number 52304) according to the modified protocol of manufacturers. All cDNA generated from reverse transcription [PrimeScript RT Reagent Kit (Perfect Real Time) (Takara Bio)] was used for quantitative PCR analysis by ABI PRISM 7900 Sequence Detection System (Applied Biosystems, Foster City, CA, United States) and SYBR Premix Ex Taq (Tli RNaseH Plus) PCR Kit (Takara Bio). The relative expression levels of eight key genes (YY1, CEBPB, HAVCR2, LGALS9, MTOR, RPS6, RPS6KB1, RPS6KB2) and reference gene were determined by the $2^{-\Delta\Delta Ct}$ method. Kruskal–Wallis test was used to identity the statistical difference of gene expression among groups.

Additionally, since the platform effect and batch effect could inevitably affect the results of the differential genes, two Affy Primeview dataset (GSE82152 and E-MTAB-5151) including normal peripheral blood samples were used as the control group for differential expression analysis. The limma package was also used to find differential expression genes (DEGs) after normalization between two batches of data (Ritchie et al., 2015). The standard of DEGs was an absolute log fold change greater than 1 and FDR *P*-value < 0.05.

Statistics Analysis

Two-sided *P*-value < 0.05 was thought to be statistical significance. All statistical analysis was conducted with R version 3.6.1 software (Institute for Statistics and Mathematics, Vienna, Austria)[5] (Package: limma, Seurat, ggplot2, SingleR, reticulate, clusterProfiler, GSEABase, GSVA).

[5]www.r-project.org

RESULTS

Differential Gene Expression Analysis

The analysis process of this study was presented in **Figure 1**. For identifying the significantly DEGs, we set the log (fold-change) >2.0 or <-2.0 and FDR < 0.05 as the cutoff and a total of 2,314 genes were identified as DEGs, including 1,152 upregulated ones and 1,162 downregulated ones (**Figure 2A**).

Functional Enrichment Analysis and Construction of Protein-Protein Interaction Network

The functional enrichment analysis for these DEGs in GO terms and KEGG pathways were shown in **Figures 2B,C**, respectively. The biological process of GO terms analysis revealed the enrichment of some remarkable immune cell related terms. Additionally, "secretory granule lumen" "ubiquitin-like protein transferase activity" and "cadherin binding" were also significantly enriched as cellular component or molecular function, which might mean the active aberrant cellular communication in the peripheral blood of SCI patients (**Figure 2B**). The KEGG enrichment analysis also suggested some critical pathways were significantly associated with cellular communication, such as "Endocytosis," "Protein processing in endoplasmic reticulum," "RNA transport," and "NF-κB signaling pathway" (**Figure 2C**).

String database was used to construct a PPI network based on the whole 2,314 DEGs, which included 4,807 PPI relationships related to 799 proteins (**Supplementary Figure S1**).

The Gene Expression Landscapes of 3,368 PBMCs

A t-SNE analysis was performed and clearly identified 13 clusters and 8 cell types (CD4 + T cells, CD14 + Monocytes, Natural killer (NK) cells, B cells, CD8 + T cells, Megakaryocytes, FCGR3A + Monocytes, Dendritic cells) (**Figures 3A,C**). The expression levels of the top 10 DEGs in each cluster and cell type were displayed in **Figures 3B,D**, respectively. The feature plots of each cell type markers reported in the CellMarker database were presented in **Figures 3E–M**.

Identification of Significant Cellular Communication Among PBMCs

The CellPhoneDB analysis was performed to identify the ligand-receptor interactions and their cell localization among single PBMCs. A total of 87 significant ligand-receptor interactions (related to 108 proteins) and their cell localization were identified. Furthermore, 5 proteins (ADRB2, LGALS9, PECAM1, HAVCR2, LRP1) were identified in the overlap of proteins in the significant ligand-receptor interactions of PBMCs and PPI network based on the DEGs. Only HAVCR2 (Hepatitis A Virus Cellular Receptor 2) was significantly associated with NeP ($P = 0.005$) (**Figures 4A,B**). Besides, a total of 87 ligand-receptor interaction relationships and a new PPI network illustrating the interaction among the five proteins were shown in **Figures 4C,D**, respectively. Additionally,

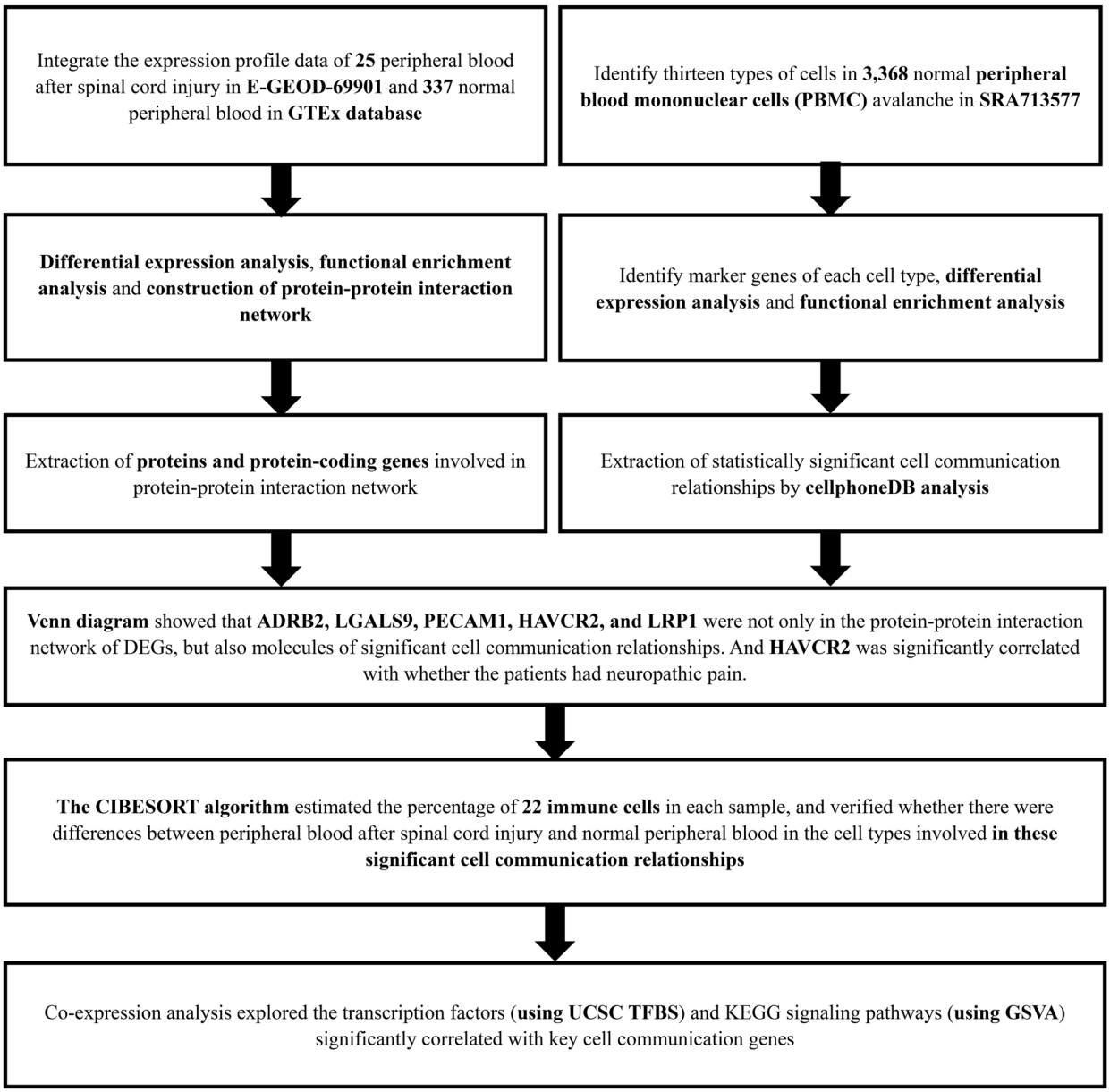


FIGURE 1 | The flow chart of the analysis process. GTEx, Genotype-Tissue Expression; PBMC, peripheral blood mononuclear cell; SRA, Sequence Read Archive; CIBERSORT, Cell type identification by estimating relative subsets of RNA transcripts; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSVA, Gene Set Variation Analysis.

the results of CellPhoneDB analysis including ADRB2, LGALS9, PECAM1, HAVCR2, and LRP1 were presented in **Table 1**.

Validation by CIBESORT Algorithm

The fraction of immune cells in each sample estimated by CIBERSORT algorithm were displayed in **Figures 5A,B**. The results of the Wilcoxon rank-sum test suggested that the fractions of B cells naive ($P < 0.001$), plasma cells ($P < 0.001$), T cells CD4 memory resting ($P < 0.001$), T cells CD4 memory activated ($P < 0.001$), T cells follicular helper ($P = 0.021$), T cells regulatory (Tregs) ($P < 0.001$), T cells gamma delta ($P < 0.001$), NK cells resting ($P < 0.001$), macrophages M1 ($P < 0.001$), macrophages M2 ($P = 0.042$), dendritic cells activated ($P = 0.001$), mast cells resting ($P = 0.001$) and eosinophils ($P < 0.001$) had significantly different cellular fractions between peripheral blood samples from SCI patients and normal control samples (**Figure 5C**). These differential immune cells covered all the cell localizations of ADRB2, LGALS9, PECAM1, HAVCR2, and LRP1. In Addition, the PCA results of all samples suggested the significant differences between the control group and experimental group (**Figure 5D**). Besides, enable batch correction and disable quantile normalization algorithm of CIBERSORTx

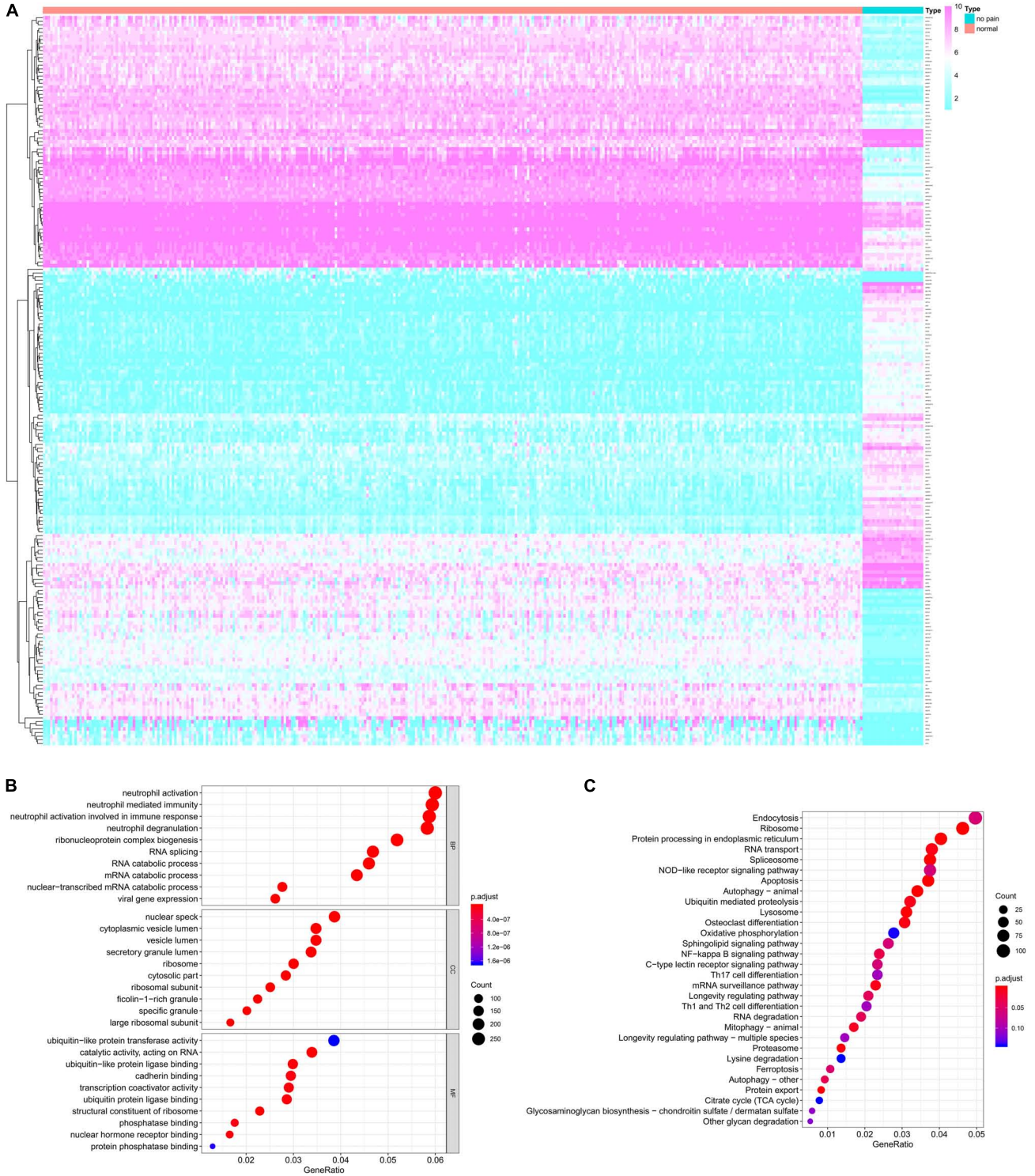


FIGURE 2 | The differentially expressed genes (DEGs) between peripheral blood samples from spinal cord injury (SCI) patients and normal control samples **(A)** and the functional enrichment analysis for these DEGs in gene ontology (GO) terms **(B)** and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways **(C)**. **(A)** The heatmap of DEGs between peripheral blood samples from SCI patients and normal control samples. **(B)** The bubble plot of top 10 significant GO terms in biological process (BP), cellular component (CC) and molecular function (MF). **(C)** The bubble plot of top 20 significant KEGG pathways. DEG, differentially expressed gene; SCI, spinal cord injury; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, biological process; CC, cellular component; MF, molecular function.

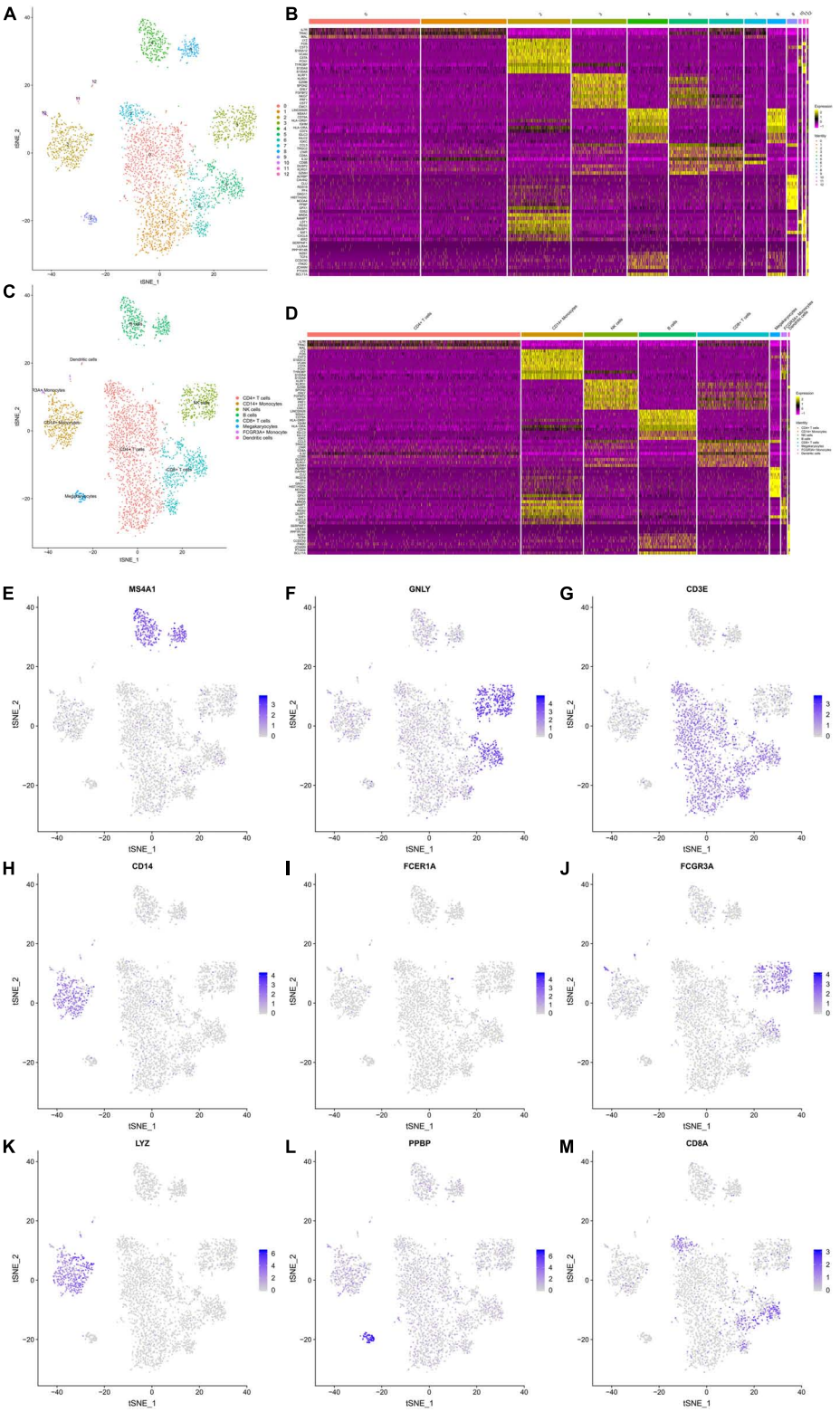


FIGURE 3 | The Gene Expression Landscapes of 3,368 peripheral blood mononuclear cells (PBMCs). A t-distributed stochastic neighbor embedding (t-SNE) analysis was performed, which clearly identified 13 clusters **(A)** and 8 cell types (CD4+ T cells, CD14+ Monocytes, NK cells, B cells, CD8+ T cells, Megakaryocytes, FCGR3A + Monocytes, Dendritic cells) **(C)**. The expression levels the top 10 differentially expressed genes (DEGs) of each cluster **(B)** and cell type **(D)** are displayed in the heatmaps. **(E–M)** illustrate the feature plots of each cell type markers reported in the CellMarker database. PBMC, peripheral blood mononuclear cell; t-SNE, t-distributed stochastic neighbor embedding; DEG, differentially expressed gene.

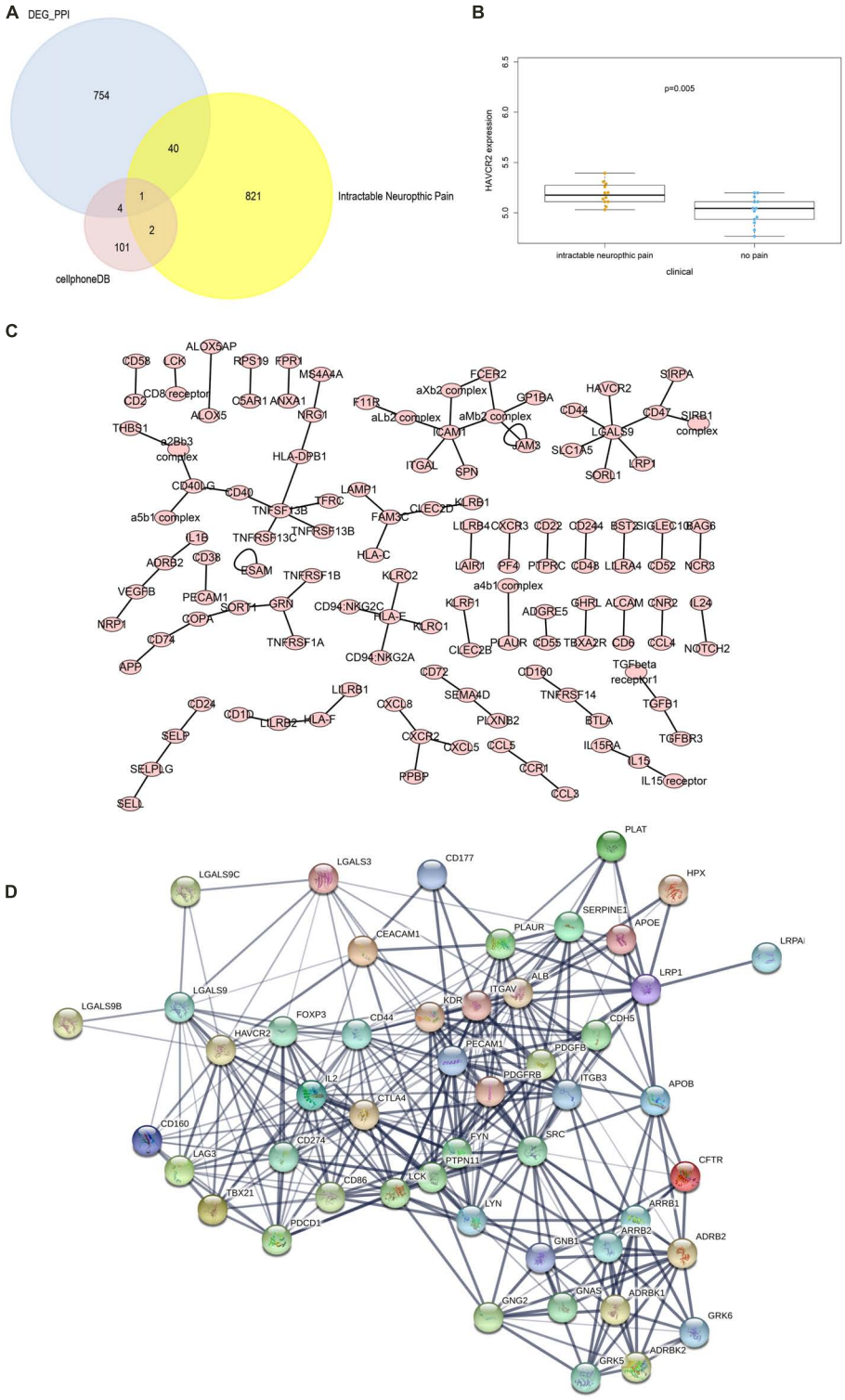


FIGURE 4 | The results of the CellPhoneDB analysis and the Venn plot illustrating five proteins (ADRB2, LGALS9, PECAM1, HAVCR2, LRP1) that not only participated in significant ligand–receptor interactions in peripheral blood mononuclear cells (PBMCs) but Protein-Protein Interaction (PPI) network based on the differentially expressed genes (DEGs). **(A)** The Venn plot illustrating five proteins (ADRB2, LGALS9, PECAM1, HAVCR2, LRP1) that not only participated in significant ligand–receptor interactions in PBMCs but PPI network based on the DEGs, and only HAVCR2 was significantly associated with neuropathic pain (P = 0.005) **(B)**. **(C)** The network of 87 significant ligand–receptor interactions (related to 108 proteins); **(D)** PPI network illustrating the interactions among the ADRB2, LGALS9, PECAM1, HAVCR2, LRP1. PBMC, peripheral blood mononuclear cell; DEG, differentially expressed gene; PPI, Protein-Protein Interaction.

TABLE 1 | The results of CellPhoneDB analysis involved ADRB2, LGALS9, PECAM1, HAVCR2, and LRP1.

ID	Interacting A	Interacting B	Secreted	Rank	Cell localization
CPI-SS098425155	ADRB2	VEGFB	TRUE	0.016	NK cells| Dendritic cells
CPI-SS0C6448B94	IL1B	ADRB2	TRUE	0.031	CD14+ Monocytes| NK cells, Dendritic cells| NK cells, CD14 + Monocytes| NK cells, Dendritic cells| NK cells
CPI-SS0E23CEB91	LGALS9	HAVCR2	TRUE	0.062	B cells| NK cells, CD14+ Monocytes| NK cells, Dendritic cells| NK cells, FCGR3A+ Monocytes| NK cells
CPI-SS0E0DEA7D5	PECAM1	CD38	FALSE	0.062	CD14 + Monocytes| NK cells, Dendritic cells| NK cells, FCGR3A + Monocytes| NK cells, Megakaryocytes| NK cells
CPI-SS0419B80C4	LGALS9	LRP1	TRUE	0.062	B cells| CD14+ Monocytes, CD14+ Monocytes| CD14+ Monocytes, Dendritic cells| CD14+ Monocytes, FCGR3A+ Monocytes| CD14+ Monocytes
CPI-SS002DF6C31	LGALS9	SLC1A5	TRUE	0.062	B cells| Dendritic cells, CD14+ Monocytes| Dendritic cells, Dendritic cells| Dendritic cells, FCGR3A+ Monocytes| Dendritic cells
CPI-SS09C52F54E	LGALS9	SORL1	TRUE	0.375	B cells| CD14+ Monocytes, B cells| CD14+ Monocytes, B cells| CD4 T cells, B cells| CD8 T cells, B cells| dendritic cells, B cells| FCGR3A+ Monocytes, B cells| NK cells, CD14+ Monocytes| CD14+ Monocytes, CD14+ Monocytes| CD4 T cells, CD14+ Monocytes| CD8 T cells, CD14+ Monocytes| Dendritic cells, CD14+ Monocytes| FCGR3A+ Monocytes, CD14+ Monocytes| NK cells, Dendritic cells| CD14+ Monocytes, Dendritic cells| CD4 T cells, Dendritic cells| CD8 T cells, Dendritic cells| Dendritic cells, Dendritic cells| FCGR3A+ Monocytes, Dendritic cells| Megakaryocytes, Dendritic cells| NK cells, FCGR3A+ Monocytes| CD14+ Monocytes, FCGR3A+ Monocytes| CD4 T cells, FCGR3A+ Monocytes| CD8 T cells, FCGR3A+ Monocytes| Dendritic cells, FCGR3A+ Monocytes| FCGR3A+ Monocytes, FCGR3A+ Monocytes| Megakaryocytes, FCGR3A+ Monocytes| NK cells
CPI-SS0703338F5	LGALS9	CD44	TRUE	0.500	B cells| B cells, B cells| CD14+ Monocytes, B cells| CD4 T cells, B cells| CD8 T cells, B cells| Dendritic cells, B cells| FCGR3A+ Monocytes, B cells| Megakaryocytes, B cells| NK cells, CD14+ Monocytes| B cells, CD14+ Monocytes| CD14+ Monocytes, CD14+ Monocytes| CD4 T cells, CD14+ Monocytes| CD8 T cells, CD14+ Monocytes| Dendritic cells, CD14+ Monocytes| FCGR3A+ Monocytes, CD14+ Monocytes| Megakaryocytes, CD14+ Monocytes| NK cells, Dendritic cells| B cells, Dendritic cells| CD14+ Monocytes, Dendritic cells| CD4 T cells, Dendritic cells| CD8 T cells, Dendritic cells| Dendritic cells, Dendritic cells| FCGR3A+ Monocytes, Dendritic cells| Megakaryocytes, Dendritic cells| NK cells, FCGR3A+ Monocytes| B cells, FCGR3A+ Monocytes| CD14+ Monocytes, FCGR3A+ Monocytes| CD4 T cells, FCGR3A+ Monocytes| CD8 T cells, FCGR3A+ Monocytes| Dendritic cells, FCGR3A+ Monocytes| FCGR3A+ Monocytes, FCGR3A+ Monocytes| Megakaryocytes, FCGR3A+ Monocytes| NK cells
CPI-SS014958F32	LGALS9	CD47	TRUE	0.500	B cells| B cells, B cells| CD14+ Monocytes, B cells| CD4 T cells, B cells| CD8 T cells, B cells| Dendritic cells, B cells| FCGR3A+ Monocytes, B cells| Megakaryocytes, B cells| NK cells, CD14+ Monocytes| B cells, CD14+ Monocytes| CD14+ Monocytes, CD14+ Monocytes| CD4 T cells, CD14+ Monocytes| CD8 T cells, CD14+ Monocytes| Dendritic cells, CD14+ Monocytes| FCGR3A+ Monocytes, CD14+ Monocytes| Megakaryocytes, CD14+ Monocytes| NK cells, Dendritic cells| B cells, Dendritic cells| CD14+ Monocytes, Dendritic cells| CD4 T cells, Dendritic cells| CD8 T cells, Dendritic cells| Dendritic cells, Dendritic cells| FCGR3A+ Monocytes, Dendritic cells| Megakaryocytes, Dendritic cells| NK cells, FCGR3A+ Monocytes| B cells, FCGR3A+ Monocytes| CD14+ Monocytes, FCGR3A+ Monocytes| CD4 T cells, FCGR3A+ Monocytes| CD8 T cells, FCGR3A+ Monocytes| Dendritic cells, FCGR3A+ Monocytes| FCGR3A+ Monocytes, FCGR3A+ Monocytes| Megakaryocytes, FCGR3A+ Monocytes| NK cells

were used to eliminate platform effect and batch effect between different dataset. Wilcoxon rank-sum test suggested that the fractions of B cells naive ($P < 0.001$), plasma cells ($P < 0.001$), T cells CD4 memory resting ($P < 0.001$), T cells CD4 memory activated ($P < 0.001$), T cells follicular helper ($P = 0.004$), T cells regulatory (Tregs) ($P < 0.001$), T cells gamma delta ($P < 0.001$), NK cells resting ($P < 0.001$), macrophages M1 ($P < 0.001$), macrophages M2 ($P = 0.002$), dendritic cells activated ($P < 0.001$), mast cells resting ($P = 0.001$), and eosinophils ($P < 0.001$) also had significantly different cellular

TABLE 2 | Transcription factors enrichment analysis of ADRB2, LGALS9, PECAM1, HAVCR2, and LRP1.

TF	Counts in DEGs	Percent of DEGs (%)	Fold Enrichment	*P*-value	FDR
ELK1	1150	49.69749352	1.211034109	7.59E-20	9.32E-17
SP1	588	25.41054451	1.288547397	9.6E-13	1.18E-09
RREB1	859	37.1218669	1.195598062	2.38E-11	2.93E-08
MZF1	1022	44.16594641	1.11604408	0.00000103	0.001270396
YY1	1468	63.43993086	1.076131756	0.00000131	0.001610992
CEBPB	1165	50.34572169	1.100197888	0.00000153	0.00187979
AHR	715	30.8988764	1.146008606	0.00000482	0.005924797
ARNT	1000	43.21521175	1.106636605	0.00000847	0.010404583
SRF	1466	63.35350043	1.068036681	0.0000121	0.014861228
GATA1	1675	72.38547969	1.051350602	0.0000355	0.043593096

TF, transcription factor; DEG, differential expression genes; FDR, false discovery rate.

fractions between peripheral blood samples from SCI patients and normal control samples (**Supplementary Figure S2**).

GSVA

Gene Set Variation Analysis was performed to estimate the expression levels of 185 KEGG pathways and 12 pathways were identified as differentially expressed pathways between peripheral blood samples from SCI patients and normal control samples (**Figures 6A,B**). Especially, some critical pathways were significantly associated with cellular communication and immune response such as "mTOR signaling pathway," "complement and coagulation cascades pathway," and "cysteine and methionine metabolism pathway."

Transcription Factor Enrichment Analysis

The TFs enrichment analysis was firstly performed based on all 2,314 DEGs and a total of 41 TFs were identified with FDR value < 0.05. Moreover, the enrichment analysis revealed that 10 significantly differentially expressed TFs might regulate the promoter regions of ADRB2, LGALS9, PECAM1, HAVCR2, and LRP1 (**Table 2**).

Co-expression Analysis

A co-expression Pearson correlation analysis was used among significantly enriched TFs, key cellular communication genes and differentially expressed KEGG pathways. A regulation network was constructed based on TFs and key cellular communication genes (**Figure 7A**). The bi-clustering heatmap and co-expression heatmap illustrated the expression levels and co-expression patterns of the three components (**Figures 7B,C**). In the co-expression heatmap, the TF Yin and Yang 1 TF (YY1) had significantly co-expression pattern with cellular communication receptor HAVCR2 ($R = -0.54$, $P < 0.001$), while HAVCR2 was also co-expressed with mTOR signaling pathway ($R = 0.57$, $P < 0.001$). Besides, the TF CEBPB was significantly co-expressed with LGALS9 ($R = -0.52$, $P < 0.001$), which was the ligand of HAVCR2 and also co-expressed with HAVCR2 ($R = 0.70$, $P < 0.001$). Moreover, the cellular localizations of the key TFs and target DEGs with co-expression patterns showed that HAVCR2 and LGALS9 were located in NK cells and CD14 + monocytes, respectively (**Figures 7D–L**). Eventually, the sketch map of the signaling axis with the most significant co-expression pattern including YY1, HAVCR2, CEBPB, LGALS9, NK cell, CD14 + monocyte and mTOR signaling pathway was shown in **Figure 8**.

Reverse Transcription Quantitative Real-Time PCR (RT-qPCR) Assays

Kruskal–Wallis test was used to identity the statistical difference of gene expression among groups. The results suggested that TF YY1 (**Figure 9A**, $P < 0.001$) and CEBPB (**Figure 9C**, $P < 0.001$) were upregulated in the peripheral blood of patients with SCI compared with patients with fractures but no SCI and normal adults. HAVCR2 (**Figure 9B**, $P < 0.001$) and LGALS9 (**Figure 9D**, $P < 0.001$) were also abnormally downregulated in peripheral blood of patients with SCI. Some key genes of the mTOR signaling pathway (MTOR, RPS6, RPS6KB1, RPS6KB2) were also identified to be significantly down-regulated in peripheral blood of patients with SCI (**Figures 9E–H**).

External Dataset Validation

Additionally, since the platform effect and batch effect could inevitably affect the results of the differential genes, two Affy Primeview dataset (GSE82152 and E-MTAB-5151) including normal peripheral blood samples were used as the control group for differential expression analysis. Two volcano plots showing the results of differential expression analysis using two Affy Primeview dataset [GSE82152 (**Supplementary Figure S4**) and E-MTAB-5151 (**Supplementary Figure S4**)] including normal peripheral blood samples as the control group. And the differential expression analysis results of YY1, CEBPB, HAVCR2, LGALS9, MTOR, RPS6, RPS6KB1, and RPS6KB2 using GSE82152 and E-MTAB-5151 as control group were summarized in **Supplementary Tables S1** and **S2**, respectively.

DISCUSSION

Spinal cord injury, one of the most devastating diseases, disrupts communication between the central and peripheral nervous, leading to the loss of essential neurological functions. Due to a large number of traffic and industrial accidents, the incidence rate

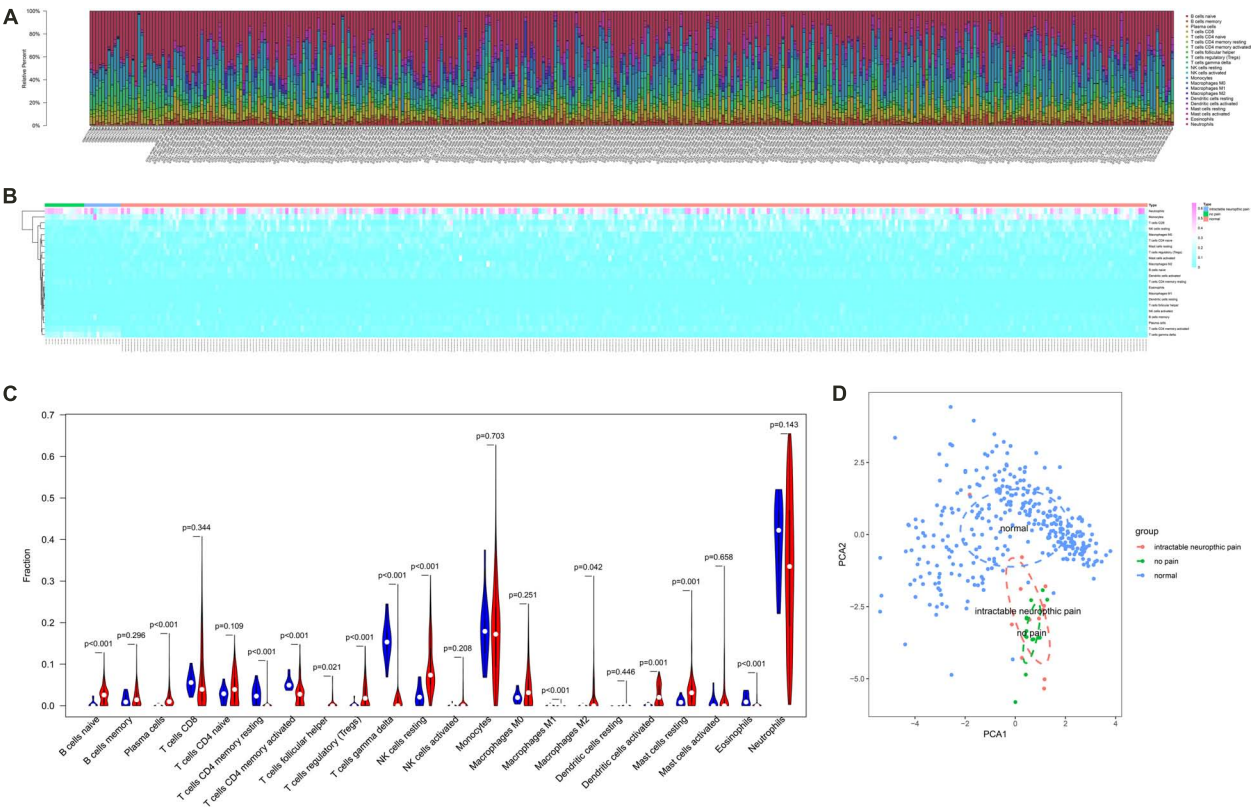


FIGURE 5 | The composition **(A)** and heat map **(B)** of immune cells estimated by Cell type identification by estimating relative subsets of RNA transcripts (CIBERSORT) algorithm in peripheral blood samples from spinal cord injury (SCI) patients and normal control samples. **(C)** The violin plot identifying immune cells different from the two groups (the blue and red bar stand for SCI group and primary normal control samples, respectively). **(D)** The Principal Component Analysis (PCA) result of all samples suggesting the significant differences between the control group and the experimental group. CIBERSORT, cell type identification by estimating relative subsets of RNA transcripts; SCI, spinal cord injury; PCA, principal component analysis.

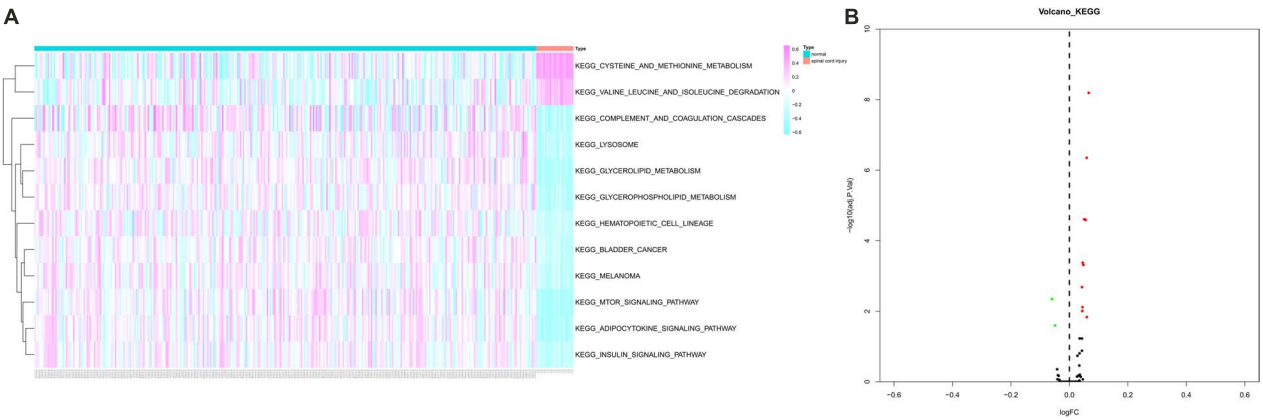


FIGURE 6 | The heat map **(A)** and volcano plot **(B)** showing Kyoto Encyclopedia of Genes and Genomes (KEGG) 12 pathways were identified as differentially expressed pathways [Quantitative by Gene Set Variation Analysis (GSVA)] between peripheral blood samples from spinal cord injury (SCI) patients and normal control samples. KEGG, Kyoto Encyclopedia of Genes and Genomes; GSVA, Gene Set Variation Analysis; SCI, spinal cord injury.

of SCI is increasing rapidly around the world (Azarhomayoun et al., 2018; Musubire et al., 2019). Chronic phase of SCI may last a long period after the acute phase, which bring physically and psychologically devastating traumas to persons with SCI (Budh and Osteraker, 2007). SCI-related NeP is one of the most common symptoms in chronic phase and severely decreases the quality of life (Bryce et al., 2012). The molecular and cellular features often have some changes during the process of

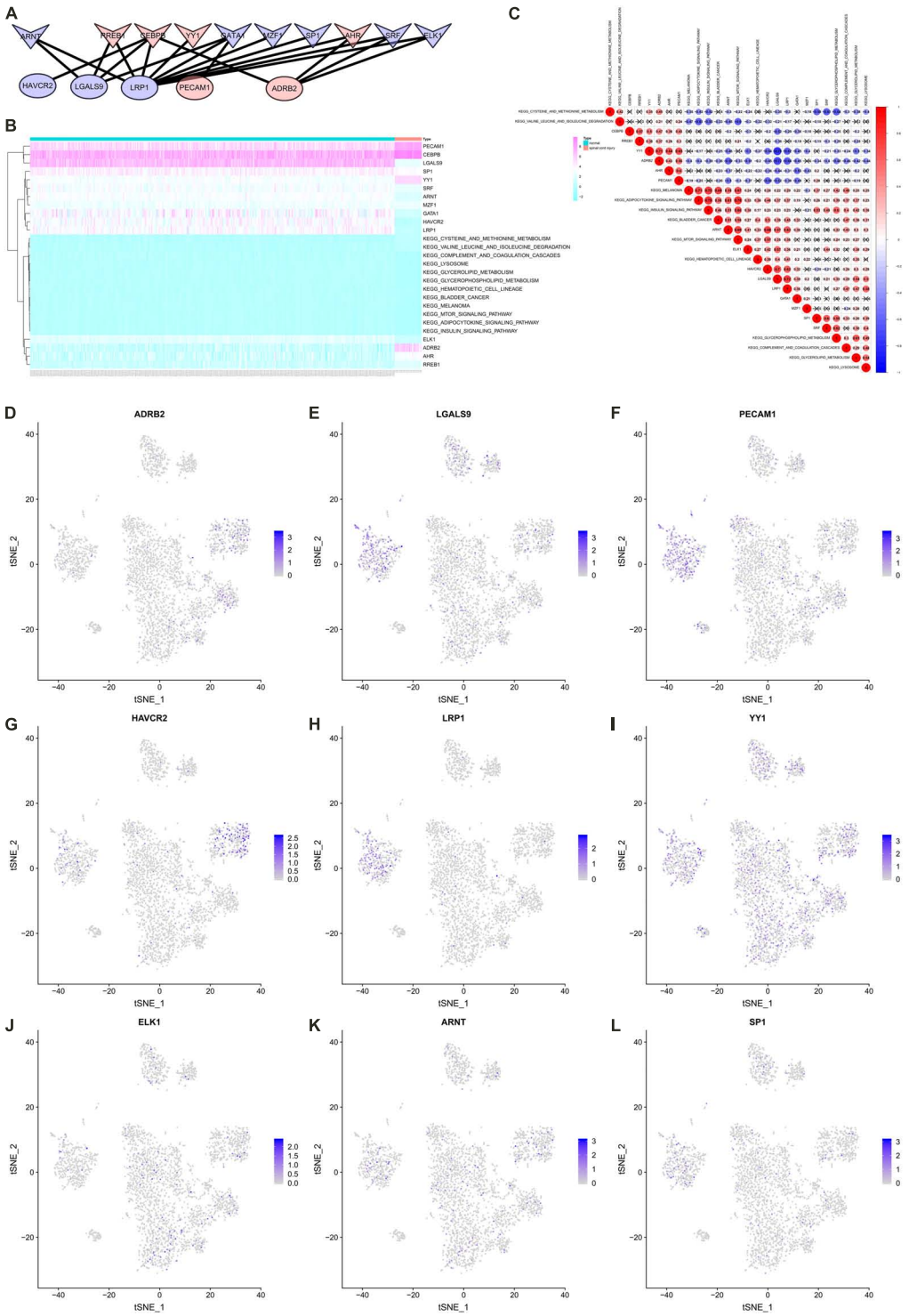


FIGURE 7 | Construct regulation network and identify co-expression patterns among transcription factors (TFs), key cellular communication genes and differentially expressed Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. **(A)** The regulation network of TFs and key cellular communication genes (the V symbols represented TFs, the ellipses represented target DEGs; Red represented significant upregulated and blue represented downregulated). **(B)** The bi-clustering heatmap illustrating the expression levels of TFs, key cellular communication genes and differentially expressed KEGG pathways. **(C)** The co-expression heatmap illustrating the co-expression patterns of TFs, key cellular communication genes and differentially expressed KEGG pathways (in the co-expression heatmap, the transcription factor YY1 had significantly co-expression pattern with cellular communication receptor HAVCR2 ($R = -0.54$, $P < 0.001$), while HAVCR2 was also co-expressed with mTOR signaling pathway ($R = 0.57$, $P < 0.001$). **(D–L)** The feature plots showing the cellular localizations of the key TFs and target DEGs with co-expression patterns. TF, transcription factors; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEG, differentially expressed gene.

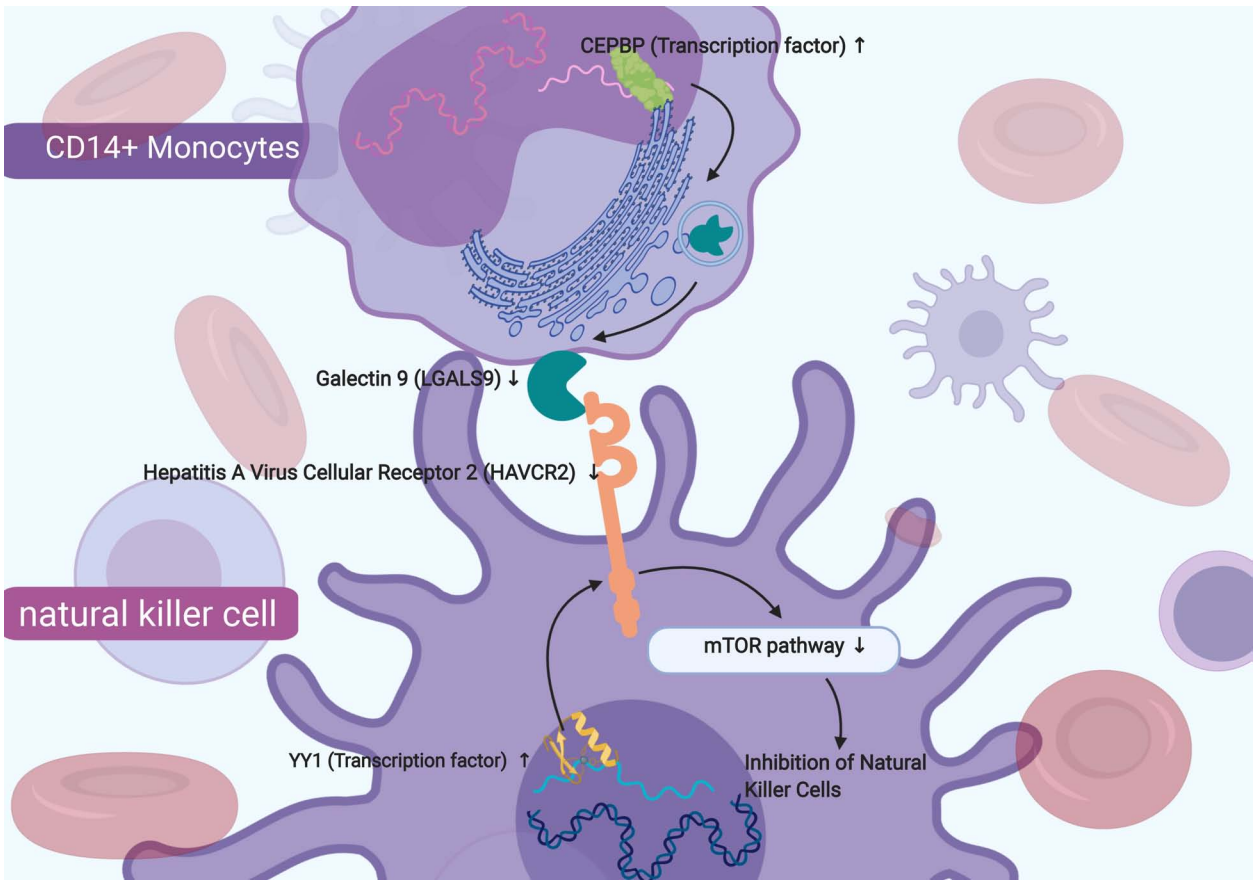


FIGURE 8 | The sketch map of the signaling axis with the most significant co-expression pattern including YY1 (Yin and Yang 1 Transcription Factor), Hepatitis A Virus Cellular Receptor 2 (HAVCR2) and mTOR signaling pathway. In conclusion, this study inferred that the mechanism of YY1 regulating HAVCR2 and mTOR signaling pathway in the NK cells and the cellular communication between NK cells and CD14 + monocytes might play an important role in chronic phase of SCI and neuropathic pain.

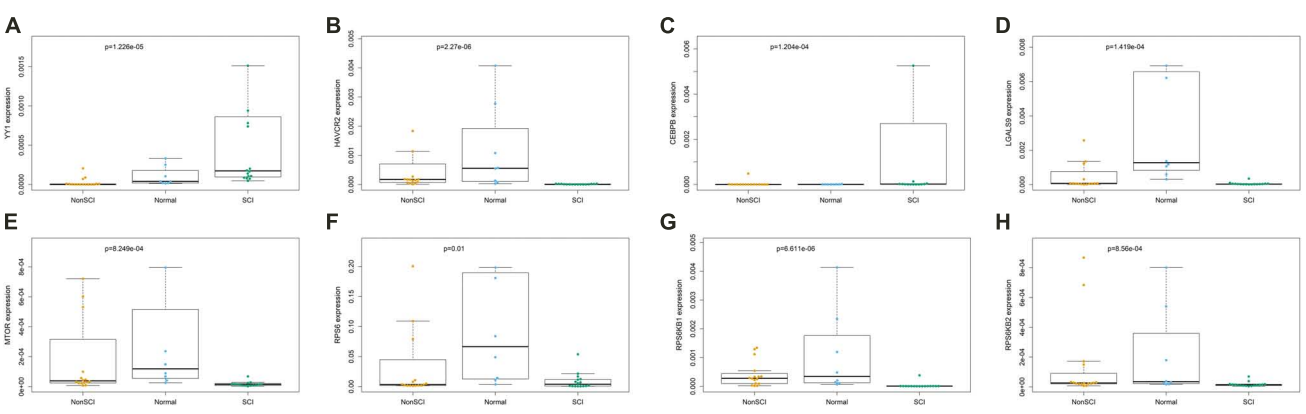


FIGURE 9 | The results of Kruskal-Wallis test identifying the statistical difference of gene expression estimated by Reverse Transcription Quantitative Real-Time PCR (RT-qPCR) Assays. Total RNA was isolated from was extracted from human whole blood of 16 patients with fractures complicated with SCI, 16 patients with fractures but no SCI and 8 normal adults. The results of Kruskal–Wallis test suggested that transcription factor YY1 (**A**, $P < 0.001$) and CEBPB (**C**, $P < 0.001$) were upregulated in the peripheral blood of patients with SCI compared with patients with fractures but no SCI and normal adults. HAVCR2 (**B**, $P < 0.001$) and LGALS9 (**D**, $P < 0.001$) were also abnormally downregulated in peripheral blood of patients with SCI. Some key genes of the mTOR signaling pathway (MTOR, RPS6, RPS6KB1, RPS6KB2) were also identified to be significantly down-regulated in peripheral blood of patients with SCI (**E–H**).

SCI-related NeP, and are often viewed as important predictors (Natinal Spinal Cord Injury Statistical Center [NSCISC], 2014). Thus, the DEGs and cellular communications in peripheral blood attract our interest, which is seldom reported by previous studies.

In the present study, an integrated transcriptome bioinformatics analysis based on bulk RNA sequence and single-cell RNA sequence was performed and the results inferred that the mechanism of YY1 regulating HAVCR2 and the downstream mTOR signaling pathway in the NK cells might be associated with SCI-related NeP. In addition, the cellular communication between NK cells and CD14+ monocytes might also play an important role in the SCI-related NeP. This hypothetical signaling axis might provide prognostic biomarkers and therapeutic targets for the SCI-related NeP.

YY1 was upregulated in the peripheral blood of patients with SCI in our study and almost distributed in all kinds of PBMCs. As a ubiquitously distributed TF, YY1 is involved in activating and repressing a diverse number of promoters. In addition, it modulates a variety of biological processes, particular in nerve and immune cells/tissues (Chen et al., 2017). NF-κB/YY1 signaling was reported to be associated with microglial activation in the progression of glaucoma, characterized with the progressive loss of retinal ganglion cells and optic nerve fibers (Lv et al., 2016). YY1 was expressed initially in pro-myelinating Sox9 + /Sox10 + Schwann cells (SCs) by E18.5 and continued to express in the early postnatal and adult SCs. Following acute nerve injury, YY1 expression was often maintained (Balakrishnan et al., 2016). In Stratton et al.'s (2017) study, YY1 was regarded as SC-associated proteins to promote axonal growth and regenerated axons and formed myelin following transplantation into the injured mouse sciatic nerve. Thus, we supposed that YY1 was a key TF in the SCI-related NeP.

HAVCR2 is also named T-Cell Immunoglobulin Mucin Family Member 3 (Tim-3). In this study, it was abnormally down-regulated in peripheral blood of patients with SCI and significantly correlated with the occurrence of SCI-related NeP. The protein coding by HAVCR2 belongs to the immunoglobulin superfamily and is involved in regulating innate and adaptive immune responses, usually mediating inhibition of target immune cells (Gorman and Colgan, 2014). HAVCR2 was reported to be abnormally expressed in T-cell lymphoma, acute myeloid leukemia, hepatitis A and injured nerve tissue. It regulated the activity of target immune cells through NF-κB signaling pathway, mTOR signaling pathway and RET signaling pathway (Prokhorov et al., 2015; Goncalves Silva et al., 2017; Avery et al., 2018). The role of HAVCR2 in nerve injury was shown in patients with spontaneous intracerebral hemorrhage whose increased HAVCR2 expression on CD14+ monocytes was associated with systemic inflammatory response and sub-acute brain injury (Xu et al., 2018). Besides our results, several previous studies revealed the close interaction between HAVCR2 and LGALS9. The interaction could inhibit the activity and promote the apoptosis of target cells, especially to the immune cells (Zhu et al., 2005; Clayton et al., 2014).

In our study, HAVCR2 was mainly distributed in NK cells in patients with SCI-related NeP, similar to the previous study which reported that HAVCR2 expressed on the surface NK cells was shown to act as a co-receptor to enhance IFN-gamma production in response to LGALS9 (Gleason et al., 2012). NK cells, originated from bone marrow derived lymphocytes, are crucial for immunoreaction against several infections and cancers (Nair et al., 2015). Post-SCI immunological changes impede neurological recovery and mediate common medical consequences of SCI, including NeP (Herman et al., 2018). It was also reported that NK cells were involved in peroneal nerve and their activation was essential in patients with traumatic SCI (Turker et al., 2012; Laginha et al., 2016; Xu et al., 2019). NK cells were observed a significant activation within 24 h after traumatic SCI regarding to the NK cell frequency and the presence of NK cells with the activated phenotype (Xu et al., 2019). During the post-acute and sub-acute phases after SCI, the function of NK cells was impaired (Laginha et al., 2016) and a marked downregulation of NK cell genes was found during chronic SCI (Herman et al., 2018). Therefore, we speculated that in peripheral blood of chronic SCI, HAVCR2 might act as a key receptor on the surface of NK cells and interact with ligand LGALS9 secreted by CD14+ monocytes, inhibiting NK cells through mTOR signaling pathway and ultimately predicting the occurrence of SCI-related NeP.

Furthermore, cell-to-cell communication across multiple cell types and tissues strictly governs proper functions of metazoans and extensively depends on the interactions between ligands and receptors (Hiramoto et al., 1993; Zhou et al., 2017; Cohen et al., 2018; Kumar, 2018; Mukherjee et al., 2018; Zhou et al., 2018). The specific communication utilized by the NK cell system and central nervous system results in conditioned response (Hiramoto et al., 1993). NK cells can engage the homotypic NK-to-NI cell interactions for optimal survival, activation and proliferation (Kim et al., 2014). However, the specific molecular mechanism utilized by the NK cell system and post-SCI central nervous system is not clearly understood. The PPI network is performed based on key genes associated with SCI-related NeP (YY1, HAVCR2, CEBPB, LGALS9), key members of mTOR signaling pathway (MTOR, AKT1, MAPK1, WNT4, PIK3CB) and the surface markers of NK cell (CD56, CD16, CD94, CD3, NKp46) (**Supplementary Figure S3**). Due to the extensive interaction between NK cell's surface markers and mTOR signaling pathway, we hypothesized that mTOR signaling pathway might be associated with the NK cells in the SCI-related NeP.

mTOR plays a crucial role in many physiological functions of the CNS, including the regulation of neuronal cell growth and the development of axon and dendrite (Gong et al., 2015). Its function in SCI are associated with the time phase following SCI (Pouw et al., 2011). With regard to the acute phase of SCI, the mTOR signaling pathway participates in the regulation of neuronal apoptosis, autophagy, activation of macrophage/microglia, and local inflammatory response (Kanno et al., 2012). During the chronic phase, mTOR signaling pathway regulate the neuroregeneration and glial scar formation (Kanno et al., 2012). Thus, Rapamycin, an inhibitor of mTOR, is supposed to be a good treatment for SCI by preventing apoptosis of nerve cells (Yuan et al., 2016), promoting axonal regeneration and inhibiting the formation of glial scar (Kanno et al., 2012;

Li et al., 2015). Moreover, mTOR regulates the development and maturation of T, B and NK cells and control the activation of macrophage/microglia (Powell et al., 2012; Xie et al., 2014; Almutairi et al., 2019; Rostamzadeh et al., 2019).

Moreover, in addition to HAVCR2 and LGALS9, the results suggested that ADRB2, PECAM1and LRP1 were also potential biomarkers associated with SCI. Adrenoceptor Beta 2 (ADRB2) in our study was also upregulated in the peripheral blood of patients after SCI and dominately distributed in the NK cells. ADRB2 (Adrenoceptor Beta 2) is a member of the G protein-coupled receptor superfamily. It is most abundantly expressed on the vasculature and modulated the release of nitric oxide and is involved in vascular function. Damage to the vasculature is universal consequences after SCI. Importantly, it has already been shown that ADRB2 agonists have neuroprotective effects and they improve the neurological and functional outcome, such as isoproterenol, salmeterol, and clenbuterol (Junker et al., 2002; Loy et al., 2002; Graumann et al., 2011). Treatment with the ADRB2 agonist can enhance the recovery in rats post-SCI (Zeman et al., 1999; Zeman et al., 2006; Brown et al., 2014). Scholpa et al. (2019) performed using an FDA-approved compound with the ability to be repurposed, reinforcing the potential clinical applicability of their findings and demonstrating the pharmacological activation of ADRB2 receptor for the treatment of SCI. However, it seemed no previous studies that reported the association between PECAM1/LRP1 and SCI. And due to the main findings of this study was that HAVCR2 might act as a key receptor on the surface of NK cells and interact with ligand LGALS9 secreted by CD14 + monocytes, inhibiting NK cells through mTOR signaling pathway and ultimately predicting the occurrence of SCI-related NeP. Thus, PECAM1/LRP1 were not discussed in details.

There are several unavoidable limitations in our study that should be taken into consideration. Firstly, although the results of the bioinformatics perdition suggested that the DNA binding domain (DDD) of YY1 could bind the promoter region of HAVCR2, no previous studies proved this interaction relationship by the direct mechanism assays. Secondly, the data released in public datasets are so limited that the clinicopathological features analyzed are not comprehensive, which might lead to potential statistical bias. Thirdly, due to the rapid progress of sequencing technology, there is heterogeneity between different batches and experimental platforms. Lastly, we must admit that there are two major limitation in this study, which are the bias between different platforms in the expression profile data and the absence of PBMC single-cell sequencing data of SCI patients. As the single-cell sequencing data originate from normal PBMCs, the results cannot well reflect the pathological changes of PBMCs following SCI. Therefore, a more comprehensive study is being conducting in our lab with data including bulk-RNA-seq and single-cell RNA-seq data of peripheral blood from patients with different time sequence SCI, and single-cell sequencing data of normal and injured spinal cord tissues in mice, which can not only validate the stability of the results in this study but also identify more biomarkers and therapeutic targets for SCI.

CONCLUSION

In peripheral blood of chronic SCI, HAVCR2 might act as a key receptor on the surface of NK cells and interact with ligand LGALS9 secreted by CD14+ monocytes, inhibiting NK cells through mTOR signaling pathway and ultimately predicting the occurrence of SCI-related NeP. This hypothetical signaling axis may provide prognostic biomarkers and therapeutic targets for SCI-related NeP.

DATA AVAILABILITY STATEMENT

All datasets for this study are included in the ArrayExpress (E-GEOD-69901, https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-69901/ and E-MTAB-5151, https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5151/), Gene Expression Omnibus (GSE82152, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE82152), Genotype-Tissue Expression Portal (GTEx, https://commonfund.nih.gov/GTEx/) and Sequence Read Archive (SRA) (SRA713577, https://www.ncbi.nlm.nih.gov/sra/SRX4149408[accn]).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Tongji Hospital, Tongji University School of Medicine. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RH, TM, RZ, LZ, DS, HY, ZH, LC, and JZ: conception, design, data analysis, interpretation, manuscript writing, and final approval of the manuscript. RH, TM, and JZ: collection and/or assembly of data.

FUNDING

We thank the ArrayExpress (E-GEOD-69901 and E-MTAB-5151), Gene Expression Omnibus (GSE82152), Genotype-Tissue Expression Portal (GTEx), and Sequence Read Archive (SRA) (SRA713577) team for using their data.

ACKNOWLEDGMENTS

We thank the ArrayExpress (E-GEOD-69901), Genotype-Tissue Expression Portal (GTEx), and Sequence Read Archive (SRA) (SRA713577) team for using their data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00101/full#supplementary-material

FIGURE S1 | The protein-protein interaction (PPI) network based on the all 2,314 differentially expressed genes (DEGs), which including 4,807 protein-protein interaction relationships related to 799 proteins. PPI, protein-protein interaction; DEG, differentially expressed gene.

FIGURE S2 | The composition **(A)** and heat map **(B)** of immune cells estimated by the next generation Cell type identification by estimating relative subsets of RNA transcripts (CIBERSORTx) algorithm in peripheral blood samples from SCI patients and normal control samples; Enable batch correction and disable quantile normalization algorithm were used to eliminate platform effect and batch effect between different dataset. **(C)** The violin plot identifying immune cells different from the two groups (the blue and red bar stand for SCI group and primary normal control samples, respectively). **(D)** The Principal Component Analysis (PCA) result of all samples suggesting the significant differences between the control group and the experimental group. **(E)** The co-expression heatmap illustrating the co-expression patterns among 22 types of immune cells. CIBERSORT, cell type identification by estimating relative subsets of RNA transcripts; SCI, spinal cord injury; PCA, principal component analysis.

FIGURE S3 | The protein-protein interaction (PPI) network comprised YY1, HAVCR2, CEBPB, LGALS9, key members of mTOR signaling pathway (MTOR, AKT1, MAPK1, WNT4, PIK3CB) and NK cell's surface markers (CD56, CD16, CD94, CD3, NKp46). PPI, protein-protein interaction.

FIGURE S4 | Two volcano plot showing the results of differential expression analysis using two Affy Primeview dataset [GSE82152 **(A)** and E-MTAB-5151 **(B)**] including normal peripheral blood samples as the control group. The standard of DEGs was an absolute log fold change greater than 1 and false discovery rate (FDR) P value < 0.05.

TABLE S1 | The results of differential expression analysis of YY1, CEBPB, HAVCR2, LGALS9, MTOR, RPS6, RPS6KB1, and RPS6KB2 using GSE82152 and E-GEOD-69901 as control group and experimental group, respectively.

TABLE S2 | The results of differential expression analysis of YY1, CEBPB, HAVCR2, LGALS9, MTOR, RPS6, RPS6KB1, and RPS6KB2 using GSE82152 and E-GEOD-69901 as control group and experimental group, respectively.

REFERENCES

Almutairi, S. M., Ali, A. K., He, W., Yang, D. S., Ghorbani, P., Wang, L., et al. (2019). Interleukin-18 up-regulates amino acid transporters and facilitates amino acid-induced mTORC1 activation in natural killer cells. *J. Biol. Chem.* 294, 4644–4655. doi: 10.1074/jbc.RA118.005892

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* 20, 163–172. doi: 10.1038/s41590-018-0276-y

Avery, L., Filderman, J., Szymczak-Workman, A. L., and Kane, L. P. (2018). Tim-3 co-stimulation promotes short-lived effector T cells, restricts memory precursors, and is dispensable for T cell exhaustion. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2455–2460. doi: 10.1073/pnas.1712107115

Azarhomayoun, A., Aghasi, M., Mousavi, N., Shokraneh, F., Vaccaro, A. R., Haj Mirzaian, A., et al. (2018). Mortality rate and predicting factors of traumatic thoracolumbar spinal cord injury; a systematic review and meta-analysis. *Bull. Emerg. Trauma* 6, 181–194. doi: 10.29252/beat-060301

Balakrishnan, A., Stykel, M. G., Touahri, Y., Stratton, J. A., Biernaskie, J., and Schuurmans, C. (2016). Temporal analysis of gene expression in the murine schwann cell lineage and the acutely injured postnatal nerve. *PLoS One* 11:e0153256. doi: 10.1371/journal.pone.0153256

Brienza, D., Krishnan, S., Karg, P., Sowa, G., and Allegretti, A. L. (2018). Predictors of pressure ulcer incidence following traumatic spinal cord injury: a secondary analysis of a prospective longitudinal study. *Spinal Cord* 56, 28–34. doi: 10.1038/sc.2017.96

Brown, A., Nabel, A., Oh, W., Etlinger, J. D., and Zeman, R. J. (2014). Perfusion imaging of spinal cord contusion: injury-induced blockade and partial reversal by beta2-agonist treatment in rats. *J. Neurosurg. Spine* 20, 164–171. doi: 10.3171/2013.10.SPINE13113

Bryce, T. N., Biering-Sorensen, F., Finnerup, N. B., Cardenas, D. D., Defrin, R., Lundeberg, T., et al. (2012). International spinal cord injury pain classification: part I. Background and description. March 6-7, 2009. *Spinal Cord* 50, 413–417. doi: 10.1038/sc.2011.156

Budh, C. N., and Osteraker, A. L. (2007). Life satisfaction in individuals with a spinal cord injury and pain. *Clin. Rehabil.* 21, 89–96. doi: 10.1177/0269215506070313

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, F., Sun, H., Zhao, Y., and Wang, H. (2017). YY1 in cell differentiation and tissue development. *Crit. Rev. Oncog.* 22, 131–141. doi: 10.1615/CritRevOncog.2017021311

Chen, K., Deng, S., Lu, H., Zheng, Y., Yang, G., Kim, D., et al. (2013). RNA-seq characterization of spinal cord injury transcriptome in acute/subacute phases: a resource for understanding the pathology at the systems level. *PLoS One* 8:e72567. doi: 10.1371/journal.pone.0072567

Cheshire, W. P., Santos, C. C., Massey, E. W., and Howard, J. F. Jr. (1996). Spinal cord infarction: etiology and outcome. *Neurology* 47, 321–330. doi: 10.1212/wnl.47.2.321

Chung, N. C., and Storey, J. D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* 31, 545–554. doi: 10.1093/bioinformatics/btu674

Clayton, K. L., Haaland, M. S., Douglas-Vail, M. B., Mujib, S., Chew, G. M., Ndhlovu, L. C., et al. (2014). T cell Ig and mucin domain-containing protein 3 is recruited to the immune synapse, disrupts stable synapse formation, and associates with receptor phosphatases. *J. Immunol.* 192, 782–791. doi: 10.4049/jimmunol.1302663

Cohen, M., Giladi, A., Gorki, A. D., Solodkin, D. G., Zada, M., Hladik, A., et al. (2018). Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. *Cell* 175, 1031.e18–1044.e18.

Consortium, G. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2019). CellPhoneDB v2.0: inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. *bioRxiv* [Preprint].

Finnerup, N. B., Attal, N., Haroutounian, S., Mcnicol, E., Baron, R., Dworkin, R. H., et al. (2015). Pharmacotherapy for neuropathic pain in adults: a systematic review and meta-analysis. *Lancet Neurol.* 14, 162–173. doi: 10.1016/S1474-4422(14)70251-0

Franzen, O., Gan, L. M., and Bjorkegren, J. L. M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* 2019:baz046. doi: 10.1093/database/baz046

Freytag, S., Tian, L., Lonnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Reserach* 7:1297. doi: 10.12688/f1000research.15809.2

Gleason, M. K., Lenvik, T. R., Mccullar, V., Felices, M., O'brien, M. S., Cooley, S. A., et al. (2012). Tim-3 is an inducible human natural killer cell receptor that enhances interferon gamma production in response to galectin-9. *Blood* 119, 3064–3072. doi: 10.1182/blood-2011-06-360321

Goncalves Silva, I., Yasinska, I. M., Sakhnevych, S. S., Fiedler, W., Wellbrock, J., Bardelli, M., et al. (2017). The Tim-3-galectin-9 secretory pathway is involved in the immune escape of human acute myeloid leukemia cells. *EBioMedicine* 22, 44–57. doi: 10.1016/j.ebiom.2017.07.018

Gong, X., Zhang, L., Huang, T., Lin, T. V., Miyares, L., Wen, J., et al. (2015). Activating the translational repressor 4E-BP or reducing S6K-GSK3beta activity

prevents accelerated axon growth induced by hyperactive mTOR *in vivo*. *Hum. Mol. Genet.* 24, 5746–5758. doi: 10.1093/hmg/ddv295

Gorman, J. V., and Colgan, J. D. (2014). Regulation of T cell responses by the receptor molecule Tim-3. *Immunol. Res.* 59, 56–65. doi: 10.1007/s12026-014-8524-1

Graumann, U., Ritz, M. F., and Hausmann, O. (2011). Necessity for re-vascularization after spinal cord injury and the search for potential therapeutic options. *Curr. Neurovasc. Res.* 8, 334–341. doi: 10.2174/156720211798121007

Herman, P., Stein, A., Gibbs, K., Korsunsky, I., Gregersen, P., and Bloom, O. (2018). Persons with chronic spinal cord injury have decreased natural killer cell and increased toll-like receptor/inflammatory gene expression. *J. Neurotrauma* 35, 1819–1829. doi: 10.1089/neu.2017.5519

Hiramoto, R., Ghanta, V., Solvason, B., Lorden, J., Hsueh, C. M., Rogers, C., et al. (1993). Identification of specific pathways of communication between the CNS and NK cell system. *Life Sci.* 53, 527–540. doi: 10.1016/0024-3205(93)90709-c

Hou, R., Denisenko, E., and Forrest, A. R. R. (2019). scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* 35, 4688–4695. doi: 10.1093/bioinformatics/btz292

Junker, V., Becker, A., Huhne, R., Zembatov, M., Ravati, A., Culmsee, C., et al. (2002). Stimulation of beta-adrenoceptors activates astrocytes and provides neuroprotection. *Eur. J. Pharmacol.* 446, 25–36. doi: 10.1016/s0014-2999(02)01814-9

Kanno, H., Ozawa, H., Sekiguchi, A., Yamaya, S., Tateda, S., Yahata, K., et al. (2012). The role of mTOR signaling pathway in spinal cord injury. *Cell Cycle* 11, 3175–3179. doi: 10.4161/cc.21262

Kim, T. J., Kim, M., Kim, H. M., Lim, S. A., Kim, E. O., Kim, K., et al. (2014). Homotypic NK cell-to-cell communication controls cytokine responsiveness of innate immune NK cells. *Sci. Rep.* 4:7157. doi: 10.1038/srep07157

Kumar, P. (2018). Future biomaterials for enhanced cell-substrate communication in spinal cord injury intervention. *Future Sci. OA* 4:Fso268. doi: 10.4155/fsoa-2017-0130

Laginha, I., Kopp, M. A., Druschel, C., Schaser, K. D., Brommer, B., Hellmann, R. C., et al. (2016). Natural killer (NK) cell functionality after human spinal cord injury (SCI): protocol of a prospective, longitudinal study. *BMC Neurol.* 16:170. doi: 10.1186/s12883-016-0681-5

Lee, J. M., Yan, P., Xiao, Q., Chen, S., Lee, K. Y., Hsu, C. Y., et al. (2008). Methylprednisolone protects oligodendrocytes but not neurons after spinal cord injury. *J. Neurosci.* 28, 3141–3149. doi: 10.1523/JNEUROSCI.5547-07.2008

Li, C. Y., Li, X., Liu, S. F., Qu, W. S., Wang, W., and Tian, D. S. (2015). Inhibition of mTOR pathway restrains astrocyte proliferation, migration and production of inflammatory mediators after oxygen-glucose deprivation and reoxygenation. *Neurochem. Int.* 83-84, 9–18. doi: 10.1016/j.neuint.2015.03.001

Loy, D. N., Crawford, C. H., Darnall, J. B., Burke, D. A., Onifer, S. M., and Whittemore, S. R. (2002). Temporal progression of angiogenesis and basal lamina deposition after contusive spinal cord injury in the adult rat. *J. Comp. Neurol.* 445, 308–324. doi: 10.1002/cne.10168

Lv, B., Huo, F., Zhu, Z., Xu, Z., Dang, X., Chen, T., et al. (2016). Crocin upregulates CX3CR1 expression by suppressing NF-kappaB/YY1 signaling and inhibiting lipopolysaccharide-induced microglial activation. *Neurochem. Res.* 41, 1949–1957. doi: 10.1007/s11064-016-1905-1

Metz, G. A., Curt, A., Van De Meent, H., Klusman, I., Schwab, M. E., and Dietz, V. (2000). Validation of the weight-drop contusion model in rats: a comparative study of human spinal cord injury. *J. Neurotrauma* 17, 1–17. doi: 10.1089/neu.2000.17.1

Mukherjee, S., Fried, A., Hussaini, R., White, R., Baidoo, J., Yalamanchi, S., et al. (2018). Phytosomal curcumin causes natural killer cell-dependent repolarization of glioblastoma (GBM) tumor-associated microglia/macrophages and elimination of GBM and GBM stem cells. *J. Exp. Clin. Cancer Res.* 37:168. doi: 10.1186/s13046-018-0792-5

Musubire, A. K., Meya, D. B., Katabira, E. T., Meyer, A. C. L., Bohjanen, P. R., Boulware, D. R., et al. (2019). Epidemiology of non-traumatic spinal cord injury in Uganda: a single center, prospective study with MRI evaluation. *BMC Neurol.* 19:10. doi: 10.1186/s12883-019-1236-3

Nair, S., Fang, M., and Sigal, L. J. (2015). The natural killer cell dysfunction of aged mice is due to the bone marrow stroma and is not restored by IL-15/IL-15Ralpha treatment. *Aging Cell* 14, 180–190. doi: 10.1111/acel.12291

Natinal Spinal Cord Injury Statistical Center [NSCISC], (2014). Spinal cord injury facts and figures at a glance. *J. Spinal Cord Med.* 37, 355–356.

New, P. W., Cripps, R. A., and Bonne Lee, B. (2014). Global maps of non-traumatic spinal cord injury epidemiology: towards a living data repository. *Spinal Cord* 52, 97–109. doi: 10.1038/sc.2012.165

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

Pinzon, A., Marcillo, A., Quintana, A., Stamler, S., Bunge, M. B., Bramlett, H. M., et al. (2008). A re-assessment of minocycline as a neuroprotective agent in a rat spinal cord contusion model. *Brain Res.* 1243, 146–151. doi: 10.1016/j.brainres.2008.09.047

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

Powell, J. D., Pollizzi, K. N., Heikamp, E. B., and Horton, M. R. (2012). Regulation of immune responses by mTOR. *Annu. Rev. Immunol.* 30, 39–68. doi: 10.1146/annurev-immunol-020711-075024

Prokhorov, A., Gibbs, B. F., Bardelli, M., Ruegg, L., Fasler-Kan, E., Varani, L., et al. (2015). The immune receptor Tim-3 mediates activation of PI3 kinase/mTOR and HIF-1 pathways in human myeloid leukaemia cells. *Int. J. Biochem. Cell Biol.* 59, 11–20. doi: 10.1016/j.biocel.2014.11.017

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

Rostamzadeh, D., Yousefi, M., Haghshenas, M. R., Ahmadi, M., Dolati, S., and Babaloo, Z. (2019). mTOR Signaling pathway as a master regulator of memory CD8(+) T-cells, Th17, and NK cells development and their functional properties. *J. Cell. Physiol.* 234, 12353–12368. doi: 10.1002/jcp.28042

Scholpa, N. E., Williams, H., Wang, W., Corum, D., Narang, A., Tomlinson, S., et al. (2019). Pharmacological stimulation of mitochondrial biogenesis using the food and drug administration-approved beta2-adrenoreceptor agonist formoterol for the treatment of spinal cord injury. *J. Neurotrauma* 36, 962–972. doi: 10.1089/neu.2018.5669

Siddall, P. J., Mcclelland, J. M., Rutkowski, S. B., and Cousins, M. J. (2003). A longitudinal study of the prevalence and characteristics of pain in the first 5 years following spinal cord injury. *Pain* 103, 249–257. doi: 10.1016/s0304-3959(02)00452-9

Stratton, J. A., Kumar, R., Sinha, S., Shah, P., Stykel, M., Shapira, Y., et al. (2017). Purification and characterization of schwann cells from adult human skin and nerve. *eNeuro* 4:ENEURO.307-16.2017. doi: 10.1523/ENEURO.0307-16.2017

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

Takeura, N., Nakajima, H., Watanabe, S., Honjoh, K., Takahashi, A., and Matsumine, A. (2019). Role of macrophages and activated microglia in neuropathic pain associated with chronic progressive spinal cord compression. *Sci. Rep.* 9:15656. doi: 10.1038/s41598-019-52234-1

Turker, B., Uz, B., Isik, M., Bektas, O., Demiroglu, H., Sayinalp, N., et al. (2012). Nasal natural killer/T-cell lymphoma with skin, eye, and peroneal nerve involvement. *Turk. J. Haematol.* 29, 413–419. doi: 10.5505/tjh.2012.03360

Van Niekerk, E. A., Tuszynski, M. H., Lu, P., and Dulin, J. N. (2016). Molecular and cellular mechanisms of axonal regeneration after spinal cord injury. *Mol. Cell. Proteom.* 15, 394–408. doi: 10.1074/mcp.R115.053751

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., et al. (2018). Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353. doi: 10.1038/s41586-018-0698-6

Werhagen, L., Budh, C. N., Hultling, C., and Molander, C. (2004). Neuropathic pain after traumatic spinal cord injury–relations to gender, spinal level,

completeness, and age at the time of injury. *Spinal Cord* 42, 665–673. doi: 10.1038/sj.sc.3101641

Xie, L., Sun, F., Wang, J., Mao, X., Xie, L., Yang, S. H., et al. (2014). mTOR signaling inhibition modulates macrophage/microglia-mediated neuroinflammation and secondary injury via regulatory T cells after focal ischemia. *J. Immunol.* 192, 6009–6019. doi: 10.4049/jimmunol.1303492

Xu, C., Ge, H., Wang, T., Qin, J., Liu, and Liu, Y. (2018). Increased expression of T cell immunoglobulin and mucin domain 3 on CD14(+) monocytes is associated with systemic inflammatory reaction and brain injury in patients with spontaneous intracerebral hemorrhage. *J. Stroke Cerebrovasc. Dis.* 27, 1226–1236. doi: 10.1016/j.jstrokecerebrovasdis.2017.11.041

Xu, L., Zhang, Y., Zhang, R., Zhang, H., Song, P., Ma, T., et al. (2019). Elevated plasma BDNF levels are correlated with NK cell activation in patients with traumatic spinal cord injury. *Int. Immunopharmacol.* 74:105722. doi: 10.1016/j.intimp.2019.105722

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, Y., Wang, Y., Hu, F. F., Jiang, C. Y., Zhang, Y. J., Yang, J. L., et al. (2016). Cadmium activates reactive oxygen species-dependent AKT/mTOR and mitochondrial apoptotic pathways in neuronal cells. *Biomed. Environ. Sci.* 29, 117–126. doi: 10.3967/bes2016.013

Yune, T. Y., Lee, J. Y., Jung, G. Y., Kim, S. J., Jiang, M. H., Kim, Y. C., et al. (2007). Minocycline alleviates death of oligodendrocytes by inhibiting pro-nerve growth factor production in microglia after spinal cord injury. *J. Neurosci.* 27, 7751–7761. doi: 10.1523/jneurosci.1661-07.2007

Zeman, R. J., Feng, Y., Peng, H., and Etlinger, J. D. (1999). Clenbuterol, a beta(2)-adrenoceptor agonist, improves locomotor and histological outcomes after spinal cord contusion in rats. *Exp. Neurol.* 159, 267–273. doi: 10.1006/exnr.1999.7146

Zeman, R. J., Peng, H., Feng, Y., Song, H., Liu, X., and Etlinger, J. D. (2006). Beta2-adrenoreceptor agonist-enhanced recovery of locomotor function after spinal cord injury is glutathione dependent. *J. Neurotrauma* 23, 170–180. doi: 10.1089/neu.2006.23.170

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

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

Zhou, X., Franklin, R. A., Adler, M., Jacox, J. B., Bailis, W., Shyer, J. A., et al. (2018). Circuit design features of a stable two-cell system. *Cell* 172, 744.e17–757.e17. doi: 10.1016/j.cell.2018.01.015

Zhu, C., Anderson, A. C., Schubart, A., Xiong, H., Imitola, J., Khoury, S. J., et al. (2005). The Tim-3 ligand galectin-9 negatively regulates T helper type 1 immunity. *Nat. Immunol.* 6, 1245–1252. doi: 10.1038/ni1271

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 © 2020 Huang, Meng, Zhu, Zhao, Song, Yin, 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fbioe.2020.00031

Transcriptomic Analyses for Identification and Prioritization of Genes Associated With Alzheimer's Disease in Humans

Yuchen Shi[1†], Hui Liu[1†], Changbo Yang[1†], Kang Xu[1†], Yangyang Cai[1], Zhao Wang[1], Zheng Zhao[2], Tingting Shao[1*] and Yixue Li[1*]*

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] Beijing Neurosurgical Institute, Capital Medical University, Beijing, China

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Yucheng Yang,
Yale University, United States
Qi Zhao,
Shenyang Aerospace University, China

***Correspondence:**
Yixue Li
biolyx@163.com
Tingting Shao
shaotingting_05@163.com
Zheng Zhao
zhaozheng0503@sina.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *11 October 2019*
Accepted: *14 January 2020*
Published: *21 February 2020*

Citation:
Shi Y, Liu H, Yang C, Xu K, Cai Y, Wang Z, Zhao Z, Shao T and Li Y (2020) Transcriptomic Analyses for Identification and Prioritization of Genes Associated With Alzheimer's Disease in Humans. Front. Bioeng. Biotechnol. 8:31. doi: 10.3389/fbioe.2020.00031

Long non-coding RNAs (lncRNAs), as important ncRNA regulators, play crucial roles in the regulation of various biological processes, and their aberrant expression is related to the occurrence and development of diseases, which is gradually validated by more and more studies. Alzheimer's disease (AD) is a chronic neurodegenerative disease that often develops slowly and gradually deteriorates over time. However, which functions the lncRNAs perform in AD are almost unknown. In this study, we performed transcriptome analysis in AD, containing 12,892 known lncRNAs and 19,053 protein-coding genes (PCGs). Further, 14 down-regulated and 39 up-regulated lncRNAs were identified, compared with normal brain samples, which indicated that these lncRNAs might play critical roles in the pathogenesis of AD. In addition, 19 down-regulated and 28 up-regulated PCGs were also detected. Using the differentially expressed lncRNAs and PCGs through the WGCNA method, an lncRNA–mRNA co-expressed network was constructed. The results showed that lncRNAs RP3-522J7, MIR3180-2, and MIR3180-3 were frequently co-expressed with known AD risk PCGs. Interestingly, PCGs in the network are significantly enriched in brain- or AD-related biological functions, including the brain renin–angiotensin system, cell adhesion, neuroprotective role of THOP1 in AD, and so on. Furthermore, it was shown that 18 lncRNAs and 7 PCGs were highly expressed in normal brain tissue relative to other normal tissue types, suggesting their potential as diagnostic markers of AD, especially RP3-522J7, MIR3180-2, MIR3180-3, and CTA-929C8. In total, our study identified a compendium of AD-related dysregulated lncRNAs and characterized the corresponding biological functions of these lncRNAs in AD, which will be helpful to understand the molecular basis and pathogenesis of AD.

Keywords: long non-coding RNA, Alzheimer's disease, transcriptomic analyses, RNA-seq, differential expression analysis

INTRODUCTION

Alzheimer's disease (AD) is an insidious progressive neurodegenerative disease, which is affected by many factors, including environmental factors and genetic and epigenetic variations (Toledo et al., 2015), but there is no doubt that aging is the biggest risk factor (Kold-Christensen and Johannsen, 2019). As the global population ages, there will be two billion people aged 60 years or over in the

world by 2050 (Harper, 2014), and two new patients will be diagnosed with AD every minute, resulting in one million new patients who will be diagnosed every year (Alzheimer's Association, 2015). Although the pathophysiology of AD still remains unclear (Xu et al., 2018), and even if there are no effective treatment methods currently, it is generally accepted that only the earliest intervention is likely to affect the progression of the disease (Sood et al., 2015).

With the development of life science research, it has been found that the human transcriptome is more complicated than previously known. Indeed, a considerable part of the human genome regions could transcribe as long non-coding RNAs (lncRNAs), a set of non-coding transcripts longer than 200 nt, which play important roles in various life activities, such as embryonic development, cell differentiation, aging, and complex diseases (Derrien et al., 2012; Rinn and Chang, 2012). Many studies have proved that lncRNAs are involved in many physiological and pathological processes through transcriptional or post-transcriptional regulatory mechanisms and thus play an important role in the process of the whole life, so it has become a hot spot of genetic and epigenetic research (Mercer and Mattick, 2013; Papait et al., 2013; Shi et al., 2013; Fatica and Bozzoni, 2014; Sultmann and Diederichs, 2014). As previous research has shown, due to the complex and transcriptional regulation mechanisms, lncRNAs could be focused on their roles in human complex diseases, including AD and cancers, to gain new insights into complex disease pathways, to identify biomarkers to improve diagnostic accuracy, and to examine the impact of treatment (Zhou et al., 2019). High-throughput sequencing technology has significant advantages, providing opportunities for insight into the genomic and transcriptomic research with large data sets, and it could be helpful to dissect the comprehensive transcriptome characterization of complex diseases (Sultan et al., 2008; Xiong et al., 2012; Prasad et al., 2016). However, the role of lncRNAs in the pathogenesis and progression of certain human complex diseases remains unclear, particularly AD. By comparing a ck-p25 AD model and control samples in mice, it was found that histone modification could regulate the differential expression of lncRNAs (Wan et al., 2016). Thus, it is indicated that lncRNAs might be helpful in investigating the transcription landscape of brain tissue in AD.

In this study, we aimed to identify the gene expression patterns in AD patients and controls, identify the lncRNAs involved in the AD-related dysregulated biological processes, and then to prioritize important lncRNAs based on the co-expression networks composed of differentially expressed lncRNAs and protein-coding genes (PCGs) in AD. Our results suggest that dysregulated expressed transcripts could affect complex disease, including lncRNAs and PCGs. Differentially expressed lncRNAs are significantly co-expressed with PCGs, which could be constructed in the co-expressed network to exhibit the complex regulatory relations in AD. Functional analysis revealed that lncRNAs might be involved in the AD-related biological processes, even regulate the AD-related function modules. Summarizing the above, our study provided a new set of AD-related dysregulated lncRNAs and identified the corresponding biological functions of these lncRNAs in AD. It will be helpful for further understanding the pathogenesis for AD.

MATERIALS AND METHODS

RNA-seq Data Analysis

NGS data used in this research were derived from public resource GEO, including RNA-seq data of human cerebral cortex tissues from nine patients and eight controls (GEO accession GSE53697) (Scheckel et al., 2016). The RNA-seq data were mapped into the human reference genome (Hg38 version) by Tophat2 (Kim et al., 2013). For the measurement of expression values for lncRNAs and PCGs, we used Cufflinks v2.1.1 with FPKM (Trapnell et al., 2012).

Differential Expression Analysis

Wilcoxon rank sum test was used to identify the differentially expressed lncRNAs and PCGs between the AD cases and the control group with the threshold of $P < 0.01$ (Han et al., 2019). Classification of lncRNAs and PCGs was according to the gene annotation file derived from GENCODE. The following biotypes are considered as known lncRNAs: "3prime_overlapping_ncrna," "ambiguous_orf," "antisense," and "antisense_RNA," "lincRNA," "ncrna_host," "non-_coding," "non-_stop_decay," "processed_transcript," "retained_intron," "sense_intronic," and "sense_overlapping." And the biotype of "protein_coding" is considered as a PCG.

Collection of Aging-Related Gene Data Set

To investigate the relationship between differentially expressed genes in AD and aging, a list of aging-related genes was collected, which were defined as the genes involved in the development or aging-associated GO biological processes. And the functional annotation for genes and GO terms was derived from the Gene2GO file in NCBI[1]. And then, a cumulative hypergeometric test was used to investigate whether the AD-associated differentially expressed PCGs were enriched in the aging gene set with statistical significance.

Co-expression Network Analysis

As important regulators, lncRNAs are considered to be involved the corresponding biological processes by regulating their target genes. For each differentially expressed lncRNA, we attempted to identify its regulatory target genes. The co-expression network for differentially expressed lncRNAs and PCGs was constructed by the WGCNA method with empirical threshold with the value equal to nine (Langfelder and Horvath, 2008).

Gene Set Enrichment Analysis

To identify biological processes and cellular components which would be regulated by aberrantly expressed lncRNAs and PCGs in AD, enrichment analyses were performed for each lncRNA target

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

PCG using the R package TCGAbiolinks (Colaprico et al., 2016), and GO/KEGG terms with adjusted p value < 0.01 by Benjamini–Hochberg methods were considered.

Brain-Elevated Expression Analysis

The tissue-specific RNA-seq data set was obtained from the Genotype-Tissue Expression (GTEx, 2013) project, in which gene-level average RPKM values are reported for each tissue sample across 30 tissues including the brain (2013). A certain differentially expressed lncRNA or PCG would be defined as brain elevated if this transcription's expression value in brain tissue is more than five times compared with the average value in all other tissues.

AD-Related miRNA

To prioritize AD-related miRNA, we used the miRanda algorithm and searched the starBase database to obtain the mRNA–miRNA targeting relationship (Betel et al., 2010; Li et al., 2014). AD-related miRNAs were collected from the HMDD database, which manually curated the experimentally validated human disease-associated miRNA information (Huang et al., 2019). We analyzed the target relationship between AD-related miRNAs and differentially expressed PCGs and then ranked candidate miRNAs according to the count of the overlap between the differently expressed PCGs targeted by known AD miRNAs and the targets of each miRNA.

RESULTS

Transcriptome Analysis in AD

A total of 17 RNA libraries were prepared from human brain samples, including nine AD and eight controls. From these samples, a total of 1,014,513,141 read pairs were generated from RNA-seq experiments, and >700 million read pairs (72.4%) were aligned to the human genome (Hg38 version). For the measurement of expression levels for lncRNAs and PCGs, we used Cufflinks. The gene expression levels were compared between AD and control samples, including lncRNAs and PCGs. The result revealed that about 12,892 lncRNAs and 19,053 PCGs were expressed. Consistently with previous studies, the expression levels of PCGs were generally higher than those of lncRNAs. In detail, there were just about 60% of lncRNAs with the FPKM value above 0.1, whereas the majority of PCGs were expressed above one (**Figure 1**). Moreover, 54.5% of lncRNAs were expressed in more than 15 samples and 1.6 times enriched for PCGs compared with lncRNAs (**Supplementary Figure S1**). These results revealed that lncRNAs were expressed widely in AD, indicating their potential important roles in the development of disease.

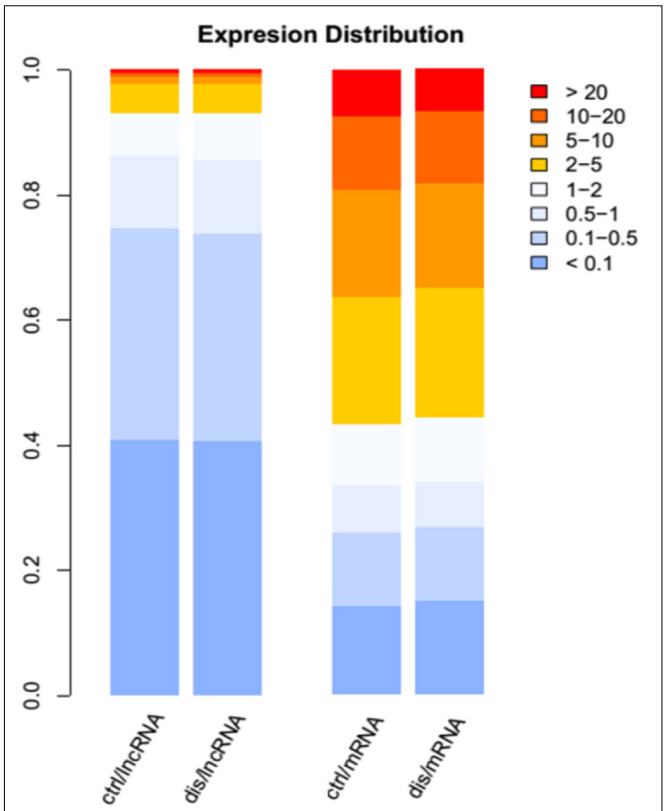


FIGURE 1 | The distribution expression levels for long non-coding RNAs (lncRNAs) and mRNAs in Alzheimer's disease (AD) and control groups. The different color legends represent the expression levels of lncRNAs and protein-coding genes (PCGs); the blue color represents low expression, and red color represents high expression.

Differential Expression of Both lncRNAs and PCGs in AD

Previous studies have confirmed that the variant events of the human brain transcriptome are involved in the pathophysiological process of AD. Thus, the identification of the aberrantly expressed genes in AD patients might be helpful to understand the molecular mechanism of AD (Twine et al., 2011; Bernstein et al., 2016; Hensman Moss et al., 2017). We re-analyzed AD transcriptome data derived from a previous independent study which focused on ELAV-like protein binding to genes in the human brain yet did not analyze differential expression analysis of lncRNAs and PCGs (Scheckel et al., 2016). In this part, we removed the lncRNAs overlapping with PCGs in order to ensure the accuracy of the results. We found 100 differentially expressed genes, including 39 up-regulated lncRNAs and 14 down-regulated lncRNAs, as well as 28 up-regulated PCGs and 19 down-regulated PCGs (p-value < 0.01) (**Supplementary Table S1**). The unsupervised hierarchical clustering results revealed that differentially expressed lncRNAs (**Figure 2A**) and PCGs (**Supplementary Figure S2**) could well distinguish AD samples from control samples, indicating that there might be existing distinct signatures in expression level for AD. Notably, several lncRNAs, including CTA-929C8, RP11-461L13, and PSMG3-AS1, have relatively high expression levels in AD or control samples. For example, the expression value of CTA-929C8 in normal tissue is 0.3665 FPKM but decreases to 0.069 FPKM in AD. Another lncRNA is PSMG3-AS1,

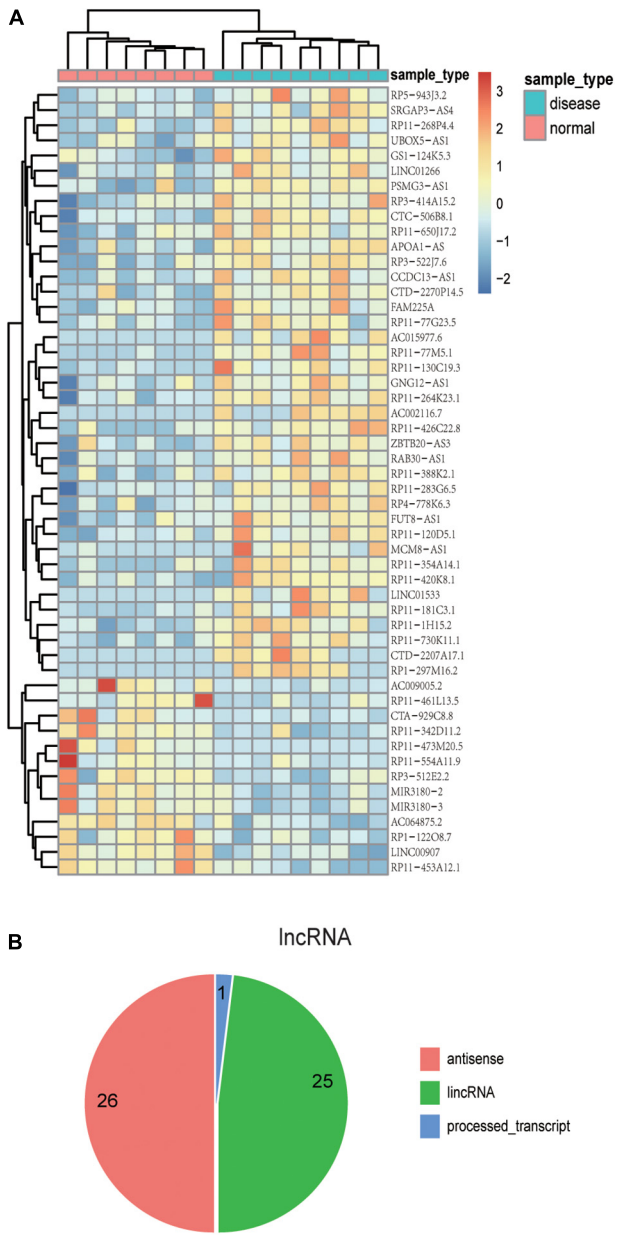


FIGURE 2 | The expression patterns and localization of lncRNAs with differential expression. **(A)** The heat map shows the expression patterns of differentially expressed lncRNAs in both AD and control samples. Rows represent lncRNAs, and columns represent samples. Turquoise color in the top bar marks AD patients, and orange marks normal samples. **(B)** Pie chart displays distinct lncRNA classifications. The majority of lncRNAs are antisense lncRNAs or lincRNAs.

which is up-regulated with the mean expression value of 1.718 FPKM in AD.

Further, among the differentially expressed PCGs in AD, eight genes have been validated to be associated with AD, including VTI1A, CUX1, S100B, AGT, CD44, NTS, IRAK4, and AQP1, which are reported by the DisGeNET database (Pinero et al., 2017). The first two genes are down-regulated, and the others are up-regulated in AD. Moreover, VTI1A was validated to have susceptibility loci for late-onset AD (Grupe et al., 2006), even in glioma, based on a genome-wide association study (Kinnersley et al., 2015). As another example, CD44, a surface antigen expressed across multiple tissues, was up-regulated in AD samples, which corresponds with previous studies (Uberti et al., 2010). The results indicated that CD44 could play crucial roles in driving the immune response into infected tissues in the central nervous system. The aberrantly expressed S100B could seriously affect one of the hallmarks of AD and neuroinflammation; thus, it plays an important role in the pathophysiological process of AD, and the drugs targeting S100B might have a great impact on the treatment of this disease (Cirillo et al., 2015). These results indicate that the differentially expressed lncRNAs and PCGs are significantly associated with AD, which might affect the occurrence and development of AD.

Moreover, we investigated the genomic localization of differentially expressed lncRNAs compared with PCGs. As the results show in **Figure 2B**, 26 differentially expressed lncRNAs are antisense lncRNAs, and 25 are lincRNAs. Next, we computed the distance between differentially expressed lncRNAs and differentially expressed PCGs. As a result, we found that there are several lncRNAs located near important PCGs in the human genome. For example, six differentially expressed lncRNAs are located within 10 MB of S100B, and five of them are up-regulated in AD corresponding with S100B. The intergenic lncRNA, RP3-522J7, is an up-regulated lncRNA in AD with a fold change value of 2.65, compared with control samples, whose genomic distance is 3.3 MB apart from S100B. Therefore, it is concluded that the lncRNA PR3-522J7 might cis-regulate the expression of S100B. Another example is the gene CD44, which had 11 lncRNAs nearby, with a distance less than 10 Mb. Especially among the 11 neighbor lncRNAs, two lncRNAs described above, CTA-929C8 and CCDC13-AS1, are 7.8 and 7.5 MB apart from CD44. Therefore, we concluded that these two lncRNAs might cis-regulate the expression of CD44. These results indicate that some of these differentially expressed lncRNAs might cis-regulate the known AD protein-coding genes to influence the occurrence and development of AD.

Additionally, 92 development or age-related GO terms were collected to obtain a total of 2,435 age-related genes (**Supplementary Table S2**), and then, we found that 11 of these differentially expressed genes were identified as aging-related genes (**Supplementary Table S3**). The results of a cumulative hypergeometric test indicated that differentially expressed genes in diseases were significantly correlated with aging with statistical significance (p = 0.004650242), including AGT and CUX1, which are involved in function "aging" and "multicellular organism development," respectively. Moreover, some differentially expressed genes are labeled as aging or longevity by the HAGR database and some studies. For example, S100B and AGT were labeled as aging genes in the HAGR database, while TOX3 was considered as a longevity gene (Tacutu et al., 2018). Another example is CLUAP1, which has been shown to vary significantly with age (Peters et al., 2015). These results indicate that these differentially expressed PCGs are associated

with aging, confirm the effect of aging on AD, and are consistent with clinical observations.

Co-expression Network Analysis Reveals Differentially Expressed Genes Involved in AD-Related Biological Functions

More and more lncRNAs have been validated to play important roles in disease-associated biological processes (Sun et al., 2014; Zhou et al., 2017); however, there are still a considerable number of lncRNAs uncovered involved in the physiological process during embryonic development and cell differentiation, even in the pathological processes of human complex diseases. Based on the knowledge of the biological functions of the coding gene, the biological processes of lncRNAs could be predicted, with the hypothesis that if the expression of a certain lncRNA is correlated with the lists of PCGs, this gene co-expression pattern may provide reliable evidence that they are involved

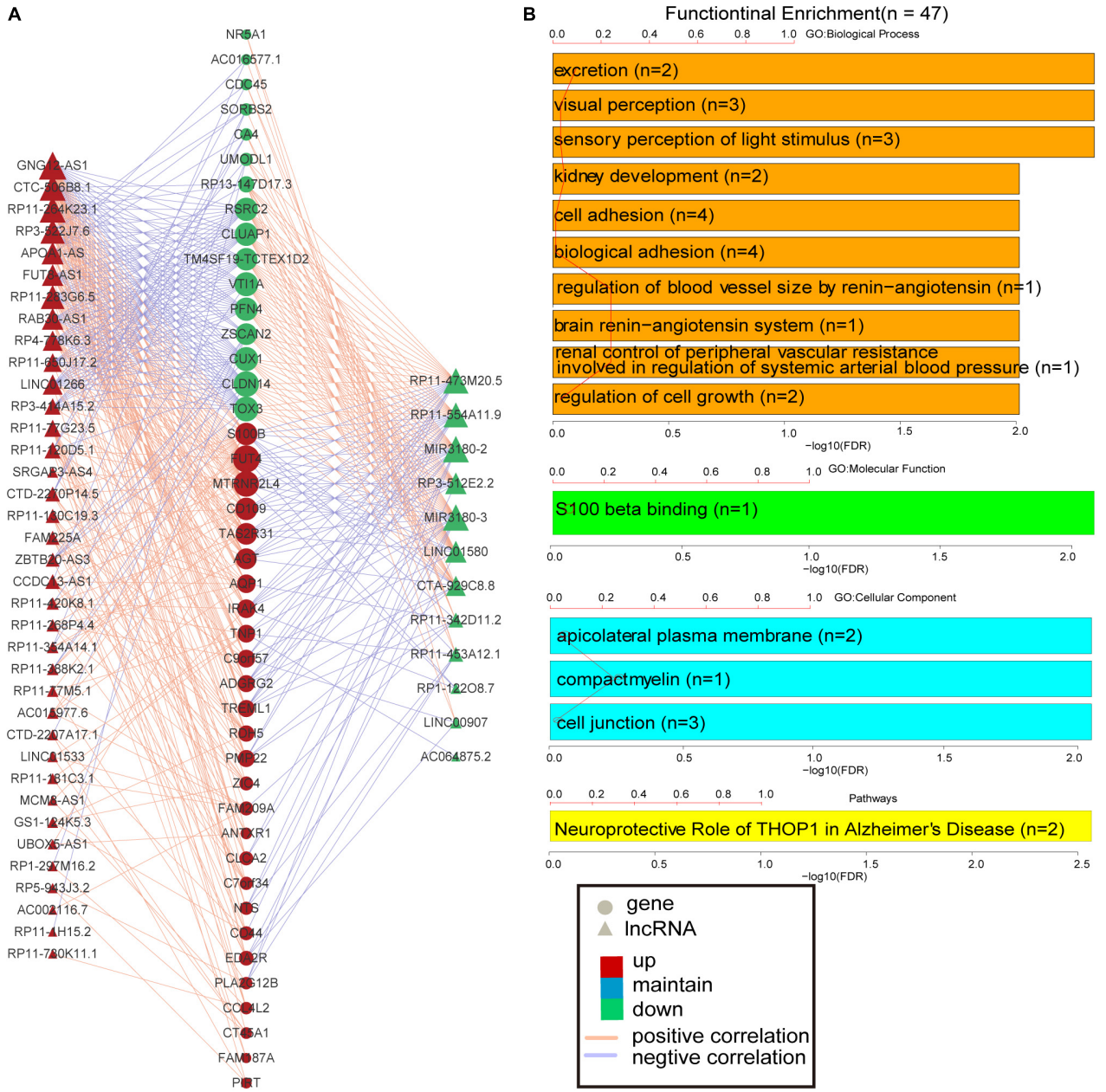


FIGURE 3 | PCG–lncRNA co-expression network and function enrichment analysis. **(A)** The construction of PCG–lncRNA co-expression network. Circle represents gene, and triangle represents lncRNA. The orange line is positive correlation between lncRNA and PCG according to the Pearson correlation coefficient, and the blue color is negative. Red node marks up-regulated entity, and green node marks down-regulated entity. **(B)** Function enrichment analysis of the differentially expressed PCGs in the co-expression network.

in the same or similar biological functions. The lncRNA–PCG co-expression network for differentially expressed lncRNAs and PCGs was constructed via R package WGCNA. As a result, the co-expression network was composed of 489 links between 49 differentially expressed lncRNAs and 44 PCGs (**Figure 3A**). These lncRNAs and PCGs are tightly connected together. Especially, S100B is co-expressed with 18 lncRNAs, including seven down-regulated lncRNAs and 11 up-regulated lncRNAs. For example, the weight value of the link between lncRNA RP3-522J7 and PCG S100B is 0.043867. Moreover, this lncRNA is located within 3.3 Mb from S100B. Therefore, we concluded that RP3-522J7 might cis-regulate the expression of S100B. In addition, we found that MIR3180-2 and MIR3180-3 shared relatively more co-expressed PCGs, including VTI1A, CUX1, S100B, AGT, NTS, and IRAK4, which are known as AD-related PCGs. Both the two lncRNAs are the host genes of hsa-miR-3180-3p and hsa-miR-3180-5p; thus, it was proposed that these two miRNAs might be up-regulated in AD.

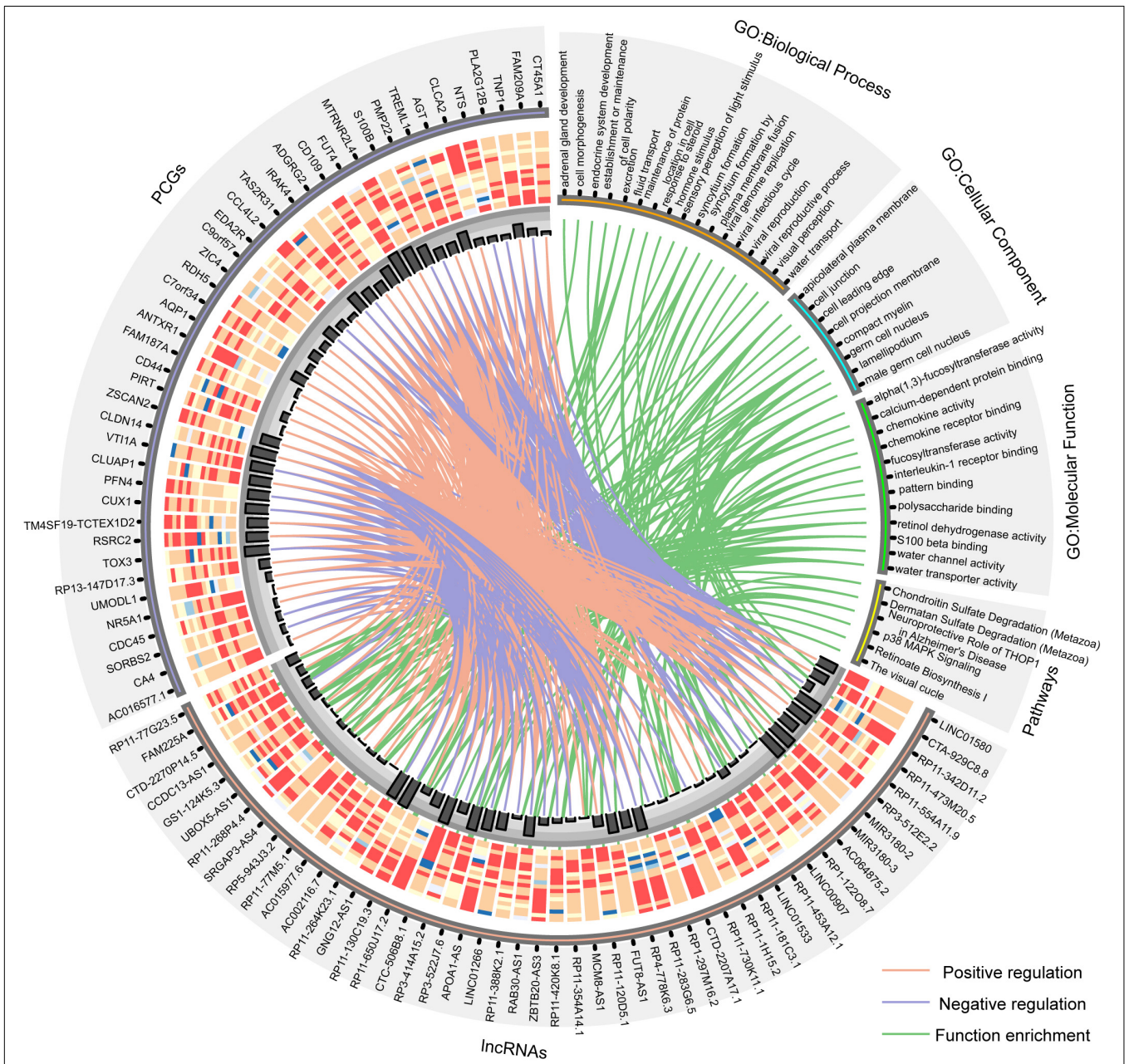


FIGURE 4 | Circos plot shows the relationship between lncRNAs, targeted PCGs, and parts of functions (FDR < 0.005). The orange line is positive correlation between lncRNA and PCG according to the Pearson correlation coefficient, the blue color is negative, and the green lines connect lncRNAs and the functions that their targeted PCGs are enriched in. Bar plot shows the number of target relationships.

Furthermore, the GO and KEGG function enrichment analysis was performed for the differentially expressed PCGs in the co-expression network by the TCGAbiolinks tool (Xiong et al., 2012). **Figure 3B** and **Supplementary Table S4** show the over-represented terms and pathways, among the most enriched of which are those involved in cell adhesion (FDR = 0.00969), regulation of cell growth (FDR = 0.00969), visual perception (FDR = 0.00460), brain renin–angiotensin system (FDR = 0.00969), neuroprotective role of THOP1 in AD (FDR = 0.00275), and so on. It's worth noting that many PCGs targeted by differentially expressed lncRNAs in AD are involved in AD- and aging-related biological functions (**Supplementary Table S5**), such as CTA-929C8 target PCGs involved in the brain renin–angiotensin system (FDR = 0.00969), regulation of inflammatory response (FDR = 0.00969), and regulation of vasoconstriction (FDR = 0.00969) (**Supplementary Figure S3**); RP3-522J7.6 target PCGs involved in the brain renin–angiotensin system (FDR = 0.00944), the renal system process involved in the regulation of systemic arterial blood pressure (FDR = 0.00944), and the neurological system process involved in regulation of systemic arterial blood pressure (FDR = 0.00944) (**Supplementary Figure S4**); and MIR3180-2 and MIR3180-3 target PCGs involved in the neuroprotective role of THOP1 in AD (FDR = 0.00130) and blood vessel remodeling (FDR = 0.00944) (**Supplementary Figures S5, S6**). The relationship between lncRNAs, targeted PCGs, and parts of functions are shown in **Figure 4** (FDR < 0.005). All these categories are associated with brain-related functions and are implicated in AD or aging. Therefore, these differentially expressed lncRNAs might be involved in AD and aging via regulating these molecular functions.

Prioritization of Candidate lncRNAs as Potential Diagnostic Markers of AD

Further, we investigated whether these differentially expressed lncRNAs and PCGs tend to be particularly highly expressed in human cerebral tissues, which might be involved in brain-related biological processes and could be diagnostic markers of AD. We obtained high-throughput data from the GTEx project including 30 human tissues, and then the mean expression of the differentially expressed lncRNAs and PCGs was calculated (2013) (**Figure 5A** and **Supplementary Figure S3**). Eighteen differentially expressed lncRNAs are specifically highly expressed in cerebral tissue, including the lncRNAs described above, such as RP3-522J7, CCDC13-AS1, MIR3180-2, MIR3180-3, CTA-929C8, and so on (**Figure 5A**). On the other hand, there are seven PCGs, including three known AD-related genes, which are S100B, AGT, and NTS (**Supplementary Figure S7**). lncRNA CTA-929C8 is highly expressed in brain tissue, and its expression level was enriched in normal brain tissue, which was more than about 1,000-fold compared with other normal tissues (**Figure 5B**). We concluded that this lncRNA might be a suitable marker in AD. In addition, for the two PCGs, S100B and PIRT, we found that they are also enriched in normal brain tissues and that their expression levels were more than 10 times versus the other tissues (**Supplementary Figure S7**). Moreover, the two genes are up-regulated in AD. Therefore, it is indicated that the brain-specific highly expressed lncRNAs and PCGs might be involved in the brain-related biological processes, and even affect the brain-associated functions, and could be potential diagnostic markers for AD.

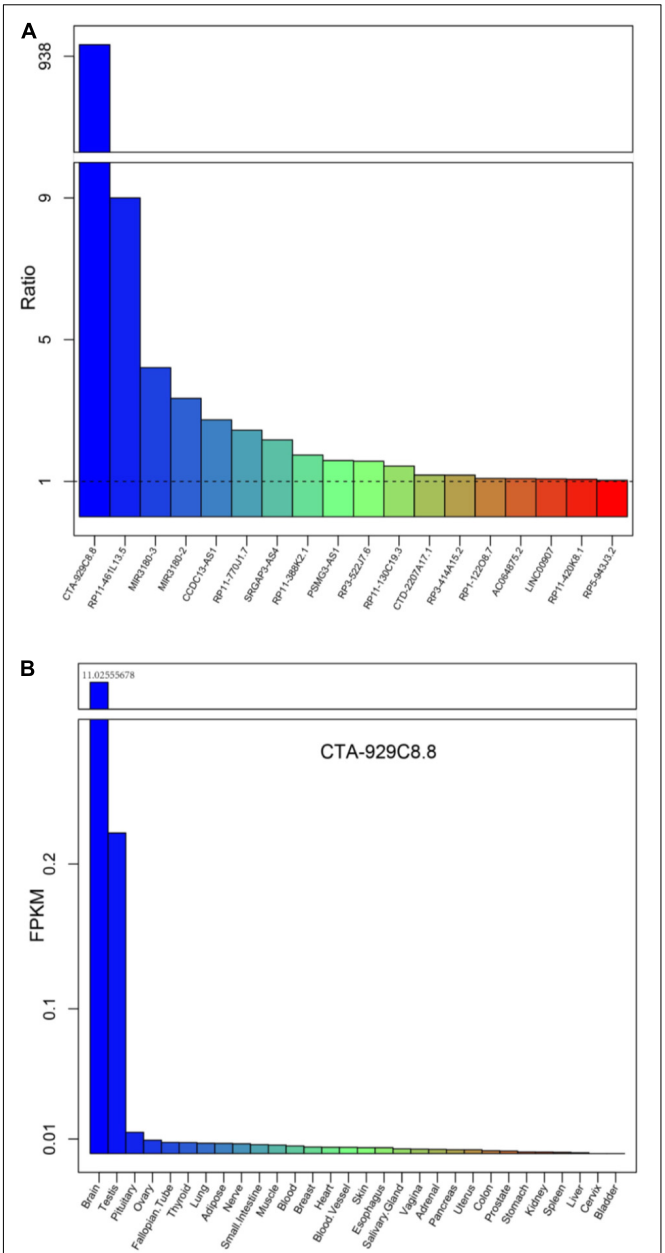


FIGURE 5 | The mean expression levels of the differentially expressed lncRNAs for each tissue from the Genotype-Tissue Expression (GTEx) project. **(A)** The expression levels of the 18 lncRNAs up-regulated, expressed only in brain tissues, compared with the other tissue types are listed in the GTEx project. **(B)** The expression levels of a typical example CTA-929C8.8 among the 30 normal tissues in the GTEx project.

DISCUSSION

RNA-seq data of AD cortex tissue was obtained to identify the differentially expressed lncRNAs and PCGs; we established a co-expression network which could be helpful to identify a list of candidate AD-related lncRNAs. Further, these differentially expressed lncRNAs are cerebral tissue–specific, highly expressed in AD patients, and also co-expressed with AD-related PCGs. The lncRNA–PCG co-expressed network was constructed by WGCNA and was enriched in the AD-associated biological processes, such as cell adhesion, brain renin–angiotensin system, neuroprotective role of THOP1 in AD, and so on. Finally, several lncRNAs were identified, including MIR3180-2, MIR3180-3, and RP3-522J7, which were most highly co-expressed with the known AD-related genes in the network, such as S100B, AGT, NTS, and so on. These results suggested that the lncRNAs might directly regulate AD-related biological functions by targeting AD risk genes. Moreover, we further prioritized a set of differentially expressed lncRNAs and PCGs, which could be acting as potential AD markers, especially RP3-522J7, MIR3180-2, MIR3180-3, CTA-929C8, S100B, and PIRT. In summary, our study identified a compendium of lncRNAs and PCGs which have never been uncharacterized for AD and normal cerebral cortex tissue in previous research and could be prioritized as diagnostic markers for AD, which are helpful in identifying the high-risk population of AD, so as to intervene in prevention and treatment as early as possible.

Furthermore, the differentially expressed lncRNAs and PCGs identified in our study were validated by an independent data set (GEO accession GSE5281) (Liang et al., 2007). We found that 23 genes are uniformly differentially expressed. Especially, several key PCGs focused on by us are significantly differentially expressed in the test set, such as VTI1A, CUX1, S100B, AGT, CD44, and AQP1, and lncRNAs PSMG3-AS1 and FUT8-AS1. Webserver AlzData[2] was also used to analyze AD-related PCG (Zhang et al., 2019). We found that the differently expressed genes are in conformity in different brain regions. And the result showed that S100B was differently expressed in the entorhinal cortex. Two PCGs, AGT and IRAK4, were expressed aberrantly in temporal cortex tissue. In addition, CD44 was differentially expressed in the entorhinal cortex, hippocampus, and temporal cortex, while another PCG AQP1 was differentially expressed in both the hippocampus and temporal cortex. These results suggested that AD-related genes may have a spatial–temporal expression pattern.

It is shown that the knowledge of the functions for the regulatory factors could be helpful for further understanding the occurrence and development of complex diseases, and the dysregulated miRNAs were validated to be associated with AD (Sadlon et al., 2019). Therefore, in our study, the AD-related miRNAs were further prioritized. By identifying the miRNA–mRNA targeting relationship using the miRanda algorithm, it was shown that 18 miRNAs regulating the differential expression of PCGs were labeled as AD-related miRNAs by HMDD, and the targeted PCGs contained seven AD-related PCGs, including VTI1A, CUX1, S100B, AGT, CD44, IRAK4, and AQP1. Then, based on the characteristics of known AD miRNAs, we ranked the miRNAs by the overlap between the differently expressed PCGs targeted by known AD miRNAs and the targets of each miRNA to provide clues for predicting AD miRNA candidates (**Supplementary Table S6**). The higher the rank of a certain miRNA is, the more it is possible for it to be considered as an AD-related miRNA. For example, as the second-ranked miRNA, Hsa-miR-1229 directly regulated the expression level of AD-related gene SORL1, and other targeted genes, which could be involved in the biological processes of nervous system development and neurological disease (Ghanbari et al., 2016). Moreover, this miRNA has been confirmed to dysregulate expression across the different brain tissue regions in AD patients (Puthiyedth et al., 2016). Hsa-mir-328 has been reported to be associated with a variety of neurologic disorders, such as autism spectrum disorder, Huntington's, Parkinson's, and Alzheimer's, and the biological functions of its targeting genes APP and BACE1 were validated experimentally in mouse brain tissues (Boissonneault et al., 2009; Provost, 2010; Nt et al., 2018). High-throughput experimental data from starBase were used to verify the results, and similar conclusions were obtained (see **Supplementary Text S1** and **Supplementary Table S7**). These results may be helpful for exploring the role of miRNAs in AD and providing novel insight for the study of the pathophysiology of AD.

[2]www.alzdata.org

DATA AVAILABILITY STATEMENT

The RNA-seq data used in this study were obtained from GSE53697 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53697), GSE5281 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281), and GTEx (https://commonfund.nih.gov/GTEx/).

AUTHOR CONTRIBUTIONS

YL, TS, and ZZ conceived and designed the experiments. YS, HL, and CY performed the experiments and analyzed the data. YS, KX, YC, and ZW wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (31601065 and 61803130). The funders had no roles 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/fbioe.2020.00031/full#supplementary-material

REFERENCES

Alzheimer's Association (2015). 2015 Alzheimers disease facts and figures. *Alzheimers Dement* 11, 332–384. doi: 10.1016/j.jalz.2015.02.003

Bernstein, A. I., Lin, Y., Street, R. C., Lin, L., Dai, Q., Yu, L., et al. (2016). 5-Hydroxymethylation-associated epigenetic modifiers of Alzheimers disease modulate Tau-induced neurotoxicity. *Hum. Mol. Genet.* 25, 2437–2450. doi: 10.1093/hmg/ddw109

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:R90. doi: 10.1186/gb-2010-11-8-r90

Boissonneault, V., Plante, I., Rivest, S., and Provost, P. (2009). MicroRNA-298 and microRNA-328 regulate expression of mouse beta-amyloid precursor protein-converting enzyme 1. *J. Biol. Chem.* 284, 1971–1981. doi: 10.1074/jbc.M807530200

Cirillo, C., Capoccia, E., Iuvone, T., Cuomo, R., Sarnelli, G., Steardo, L., et al. (2015). S100B inhibitor pentamidine attenuates reactive gliosis and reduces neuronal loss in a mouse model of Alzheimer's Disease. *Biomed. Res. Int.* 2015:508342. doi: 10.1155/2015/508342

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

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606

Ghanbari, M., Ikram, M. A., de Looper, H. W. J., Hofman, A., Erkeland, S. J., Franco, O. H., et al. (2016). Genome-wide identification of microRNA-related variants associated with risk of Alzheimer's disease. *Sci. Rep.* 6:28387. doi: 10.1038/srep28387

Grupe, A., Li, Y., Rowland, C., Nowotny, P., Hinrichs, A. L., Smemo, S., et al. (2006). A scan of chromosome 10 identifies a novel locus showing strong association with late-onset Alzheimer disease. *Am. J. Hum. Genet.* 78, 78–88. doi: 10.1086/498851

GTEx, (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi: 10.1038/ng.2653

Han, S., Zhuang, H., Lee, P. Y., Li, M., Yang, L., Nigrovic, P. A., et al. (2019). Differential responsiveness of monocyte and macrophage subsets to interferon. *Arthritis Rheumatol.* 72, 100–113. doi: 10.1002/art.41072

Harper, S. (2014). Economic and social implications of aging societies. *Science* 346, 587–591. doi: 10.1126/science.1254405

Hensman Moss, D. J., Flower, M. D., Lo, K. K., Miller, J. R., van Ommen, G. B., t Hoen, P. A., et al. (2017). Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. *Sci. Rep.* 7:44849. doi: 10.1038/srep44849

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 47, D1013–D1017. doi: 10.1093/nar/gky1010

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

Kinnersley, B., Labussiere, M., Holroyd, A., Di Stefano, A. L., Broderick, P., Vijayakrishnan, J., et al. (2015). Genome-wide association study identifies multiple susceptibility loci for glioma. *Nat. Commun.* 6:8559. doi: 10.1038/ncomms9559

Kold-Christensen, R., and Johannsen, M. (2019). Methylglyoxal metabolism and aging-related disease: moving from correlation toward causation. *Trends Endocrinol. Metab.* 31, 81–92. doi: 10.1016/j.tem.2019.10.003

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, 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

Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Walker, D. G., et al. (2007). Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics* 28, 311–322. doi: 10.1152/physiolgenomics.00208.2006

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

Nt, P., Ds, M., Mm, N., In, M., and Ti, V. (2018). Investigation of circulating serum MicroRNA-328-3p and MicroRNA-3135a expression as promising novel biomarkers for autism spectrum disorder. *Balkan J. Med. Genet.* 21, 5–12. doi: 10.2478/bjmg-2018-0026

Papait, R., Kunderfranco, P., Stirparo, G. G., Latronico, M. V., and Condorelli, G. (2013). Long noncoding RNA: a new player of heart failure? *J. Cardiovasc. Transl. Res.* 6, 876–883. doi: 10.1007/s12265-013-9488-6

Peters, M. J., Joehanes, R., Pilling, L. C., Schurmann, C., Conneely, K. N., Powell, J., et al. (2015). The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* 6:8570. doi: 10.1038/ncomms9570

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, 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

Prasad, K., Abdel-Hameed, A. A. E., Xing, D., and Reddy, A. S. N. (2016). Global gene expression analysis using RNA-seq uncovered a new role for SR1/CAMTA3 transcription factor in salt stress. *Sci. Rep.* 6:27021. doi: 10.1038/srep27021

Provost, P. (2010). Interpretation and applicability of microRNA data to the context of Alzheimer's and age-related diseases. *Aging* 2, 166–169. doi: 10.18632/aging.100131

Puthiyedth, N., Riveros, C., Berretta, R., and Moscato, P. (2016). Identification of differentially expressed genes through integrated study of Alzheimer's Disease affected brain regions. *PLoS One* 11:e0152342. doi: 10.1371/journal.pone.0152342

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166. doi: 10.1146/annurev-biochem-051410-092902

Sadlon, A., Takousis, P., Alexopoulos, P., Evangelou, E., Prokopenko, I., and Perneczky, R. (2019). miRNAs identify shared pathways in Alzheimer's and Parkinson's Diseases. *Trends Mol. Med.* 25, 662–672. doi: 10.1016/j.molmed.2019.05.006

Scheckel, C., Drapeau, E., Frias, M. A., Park, C. Y., Fak, J., Zucker-Scharff, I., et al. (2016). Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. *Elife* 5:e10421. doi: 10.7554/eLife.10421

Shi, X., Sun, M., Liu, H., Yao, Y., and Song, Y. (2013). Long non-coding RNAs: a new frontier in the study of human diseases. *Cancer Lett.* 339, 159–166. doi: 10.1016/j.canlet.2013.06.013

Sood, S., Gallagher, I. J., Lunnon, K., Rullman, E., Keohane, A., Crossland, H., et al. (2015). A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome Biol.* 16:185. doi: 10.1186/s13059-015-0750-x

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960. doi: 10.1126/science.1160342

Sultmann, H., and Diederichs, S. (2014). Long noncoding RNA: "LNCs" to cancer. *Eur. Urol.* 65, 1152–1153. doi: 10.1016/j.eururo.2014.01.032

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, 2074–2081. doi: 10.1039/c3mb70608g

Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., et al. (2018). Human ageing genomic resources: new and updated databases. *Nucleic Acids Res.* 46, D1083–D1090. doi: 10.1093/nar/gkx1042

Toledo, J. B., Zetterberg, H., van Harten, A. C., Glodzik, L., Martinez-Lage, P., Bocchio-Chiavetto, L., et al. (2015). Alzheimer's disease cerebrospinal fluid biomarker in cognitively normal subjects. *Brain* 138(Pt 9), 2701–2715. doi: 10.1093/brain/awv199

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

Twine, N. A., Janitz, K., Wilkins, M. R., and Janitz, M. (2011). Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS One* 6:e16266. doi: 10.1371/journal.pone.0016266

Uberti, D., Cenini, G., Bonini, S. A., Barcikowska, M., Styczynska, M., Szybinska, A., et al. (2010). Increased CD44 gene expression in lymphocytes derived from Alzheimer disease patients. *Neurodegener. Dis.* 7, 143–147. doi: 10.1159/000289225

Wan, G., Zhou, W., Hu, Y., Ma, R., Jin, S., Liu, G., et al. (2016). Transcriptional regulation of lncRNA genes by histone modification in Alzheimer's Disease. *Biomed. Res. Int.* 2016:3164238. doi: 10.1155/2016/3164238

Xiong, J., Lu, X., Zhou, Z., Chang, Y., Yuan, D., Tian, M., et al. (2012). Transcriptome analysis of the model protozoan, Tetrahymena thermophila, using Deep RNA sequencing. *PLoS One* 7:e30630. doi: 10.1371/journal.pone.0030630

Xu, M., Zhang, D. F., Luo, R., Wu, Y., Zhou, H., Kong, L. L., et al. (2018). A systematic integrated analysis of brain expression profiles reveals YAP1 and other prioritized hub genes as important upstream regulators in Alzheimer's disease. *Alzheimers Dement* 14, 215–229. doi: 10.1016/j.jalz.2017.08.012

Zhang, D. F., Fan, Y., Xu, M., Wang, G., Wang, D., Li, J., et al. (2019). Complement C7 is a novel risk gene for Alzheimer's disease in Han Chinese. *Natl. Sci. Rev.* 6, 257–274. doi: 10.1093/nsr/nwy127

Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2019). Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Brief Bioinform* 20, 598–608. doi: 10.1093/bib/bby021

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

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 © 2020 Shi, Liu, Yang, Xu, Cai, Wang, Zhao, Shao 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 26 February 2020
doi: 10.3389/fbioe.2020.00134

RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites

Zhibin Lv[1], Jun Zhang[2], Hui Ding[3] and Quan Zou[1,3]**

[1] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China,
[2] Rehabilitation Department, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China,
[3] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China

One of the ubiquitous chemical modifications in RNA, pseudouridine modification is crucial for various cellular biological and physiological processes. To gain more insight into the functional mechanisms involved, it is of fundamental importance to precisely identify pseudouridine sites in RNA. Several useful machine learning approaches have become available recently, with the increasing progress of next-generation sequencing technology; however, existing methods cannot predict sites with high accuracy. Thus, a more accurate predictor is required. In this study, a random forest-based predictor named RF-PseU is proposed for prediction of pseudouridylation sites. To optimize feature representation and obtain a better model, the light gradient boosting machine algorithm and incremental feature selection strategy were used to select the optimum feature space vector for training the random forest model RF-PseU. Compared with previous state-of-the-art predictors, the results on the same benchmark data sets of three species demonstrate that RF-PseU performs better overall. The integrated average leave-one-out cross-validation and independent testing accuracy scores were 71.4% and 74.7%, respectively, representing increments of 3.63% and 4.77% versus the best existing predictor. Moreover, the final RF-PseU model for prediction was built on leave-one-out cross-validation and provides a reliable and robust tool for identifying pseudouridine sites. A web server with a user-friendly interface is accessible at http://148.70.81.170:10228/rfpseu.

Keywords: pseudouridine sites, light gradient boosting, random forest, machine learning, RNA

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Hongmin Cai,
South China University of Technology, China

****Correspondence:***
Jun Zhang
zhangjun13902003@163.com
Quan Zou
zouquan@nclab.net

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *17 January 2020*
Accepted: *10 February 2020*
Published: *26 February 2020*

Citation:
Lv Z, Zhang J, Ding H and Zou Q (2020) RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites. Front. Bioeng. Biotechnol. 8:134. doi: 10.3389/fbioe.2020.00134

INTRODUCTION

More than 150 types of chemical modification have been identified in cellular RNA, including adenosine methylation, cytosine modification, isomerization of uridine, and ribose modification (Boccaletto et al., 2018). These modifications have critical roles in cellular biological and physiological processes (Song and Yi, 2017). For instance, one of the most prevalent RNA modifications in eukaryotes, N^6-methyladenosine (m6A), affects RNA stability (Wang et al., 2014), RNA-protein interaction (Liu et al., 2015b), RNA splicing and translation (Meyer and Jaffrey, 2014), the circadian clock (Fustin et al., 2013), immune response (Winkler et al., 2019), etc. Another widespread RNA modification is 5-methylcytosine (m5C), which has functions including preservation of the secondary structure of tRNA (Motorin and Helm, 2010), control of amino-acylation (Helm, 2006), codon identification and metabolic stability (Agris, 2008; Li et al., 2017). The pseudouridine modification is another common post-transcriptional modification in

various living organisms (Zaringhalam and Papavasiliou, 2016). In 1951, pseudouridine was first identified, and experiments in 1960 revealed that it was abundant in tRNA and rRNA (Cohn, 1960). Pseudouridine results from an isomerization of uridine by breaking the glycosidic bond with 180° base rotation (Karijolich et al., 2015). This modification has been shown to have vital roles, for instance, in stabilizing RNA and in the stress response (Zhao and He, 2015; Cheng et al., 2019a; Wang et al., 2019b).

Although RNA pseudouridylation was discovered decades ago, the first transcriptome-wide RNA pseudouridylation map was not published until 2014, following the rapid development of next-generation sequencing technology (Goodwin et al., 2016). Carlile et al. (2014) developed the PseudoU-seq technology, which they used to identify more than 200 pseudouridylation sites in the regulated mRNA of yeast and human cells; in the same year, Schwartz et al. (2014) performed transcriptome-wide mapping using a similar protocol, finding more than 300 dynamic-regulated pseudouridine sites in non-coding RNA and mRNA. Li et al. (2015a) presented a chemical labeling method (CeU-Seq) that they used to pull down more than 2000 pseudouridine sites in human mRNA. Other RNA pseudouridylation sequencing protocols were also developed (Carlile et al., 2015).

As an alternative to costly and labor-intensive laboratory experiments, robust, swift, and inexpensive computational methods for RNA chemical modification prediction have emerged recently, owing to the increasing amount of data generated in this post-genomics era (Libbrecht and Noble, 2015). A large number of m6A (Chen et al., 2015, 2018a,b, 2019a; Zhou et al., 2016; Zhao et al., 2019; Zou et al., 2019) and m5C (Feng et al., 2016; Qiu et al., 2017; Li et al., 2018; Sabooh et al., 2018; Zhang et al., 2018; Yin et al., 2019) site predictors based on traditional machine learning and emerging deep learning algorithms have been proposed. However, few computational tools have been developed to predict pseudouridine sites. Li et al. (2015b) used a support vector machine (SVM) classifier to design a web server called PPUS for the identification of pseudouridine sites in *Saccharomyces cerevisiae* and *Homo sapiens*. Chen et al. (2016) constructed another SVM-based web server for pseudouridine site prediction, using the frequency composition of the nucleotides and pseudo K-tuple nucleotide composition (PseKNC) for feature representation. He et al. (2018) presented another model, PseUI, to identify pseudouridine sites in RNA sequences from three species (*H. sapiens*, *S. cerevisiae*, and *M. musculus*); this was an SVM-based model incorporating multiple feature-extraction technologies. Tahir et al. (2019) used convolutional neural networks to design a new predictor, iPseU-CNN; and Liu et al. (2019b) developed the eXtreme gradient boosting (XGboost) method for RNA pseudouridine site prediction (XG-PseU). Cross-validation scores for RNA pseudouridine site identification in the abovementioned three species showed the best accuracy for iPseU-CNN (66.9%) in *H. sapiens*, whereas XG-PseU and iPseU-CNN had the best accuracy (68.2%) in *S. cerevisiae*, and XG-PseU was the most accurate (72.0%) in *M. musculus*. According to independent testing scores, iPseU-CNN outperformed the other models, with 69.0% accuracy in *H. sapiens* and 73.6% accuracy in *S. cerevisiae*. Although the iPseU-CNN predictor had a high average cross-validation accuracy (68.9%) and independent testing accuracy (71.3%) scores, there was still room for improvement in comparison with some high-performing m6A site predictors (Chen et al., 2019a; Zou et al., 2019).

In this work, a model is developed based on the random forest algorithm, RF-PseU, for pseudouridine site recognition. The modeling overview is shown in **Figure 1**. RF-PseU incorporates multiple sequence feature representation technologies, and the light gradient boosting machine (LGBM) algorithm is employed to remove redundant features and rank the remaining features. Evaluation with leave-one-out (LOO) cross-validation demonstrated the robustness of the model. The average cross-validation accuracy (71.3% for 10-Fold and 71.4% for LOO) of RF-PseU was improved by 3.48–10.3% compared with existing state-of-the-art predictors, and the average independent testing accuracy (74.7%) showed a 4.8–19% increase. A user-friendly web server was also implemented, which can be accessed at http://148.70.81.170:10228/rfpseu. RF-PseU is expected to be a useful supplement to the existing tools for pseudouridine site identification.

MATERIALS AND METHODS

Data Sets

Given that there were small differences between the benchmark data sets used in the studies of Chen et al. (2018a) and Liu et al. (2019b), data sets obtained from Chen et al. (2018a) were used to train and test our models. The training data sets included data for three species. That is, *H. sapiens* training dataset with 495 psedouridine-sites-containing sequences and 495 non-psedouridine-sites-containing; *S. cerevisiae* training dataset contains 314 psedouridine-sites-sequences and 314 non-psedouridine-sites-sequences; *M. musculus* training dataset consists of 944 sequences, half of which is positive samples. Whereas the independent testing data sets covered only two species, *H. sapiens* and *S. cerevisiae*, both of which contain 100 positive samples and 100 negative samples. For the *H. sapiens* and *M. musculus* data sets, the window size was 21, i.e. the positive samples were psedouridine site centroid sequences of 21 base pairs each, whereas those for the *S. cerevisiae* samples window site was 31, with psedouridine site centroid sequences of 31 base pairs. Negative samples, in which no psedouridine sites were detected, consisted of 21 base pairs for *H. sapiens* and *M. musculus*, and 31 base pairs for *S. cerevisiae*. The benchmark data sets can be downloaded from http://lin-group.cn/server/iRNAPseu/data.

Feature Representation

Several widely used and convenient bio-sequence feature representation tools have been developed (Mrozek et al., 2013; Liu et al., 2015a, 2019c; Yu et al., 2015, 2019; Cheng and Hu, 2018; Hu et al., 2019; Muhammod et al., 2019). The two main tools used in this work were iLearn (Hu et al., 2019) and PyFeat (Muhammod et al., 2019).

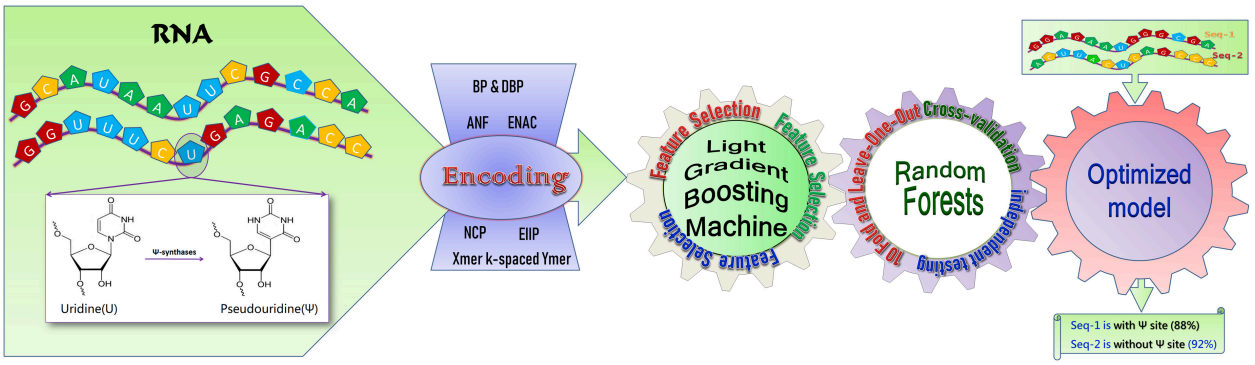


FIGURE 1 | A schematic diagram of RF-PseU. RNA sequences with or without pseudouridine sites were encoded via seven RNA coding technologies; following removal of redundant features by light gradient boosting machine feature selection, the random forest model was trained on smaller but more relevant feature vector spaces, and was evaluated through cross-validation and independent testing to obtain an optimized model for prediction.

Nucleotide Binary Profiles

Binary profiles encode the four bases (ACGU) as (1,0,0,0), (0,1,0,0), (0,0,1,0), and (0,0,0,1), whereas dibinary profiles encode the 16 dinucleotides (AA, AC,AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, and UU) as (0,0,0,0), (0,0,0,1), (0,0,1,0), (0,0,1,1),. . ., (1,1,1,1).

Accumulated Nucleotide Frequency

Suppose s_i is a base (ACGU) at the ith position of a RNA sequence. Then we can determine the s_i density d_i of the ith prefix subsequence of a RNA sequence as follows:

$$d_i = \frac{i}{|s_i|}\sum_{j=1}^{L} f(s_i), \quad where\, f(q) = \begin{cases} 1, & if\ s_i = q \\ 0, & otherwise \end{cases},$$

where L is the sequence length and q is one of the four nucleotides (ACGU).

Nucleotide Chemical Properties

The four RNA nucleotides (ACGU) are different from each other in terms of chemical structure and chemical bonds. On the basis of these differences, AGCU can be categorized into three different classes (**Table 1**) and encoded using a three-dimensional coordinate, i.e. A is denoted by (1,1,1), C by (0,1,0), G by (1,0,0), and U by (0,0,1).

Electron-Ion Interaction Pseudopotentials (EIIP)

Nair and Sreenadhan (2006) used the EIIP values of A, G, C, and T (A: 0.1260, G: 0.0806, C: 0.1340, T: 0.1335) to directly represent the nucleotides in a DNA sequence. Here, iLearn was used to encode each nucleotide in the RNA sequences into EIIP feature vectors.

TABLE 1 | ACGU categories based on chemical properties.

Nucleotides	Chemical property
C,U	Pyrimidine and ring structure
A,G	Purine and ring structure
A,U	Weak and hydrogen bond
C,G	Strong and hydrogen bond
G,U	Keto and functional group
A,C	Amino and functional group

Enhanced Nucleic Acid Composition

The nucleotide composition was calculated for a fixed-length window of the RNA sequence, allowing the fixed window (length = 5) to continuously slide from the 5′ to the 3′ terminus. RNA sequences were then encoded into feature vectors of equal length.

Xmer k-Spaced Ymer Composition Frequency

This method is used to count the composition of a subsequence of X and Y consecutive nucleotides with intervals k, e.g. AGU@AU, A@CU, GU@@@A, where @ indicates a one-interval space, @@ a two-interval space, and so on. Generally, using Xmer k-spaced Ymer to encode an RNA sequence will generate a $4^X \times 4^Y$ feature vector. In this study, X, Y, and k were set to 1, 2, or 3; and eight XYK combinations (except for 3mer-kspaced-3mer) were used for encoding. The PyFeat tool developed by Rafsanjani et al. (Muhammod et al., 2019) was used to convert RNA sequences into vectors.

Feature Selection

Feature selection is an effective way to remove redundant information and prevent over-fitting in machine learning modeling (Tang et al., 2017; Xu et al., 2018a; Cheng et al., 2019a; Liu, 2019; Sun et al., 2019; Yu et al., 2019). Several feature selection technologies, including ANOVA (Lv et al., 2019b) and MRMD (Zou et al., 2016), have been developed and are widely used for DNA, RNA, and protein identification (Xu et al., 2018b). In this work, an LGBM (Ke et al., 2017)[1] wrapper was used to select appropriate feature spaces for model training. In this process, raw training data were fed into the LGBM model and their features were ranked by importance value as calculated with the LGBM algorithm. Features with importance values greater than the average were selected to compose the feature space for modeling.

[1]https://lightgbm.readthedocs.io

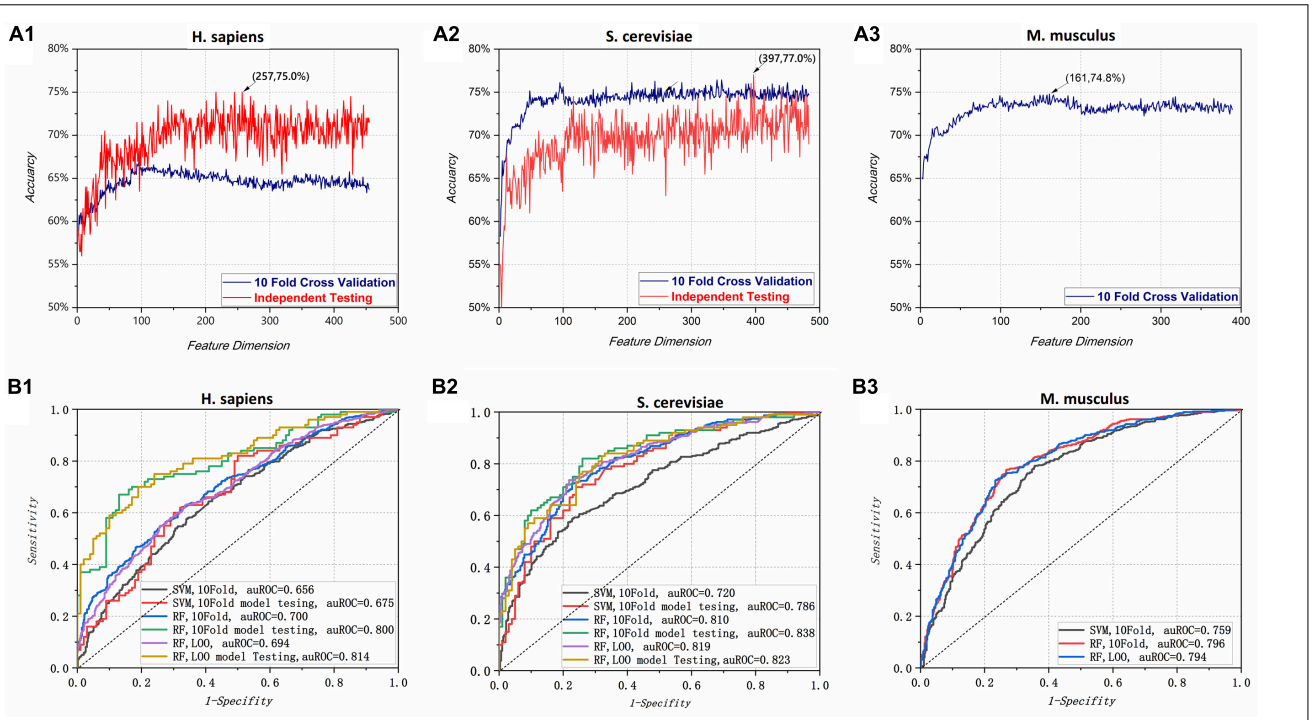


FIGURE 2 | (A) Accuracy of the random forest predictor varied with feature dimension for all three species: **(A1)** *H. sapiens*; **(A2)** *S. cerevisiae*; **(A3)** *M. musculus*. The best independent accuracies for *H. sapiens* and *S. cerevisiae* were 75.0% with 257 features and 77.0% with 397 features, respectively, and the best 10-Fold cross-validated accuracy for *M. musculus* was 74.8% with 161 features. **(B)** Receiver operating characteristic curve (ROC) and area under the ROC curve (auROC) for different species under various conditions. **(B1)** is for *H. sapiens*, **(B2)** is for *S. cerevisiae* and **(B3)** is *M. musculus*. A support vector machine (SVM) was used for comparison with the random forest (RF) model. 10-Fold (10-Fold) model testing and leave-one-out (LOO) model testing indicate the model with the best 10-Fold and LOO cross-validation scores in independent testing. In cross-validation (10-Fold and LOO) and testing process, the training datasets have divided into training part and validation part. That is, they have used the general machine learning evaluation methods (training, validation and testing) for model optimization. In the figure, the 10-fold cross-validation and LOO cross-validation metric values are obtained from the validation part of training part, while the independent testing metric values are obtained from the independent testing datasets.

Model Evaluation Metrics and Methods

The proposed models were evaluated by five commonly used metrics, accuracy (ACC), sensitivity (Sn), specificity (Sp), Matthew correlation coefficient (MCC), and integral area under the receiver operating characteristic curve (auROC). These metrics were calculated using the following equations, where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively (Cheng et al., 2016, 2019b,c; Wei et al., 2017d,e; Liu et al., 2019a). For the ROC curve, 1-specificity was plotted on the horizontal axis, and sensitivity on the vertical axis.

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$S_n = \frac{TP}{FN + TP}$$

$$S_n = \frac{TN}{FP + TN}$$

$$\mathrm{MCC} = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}}$$

LOO, K-Fold cross-validation, and independent testing are the most widely used methods for predictor evaluation (Mrozek et al., 2015; Cao and Cheng, 2016; Chen et al., 2017, 2018a, 2019b; Pan et al., 2017; He et al., 2018, 2019; Jiang et al., 2018; Xiong et al., 2018; Yu et al., 2018; Zhang et al., 2018; Ding et al., 2019; Feng et al., 2019; Kong and Zhang, 2019; Li and Liu, 2019; Lv et al., 2019a; Manavalan et al., 2019; Shan et al., 2019; Wang et al., 2019a; Wei et al., 2019a,b; Xu et al., 2019; Yu and Dai, 2019). That is the general machine learning evaluation methods (training, validation and testing) are used for optimized model evaluation. To test the efficiency of the classification, LOO cross-validation was performed for a data set containing n items, of which n-1 items were used for training and the remaining one was used for validation. This procedure was repeated until every sequence in the training data set had been used once as a validation testing sample. LOO cross-validation is robust but time-consuming for a large data set. To compare the performance of the model with that of existing predictors, 10-Fold cross-validation was also used. The training data set was stochastically divided into 10 subsets, with one subset for validation and the remaining nine for training. This process was repeated 10 times and the average results were used to evaluate the model. Finally, independent testing was performed to obtain a data set that was

completely distinct from the training data set for evaluation of the trained model.

Algorithm

The random forest method is a bagging-type ensemble learning algorithm (Cheng et al., 2018a,b). By combining multiple weak classifiers, the final results can be voted or averaged to obtain an overall model with higher accuracy, better general performance, and resistance to overfitting. This algorithm has been extensively used in bioinformatics and other areas, and has been confirmed to be an effective modeling technique in various domains (Ding et al., 2016a,b; Mrozek et al., 2016; Qiu et al., 2016; Wang et al., 2017; Wei et al., 2017a,b,c; Yu et al., 2017a; Zheng et al., 2017; Tang et al., 2018, 2019a; Xue et al., 2018; Degenhardt et al., 2019; Xu et al., 2019). In this study, the scikit-learn toolkit, available at https://scikit-learn.org, was used to establish the models.

Support vector machine (Cortes and Vapnik, 1995) is a generalized linear classifier that classifies data based on

TABLE 2 | Cross-validation and independent testing scores of two different classifiers for three species.

Species	Algorithm	10 fold cross-validation					Independent testing				
		ACC	MCC	Sn	Sp	auROC	ACC	MCC	Sn	Sp	auROC
H. sapiens	SVM	62.0%	0.240	61.4%	62.6%	0.656	64.0%	0.280	66.0%	62.0%	0.679
	RF	64.3%	0.287	66.1%	62.6%	0.700	75.0%	0.501	78.0%	72.0%	0.800
S. cerevisiae	SVM	67.5%	0.352	73.7%	61.2%	0.720	72.5%	0.45	73.0%	73.0%	0.786
	RF	74.8%	0.497	77.2%	72.4%	0.810	77.0%	0.540	75.0%	79.0%	0.838
M. musculus	SVM	70.7%	0.42	65.9%	75.4%	0.759	/	/	/	/	/
	RF	74.8%	0.50	73.1%	76.5%	0.796	/	/	/	/	/

TABLE 3 | Comparison of cross-validation and independent testing scores of existing state-of-the-art pseudouridine site predictors and RF-PseU.

Species	Classifier	Cross-validation					Independent testing				
		ACC	MCC	Sn	Sp	auROC	ACC	MCC	Sn	Sp	auROC
H. sapiens	iRNA-PseU(LOO)[a]	60.4%	0.21	61.0%	59.8%	0.640	65.0%	0.30	60.0%	70.0%	/
	PseUI(LOO)[a]	64.2%	0.28	64.9%	63.6%	0.68	65.5%	0.31	63.0%	68.0%	/
	iPseU-CNN(5F)[b]	66.7%	0.34	65.0%	68.8%	/	69.0%	0.40	77.7%	60.8%	/
	XG-PseU (10F)[c]	66.1%	0.32	63.5%	68.7%	0.700	67.5%	/	/	/	/
	RF-PseU(10F)[d]	64.3%	0.29	66.1%	62.6%	0.700	75.0%	0.50	78.0%	72.0%	0.800
	RF-PseU(LOO)[e]	64.0%	0.29	65.9%	62.6%	0.694	74.0%	0.48	74.0%	74.0%	0.814
S. cerevisiae	iRNA-PseU(LOO)	64.5%	0.29	64.7%	64.3%	0.81	60.0%	0.20	63.0%	57.0%	/
	PseUI(LOO)	64.1%	0.30	64.7%	67.5%	0.69	68.5%	0.37	65.0%	72.0%	/
	iPseU-CNN(5F)	68.2%	0.37	66.4%	70.5%	/	73.5%	0.47	68.8%	77.8%	/
	XG-PseU(10F)	68.2%	0.37	66.8%	69.5%	0.77	71.0%	/	/	/	/
	RF-PseU(10F)	74.8%	0.49	77.2%	72.4%	0.810	77.0%	0.54	75.0%	79.0%	0.838
	RF-PseU(LOO)	75.8%	0.52	78.2%	73.4%	0.819	74.5%	0.49	70.0%	79.0%	0.823
M. musculus	iRNA-PseU(LOO)	69.1%	0.38	73.3%	64.8%	0.75	/	/	/	/	/
	PseUI(LOO)	70.4%	0.41	79.9%	70.3%	0.71	/	/	/	/	/
	iPseU-CNN(5F)	71.8%	0.44	74.8%	69.1%	/	/	/	/	/	/
	XG-PseU(10F)	72.0%	0.45	76.5%	67.6%	0.74	/	/	/	/	/
	RF-PseU(10F)	74.8%	0.50	73.1%	76.5%	0.796	/	/	/	/	/
	RF-PseU(LOO)	74.5%	0.48	72.7%	75.2%	0.794	/	/	/	/	/

[a]Predictors based on support vector machine, Leave-One-Out Cross-Validation (LOO); [b]Predictors based on convolutional neural nets, five-fold cross-validation (5F); [c]Predictors based on XGboost,10-fold cross-validation (10F); [d]Predictors based on Random Forest, 10-fold cross-validation; [e]LOO:Leave-One-Out Cross-Validation

TABLE 4 | Comparison of average accuracies for state-of-the-art predictors.

Scores type	RF-PseU (10 Fold[c])	RF-PseU (LOO[d])	iRNA-PseU (LOO)	PseUI (LOO)	iPseU-CNN (5 Fold[e])	XG-PseU (10 Fold)
Cross-validation[a]	71.3%	71.4%	64.7%	66.2%	68.9%	68.7%
Independent testing[b]	76.0%	74.7%	62.5%	67.0%	71.3%	69.3%

[a]Average values of H. sapiens, S. cerevisiae and M. musculus; [b]Average values of H. sapiens and S. cerevisiae; [c]model with 10-fold cross-validation; [d]model with leave-one-out cross-validation; [e]model with five-fold cross-validation.

supervised learning; its decision boundary is the maximum-margin hyperplane required to solve the learning sample. SVM has been widely used in a variety of fields (Xiong et al., 2012; Ding et al., 2017; Yu et al., 2017b; Fu et al., 2018; Fang et al., 2019; Lai et al., 2019; Meng et al., 2019; Shen et al., 2019; Tang et al., 2019b; Zhang et al., 2019; Zhu et al., 2019). Here, it was used for modeling comparisons. SVM was also implemented via the scikit-learn toolkit, using the Gaussian radial basis functions, with the critical hyper-parameters (C and γ) of SVM optimized in a range from 10^{-6} to 10^{6} with exponent step $10^{0.5}$.

RESULTS AND DISCUSSION

Optimization With Different Feature Spaces

To determine optimal feature spaces, we first used the LGBM algorithm to sort the features from maximum to minimum according to their importance value. All the features with importance value greater than the average were kept. Second, we used an incremental feature selection strategy; as shown in **Figure 2A**, the 10-Fold cross-validation and independent testing accuracy varied as features were added. Initially, the accuracy increased rapidly for each species. As shown in **Figure 2** (A1) and **Figure 2** (A2), when the feature dimensions for *H. sapiens* and *S. cerevisiae* reached 257 and 397, the model achieved maximum independent testing accuracies of 75.0 and 77.0%, respectively. Owing to the lack of independent test data sets for *M. musculus*, **Figure 2** (A3) shows only the cross-validation accuracy curve, with its peak value (74.8%) at a feature dimension of 161. The optimal feature space dimensions selected for each species were 257, 397, and 161, respectively. These values were used for further experiments and optimization.

Comparison With SVM Predictors

Given that PPUS (Li et al., 2015b), iRNA-PseU (Chen et al., 2016), and PseUI (He et al., 2018) were all based on SVM, an optimized SVM model for pseudouridine site identification with the same feature spaces as the RF model was constructed to determine the effects of the SVM and RF on prediction performance. The performances of the two models are shown in **Table 2**. Overall, the models based on RF showed markedly better performance than those based on SVM. For instance, in terms of 10-Fold cross-validation accuracy, the RF models for *H. sapiens*, *S. cerevisiae*, and *M. musculus* outperformed the corresponding SVM models by 3.71%, 10.8%, and 5.80%, respectively. The independent testing accuracy scores showed an even greater contrast. For example, the RF model had 75.0% accuracy for *H. sapiens*, exactly 1.17 times that of the SVM model. The ROC curve and auROC value shown in **Figure 2B** also demonstrate that the optimized RF models performed better than the optimized SVM models for the same feature spaces. Thus, non-SVM models such as XG-PseU (Liu et al., 2019b), iPseU-CNN (Tahir et al., 2019), and our RF-PseU model might be more suitable for distinguishing pseudouridine sites from non-pseudouridine sites.

Comparison With Previous Predictors

The performance of RF-PseU was also compared with that of state-of-the-art predictors including iRNA-PseU (Chen et al., 2016), PseUI (He et al., 2018), iPseU-CNN (Tahir et al., 2019),

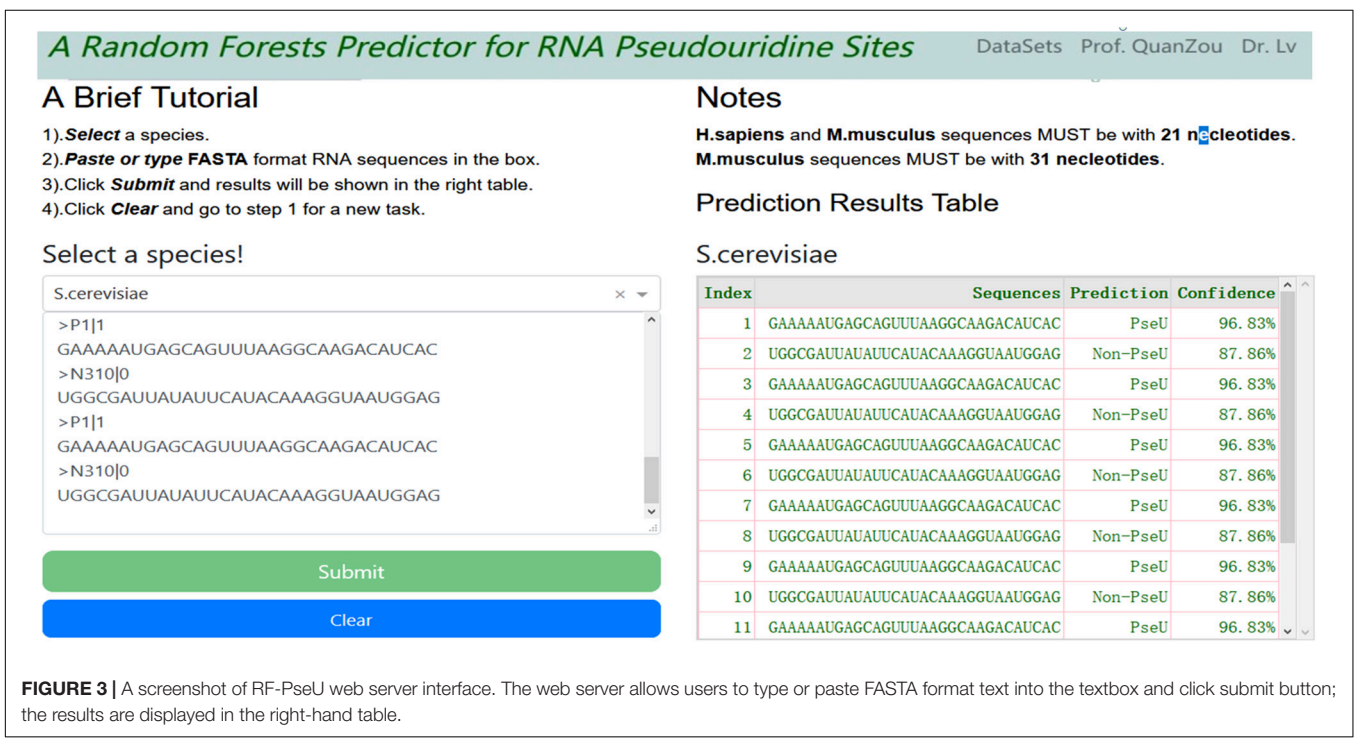


FIGURE 3 | A screenshot of RF-PseU web server interface. The web server allows users to type or paste FASTA format text into the textbox and click submit button; the results are displayed in the right-hand table.

and XG-PseU (Liu et al., 2019b). First, we compared the evaluation scores for the three species. **Table 3** compares the cross-validation and independent testing scores for the state-of-the-art pseudouridine sites predictors with those of RF-PseU. In terms of cross-validation scores, the LOO accuracy values for *S. cerevisiae* and *M. musculus* were 75.4% and 74.5%, respectively, representing increments of approximately 10.5% and 3.47% over the values for the existing predictor (XG-PseU) with the best cross-validation score. However, the LOO accuracy of RF-PseU for *H. sapiens*, at 64.0%, showed a decrease of 4.0% compared with the best *H. sapiens* pseudouridine site predictor, PseU-CNN. In terms of independent testing, as shown in **Table 3**, RF-PseU scored higher than the existing predictors in all aspects. For comprehensive comparison, the average scores for different species were calculated. The results, shown in **Table 4**, demonstrate that RF-PseU performed better overall than the other four predictors. The cross-validation accuracy scores of RF-PseU were 3.48% higher than those of the best existing predictor, iPseU-CNN; in terms of independent testing scores, RF-PseU showed a marked improvement of 4.7–10.6% compared with iPseU-CNN. The overall performance of RF-PseU was also significantly better than those of the other predictors, indicating that RF-PseU can discriminate true pseudouridine sites from non-pseudouridine sites more precisely than the existing predictors.

Web Server Implementation

For convenience, a webserver with an easy-to-use interface was developed (see screenshot in **Figure 3**), which can be accessed freely at http://148.70.81.170:10228/rfpseu. A step-by-step user guide is given here. First, users select a species from the drop-down box and paste or type the query RNA sequences in FASTA format into the textbox. Second, after clicking the submit button, the query results will be shown in a table on the same page after a wait. Note that once a query task has been submitted, the submit button will be disabled. Third, the user can click the clear button to empty the input text box and enable the submit button, and return to step one to enter a new query task.

CONCLUSION

In this study, a new model, named RF-PseU, for predicting RNA pseudouridine sites in multiple species is presented. For given feature spaces, the random forest algorithm was shown to be more efficient than SVM models for discriminating pseudouridine sites from non-pseudouridine sites. In terms of average cross-validation and independent testing accuracy scores, RF-PseU showed improvements of 3.6–10% and 4.8–21%, respectively, compared with state-of-the-art predictors. Moreover, a web server with a user-friendly interface is available. It is anticipated that RF-PseU will be a useful tool for RNA pseudouridine site analysis. However, the model requires further development via combination with other technologies before it is suitable for use as a classifier for RNA pseudouridine sites. Future work will explore emerging methods such as Gene2Vec (Zou et al., 2019), m6Acomet (Wu et al., 2019), and iterative feature representation (Wei et al., 2019b) to improve the model's performance.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://lin-group.cn/server/iRNAPseu/data.

AUTHOR CONTRIBUTIONS

ZL and JZ were responsible for experiments and manuscripts preparation. HD participated in discussions. QZ worked as supervisor for all procedures.

FUNDING

The work was supported by the National Natural Science Foundation of China (Nos. 61922020, 61771331, and 91935302), and the Scientific Research Foundation in Shenzhen (JCYJ20170818100431895 and JCYJ20180306172207178).

REFERENCES

Agris, P. F. (2008). Bringing order to translation: the contributions of transfer RNA anticodon-domain modifications. *Embo Rep.* 9, 629–635. doi: 10.1038/embor.2008.104

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., and Wirecki, T. K. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Cao, R., and Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Sci. Rep.* 6:23990. doi: 10.1038/srep23990

Carlile, T. M., Rojas-Duran, M. F., and Gilbert, W. V. (2015). Pseudo-seq: genome-wide detection of pseudouridine modifications in RNA. *Methods Enzymol.* 560, 219-245. doi: 10.1016/bs.mie.2015.03.011

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* 515, 143–146. doi: 10.1038/nature13802

Chen, K. Q., Wei, Z., Zhang, Q., Wu, X. Y., Rong, R., Lu, Z. L., et al. (2019a). WHISTLE: a high-accuracy map of the human N-6-methyladenosine (m(6)A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* 47:e41. doi: 10.1093/nar/gkz074

Chen, W., Ding, H., Zhou, X., Lin, H., and Chou, K. C. (2018a). iRNA(m6A)-PseDNC: identifying N-6-methyladenosine sites using pseudo dinucleotide composition. *Anal. Biochem.* 561–562, 59–65. doi: 10.1016/j.ab.2018.09.002

Chen, W., Feng, P. M., Ding, H., Lin, H., and Chou, K. C. (2015). iRNA-Methyl: identifying N-6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* 490, 26–33. doi: 10.1016/j.ab.2015.08.021

Chen, W., Feng, P. M., Yang, H., Ding, H., Lin, H., and Chou, K. C. (2018b). iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. *Mol. Ther. Nucleic Acids* 11, 468–474. doi: 10.1016/j.omtn.2018.03.012

Chen, W., Lv, H., Nie, F., and Lin, H. (2019b). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics (Oxf. Engl.)* 35, 2796–2800. doi: 10.1093/bioinformatics/btz015

Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K. C. (2016). iRNA-PseU: identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* 5:e332. doi: 10.1038/mtna.2016.37

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N-4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Cheng, L., and Hu, Y. (2018). Human disease system biology. *Curr. Gene Ther.* 18, 255–256. doi: 10.2174/1566523218666181101143116

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). 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., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018b). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19(Suppl. 1):919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019a). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48(Suppl. 1):gkz843. doi: 10.1093/nar/gkz843

Cheng, L., Sun, J., Xu, W. Y., Dong, L. X., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep34820

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019b). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief. Bioinform.* 20, 203–209. doi: 10.1093/bib/bbx103

Cheng, L., Zhuang, H., Ju, H., Yang, S., Han, J., Tan, R., et al. (2019c). Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front. Genet.* 10:94. doi: 10.3389/fgene.2019.00094

Cohn, W. E. (1960). Pseudouridine, a carbon-carbon linked ribonucleoside in ribonucleic acids: isolation, structure, and chemical characteristics. *J. Biol. Chem.* 235, 1488–1498. doi: 10.1002/jbmte.390020410

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Degenhardt, F., Seifert, S., and Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Brief. Bioinform.* 20, 492–503. doi: 10.1093/bib/bbx124

Ding, Y., Tang, J., and Guo, F. (2016a). Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information. *Int. J. Mol. Sci.* 17:1623. doi: 10.3390/ijms17101623

Ding, Y., Tang, J., and Guo, F. (2016b). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinformatics* 17:398. doi: 10.1186/s12859-016-1253-9

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

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

Fang, T., Zhang, Z., Sun, R., Zhu, L., He, J., Huang, B., et al. (2019). RNAm5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition. *Mol. Ther. Nucleic Acids* 18, 739–747. doi: 10.1016/j.omtn.2019.10.008

Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.-C. (2019). iDNA6mA-PseKNC: identifying DNA N-6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 111, 96–102. doi: 10.1016/j.ygeno.2018.01.005

Feng, P. M., Ding, H., Chen, W., and Lin, H. (2016). Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. *Mol. Biosyst.* 12, 3307–3311. doi: 10.1039/c6mb00471g

Fu, J., Tang, J., Wang, Y., Cui, X., Yang, Q., Hong, J., et al. (2018). Discovery of the consistently well-performed analysis chain for SWATH-MS Based pharmacoproteomic quantification. *Front. Pharmacol.* 9:681. doi: 10.3389/fphar.2018.00681

Fustin, J.-M., Doi, M., Yamaguchi, Y., Hida, H., Nishimura, S., Yoshida, M., et al. (2013). RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* 155, 793–806. doi: 10.1016/j.cell.2013.10.026

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

He, J. J., Fang, T., Zhang, Z. Z., Huang, B., Zhu, X. L., and Xiong, Y. (2018). PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics* 19:11. doi: 10.1186/s12859-018-2321-0

He, W., Jia, C., and Zou, Q. (2019). 4mCPred: machine learning methods for DNA N-4-methylcytosine sites prediction. *Bioinformatics* 35, 593–601. doi: 10.1093/bioinformatics/bty668

Helm, M. (2006). Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic Acids Res.* 34, 721–733. doi: 10.1093/nar/gkj471

Hu, Y., Zhao, T., Zhang, N., Zhang, Y., and Cheng, L. (2019). A review of recent advances and research on drug target identification methods. *Curr. Drug Metab.* 20, 209–216. doi: 10.2174/1389200219666180925091851

Jiang, L., Ding, Y., Tang, J., and Guo, F. (2018). MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association. *Front. Genet.* 9:618. doi: 10.3389/fgene.2018.00618

Karijolich, J., Yi, C. Q., and Yu, Y. T. (2015). Transcriptome-wide dynamics of RNA pseudouridylation. *Nat. Rev. Mol. Cell Biol.* 16, 581–585. doi: 10.1038/nrm4040

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, (Red Hook, NY: Curran Associates, Inc), 3146–3154.

Kong, L., and Zhang, L. (2019). i6mA-DNCP: computational identification of DNA N6-methyladenine sites in the rice genome using optimized dinucleotide-based features. *Genes* 10:828. doi: 10.3390/genes10100828

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

Li, B., Tang, J., Yang, Q., Li, S., Cui, X., and Li, Y. (2017). NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res.* 45, W162–W170. doi: 10.1093/nar/gkx449

Li, C.-C., and Liu, B. (2019). MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Brief. Bioinform.* bbz133. doi: 10.1093/bib/bbz133

Li, X. Y., Zhu, P., Ma, S. Q., Song, J. H., Bai, J. Y., Sun, F. F., et al. (2015a). Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.* 11, 592–597. doi: 10.1038/nchembio.1836

Li, Y. H., Zhang, G. G., and Cui, Q. H. (2015b). PPUS: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics* 31, 3362–3364. doi: 10.1093/bioinformatics/btv366

Li, Y. Z., Fan, Y.-X., and Yang, H.-H. (2018). KELMPSP: pseudouridine sites identification based on kernel extreme learning machine. *Chin. J. Biochem. Mol. Biol.* 34, 785–793. doi: 10.13865/j.cnki.cjbmb2018.07.14

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Li, C.-C., and Yan, K. (2019a). DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinform.* bbz098. doi: 10.1093/bib/bbz098

Liu, B., Liu, F. L., Wang, X. L., Chen, J. J., Fang, L. Y., and Chou, K. C. (2015a). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Liu, K., Chen, W., and Lin, H. (2019b). XG-PseU: an eXtreme gradient boosting based method for identifying pseudouridine sites. *Mol. Genet. Genomics* 295, 13–21. doi: 10.1007/s00438-019-01600-9

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015b). N-6-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* 518, 560–564. doi: 10.1038/nature14234

Liu, S., Zheng, B., Sheng, Y., Kong, Q., Jiang, Y., Yang, Y., et al. (2019c). Identification of cancer dysfunctional subpathways by integrating DNA methylation, copy number variation, and gene-expression data. *Front. Genet.* 10:441. doi: 10.3389/fgene.2019.00441

Lv, H., Dao, F.-Y., Guan, Z.-X., Zhang, D., Tan, J.-X., and Jiu-Xin, Tan (2019a). iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Front. Genet.* 10:793. doi: 10.3389/fgene.2019.00793

Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019b). A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

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

Meng, C. L., Jin, S. S., Wang, L., Guo, F., and Zou, Q. (2019). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7:10. doi: 10.3389/fbioe.2019.00224

Meyer, K. D., and Jaffrey, S. R. (2014). The dynamic epitranscriptome: N-6-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.* 15, 313–326. doi: 10.1038/nrm3785

Motorin, Y., and Helm, M. (2010). tRNA stabilization by modified nucleotides. *Biochemistry* 49, 4934–4944. doi: 10.1021/bi100408z

Mrozek, D., Gosk, P., and Malysiak-Mrozek, B. (2015). Scaling Ab initio predictions of 3D protein structures in microsoft azure cloud. *J. Grid Comput.* 13, 561–585. doi: 10.1007/s10723-015-9353-8

Mrozek, D., Malysiak-Mrozek, B., and Siaznik, A. (2013). search GenBank: interactive orchestration and ad-hoc choreography of web services in the exploration of the biomedical resources of the national center For biotechnology information. *BMC Bioinformatics* 14:18. doi: 10.1186/1471-2105-14-73

Mrozek, D., Socha, B., Kozielski, S., and Malysiak-Mrozek, B. (2016). An efficient and flexible scanning of databases of protein secondary structures. *J. Intell. Inf. Syst.* 46, 213–233. doi: 10.1007/s10844-014-0353-0

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 (Oxf. Engl.)* 35, 3831–3833. doi: 10.1093/bioinformatics/btz165

Nair, A. S., and Sreenadhan, S. P. (2006). A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation* 1, 197–202.

Pan, G., Tang, J., and Guo, F. (2017). Analysis of co-associated transcription factors via ordered adjacency differences on motif distribution. *Sci. Rep.* 7:43597. doi: 10.1038/srep43597

Qiu, W.-R., Jiang, S.-Y., Xu, Z.-C., Xiao, X., and Chou, K.-C. (2017). iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. *Oncotarget* 8, 41178–41188. doi: 10.18632/oncotarget.17104

Qiu, W. R., Sun, B. Q., Xiao, X., Xu, Z. C., and Chou, K. C. (2016). iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics* 32, 3116–3123. doi: 10.1093/bioinformatics/btw380

Sabooh, M. F., Iqbal, N., Khan, M., Khan, M., and Maqbool, H. F. (2018). Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. *J. Theor. Biol.* 452, 1–9. doi: 10.1016/j.jtbi.2018.04.037

Schwartz, S., Bernstein, D. A., Mumbach, M. R., Jovanovic, M., Herbst, R. H., and León-Ricardo, B. X. (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* 159, 148–162. doi: 10.1016/j.cell.2014.08.028

Shan, X., Wang, X., Li, C. D., Chu, Y., Zhang, Y., and Xiong, Y. I. (2019). Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method. *J. Chem. Inf. Model.* 59, 4577–4586. doi: 10.1021/acs.jcim.9b00749

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

Song, J. H., and Yi, C. Q. (2017). Chemical modifications to RNA: a new layer of gene expression regulation. *ACS Chem. Biol.* 12, 316–325. doi: 10.1021/acschembio.6b00960

Sun, W., Han, Y., Yang, S., Zhuang, H., Zhang, J., and Cheng, L. (2019). The assessment of Interleukin-18 on the risk of coronary heart disease. *Med. Chem.* doi: 10.2174/1573406415666191004115128 [Epub ahead of print].

Tahir, M., Tayara, H., and Chong, K. T. (2019). iPseU-CNN: identifying RNA pseudouridine sites using convolutional neural networks. *Mol. Ther. Nucleic Acids* 16, 463–470. doi: 10.1016/j.omtn.2019.03.010

Tang, H., Cao, R.-Z., Wang, W., Liu, T.-S., Wang, L.-M., and He, C.-M. (2017). A two-step discriminated method to identify thermophilic proteins. *Int. J. Biomath.* 10:1750050. doi: 10.1142/s1793524517500504

Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., and Yang, Q. (2019a). ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief. Bioinform.* 10:bby127. doi: 10.1093/bib/bby127

Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., et al. (2019b). Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol. Cell. Proteomics* 18, 1683–1699. doi: 10.1074/mcp.RA118.001169

Tang, W., Wan, S. X., 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

Wang, P., Zhang, X., Fu, T., Li, S., Li, B., Xue, W., et al. (2017). Differentiating physicochemical properties between addictive and nonaddictive ADHD drugs revealed by molecular dynamics simulation studies. *ACS Chem. Neurosci.* 8, 1416–1428. doi: 10.1021/acschemneuro.7b00173

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., and Han, D. (2014). N-6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* 505, 117–120. doi: 10.1038/nature12730

Wang, X., Zhu, X., Ye, M., Wang, Y., Li, C.-D., and Xiong, Y. (2019a). STS-NLSP: a network-based label space partition method for predicting the specificity of membrane transporter substrates using a hybrid feature of structural and semantic similarity. *Front. Bioeng. Biotechnol.* 7:306. doi: 10.3389/fbioe.2019.00306

Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., and Wang, Z. (2019b). Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* 48, D1031–D1041. doi: 10.1093/nar/gkz981

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2019a). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35, 1326–1333. doi: 10.1093/bioinformatics/bty824

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019b). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z., and Zou, Q. (2017a). CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Wei, L., Xing, P., Tang, J., and Zou, Q. (2017b). PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Trans. Nanobioscience* 16, 240–247. doi: 10.1109/TNB.2017.2661756

Wei, L. Y., Tang, J. J., and Zou, Q. (2017c). 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. Y., Wan, S. X., Guo, J. S., and Wong, K. K. L. (2017d). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L. Y., Xing, P. W., Zeng, J. C., Chen, J. X., Su, R., and Guo, F. (2017e). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Winkler, R., Gillis, E., Lasman, L., Safra, M., Geula, S., and Soyris, C. (2019). m(6)A modification controls the innate immune response to infection by targeting type I interferons. *Nat. Immunol.* 20, 173–182. doi: 10.1038/s41590-018-0275-z

Wu, X. Y., Wei, Z., Chen, K. Q., Zhang, Q., Su, J. L., Liu, H., et al. (2019). m6Acomet: large-scale functional prediction of individual m(6)A RNA methylation sites from an RNA co-methylation network. *BMC Bioinformatics* 20:223. doi: 10.1186/s12859-019-2840-3

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Sci.* 10(Suppl. 1):S20. doi: 10.1186/1477-5956-10-S1-S20

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., and Chang, C.-C. (2018a). An efficient classifier for Alzheimer's disease genes identification. *Molecules* 23:3140. doi: 10.3390/molecules23123140

Xu, L., Liang, G., Liao, C., Chen, G.-D., and Chang, C.-C. (2019). k-Skip-n-Gram-RF: a random forest based method for Alzheimer's disease protein identification. *Front. Genet.* 10:33. doi: 10.3389/fgene.2019.00033

Xu, L., Liang, G., Wang, L., and Liao, C. (2018b). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* 9:158. 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, 1128–1140. doi: 10.1021/acschemneuro.7b00490

Yin, J., Sun, W., Li, F., Hong, J., Li, X., and Zhou, Y. (2019). VARIDT 1.0: variability of drug transporter database. *Nucleic Acids Res.* 48, D1042–D1050. doi: 10.1093/nar/gkz779

Yu, H., and Dai, Z. (2019). SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Front. Genet.* 10:1071. doi: 10.3389/fgene.2019.01071

Yu, L., Huang, J. B., Ma, Z. X., Zhang, J., Zou, Y. P., and Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomics* 8:13. doi: 10.1186/1755-8794-8-s2-s2

Yu, L., Su, R., Wang, B., Zhang, L., Zou, Y., Zhang, J., et al. (2017a). Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. *IEEE ACM Trans. Comput. Biol. Bioinform.* 14, 966–977. doi: 10.1109/tcbb.2016.2550453

Yu, L., Yao, S., Gao, L., and Zha, Y. (2019). Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments. *Front. Genet.* 9:745. doi: 10.3389/fgene.2018.00745

Yu, L., Zhao, J., and Gao, L. (2017b). Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome. *Artif. Intell. Med.* 77, 53–63. doi: 10.1016/j.artmed.2017.03.009

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, 971–980. doi: 10.7150/ijbs.23350

Zaringhalam, M., and Papavasiliou, F. N. (2016). Pseudouridylation meets next-generation sequencing. *Methods* 107, 63–72. doi: 10.1016/j.ymeth.2016.03.001

Zhang, M., Li, F. Y., Marquez-Lago, T. T., Leier, A., Fan, C., Kwoh, C. K., et al. (2019). MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. *Bioinformatics* 35, 2957–2965. doi: 10.1093/bioinformatics/btz016

Zhang, M., Xu, Y., Li, L., Liu, Z., Yang, X., and Yu, D.-J. (2018). Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble. *Anal. Biochem.* 550, 41–48. doi: 10.1016/j.ab.2018.03.027

Zhao, B. S., and He, C. (2015). Pseudouridine in a new era of RNA modifications. *Cell Res.* 25, 153–154. doi: 10.1038/cr.2014.143

Zhao, X. W., Zhang, Y., Ning, Q., Zhang, H. R., Ji, J. C., and Yin, M. H. (2019). Identifying N-6-methyladenosine sites using extreme gradient boosting system optimized by particle swarm optimizer. *J. Theor. Biol.* 467, 39–47. doi: 10.1016/j.jtbi.2019.01.035

Zheng, G., Xue, W., Yang, F., Zhang, Y., Chen, Y., Yao, X., et al. (2017). Revealing vilazodone's binding mechanism underlying its partial agonism to the 5-HT1A receptor in the treatment of major depressive disorder. *Phys. Chem. Chem. Phys.* 19, 28885–28896. doi: 10.1039/c7cp05688e

Zhou, Y., Zeng, P., Li, Y. H., Zhang, Z. D., and Cui, Q. H. (2016). SRAMP: prediction of mammalian N-6-methyladenosine (m(6)A) sites based on sequence-derived features. *Nucleic Acids Res.* 44:e91. doi: 10.1093/nar/gkw104

Zhu, X., He, J., Zhao, S., Tao, W., Xiong, Y., and Bi, S. (2019). A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of *Saccharomyces cerevisiae*. *Brief. Funct. Genomics* 18, 367–376. doi: 10.1093/bfgp/elz018

Zou, Q., Xing, P. W., Wei, L. Y., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N-6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

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

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 © 2020 Lv, Zhang, Ding and Zou. 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 Genetics

ORIGINAL RESEARCH
published: 27 February 2020
doi: 10.3389/fgene.2020.00099

A FITM1-Related Methylation Signature Predicts the Prognosis of Patients With Non-Viral Hepatocellular Carcinoma

Jie Chen[1,2†], *Xicheng Wang*[1,2†], *Xining Wang*[1,2], *Wenxin Li*[3], *Changzhen Shang*[1,2], *Tao Chen*[1,2*] *and Yajin Chen*[1,2*]

[1] Department of Hepatobiliary Surgery, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China, [2] Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Medical Research Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China, [3] Department of Cardiology, The Eight Affiliated Hospital, Sun Yat-sen University, Shenzhen, China

OPEN ACCESS

Edited by:
Xiaotian Zhang,
Van Andel Research Institute (VARI),
United States

Reviewed by:
Rui Benfeitas,
Uppsala University,
Sweden
Biju Issac,
Leidos Biomedical Research, Inc.,
United States

***Correspondence:**
Tao Chen
ct0498@126.com
Yajin Chen
cyj0509@126.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 05 October 2019
Accepted: 29 January 2020
Published: 27 February 2020

Citation:
Chen J, Wang X, Wang X, Li W, Shang C, Chen T and Chen Y (2020) A FITM1-Related Methylation Signature Predicts the Prognosis of Patients With Non-Viral Hepatocellular Carcinoma. Front. Genet. 11:99. doi: 10.3389/fgene.2020.00099

Although great progress has been made in treatment against hepatitis virus infection, the prognosis of hepatocellular carcinoma (HCC) remains unsatisfied. Therefore, there is an unmet need to explore biomarkers or prognostic models for monitoring non-viral hepatocellular carcinoma. Accumulating evidence indicates that DNA methylation participates in carcinogenesis of malignancies. In the present study, we analyzed 101 non-viral HCC patients from TCGA database to figure out methylation-driven genes (MDGs) that might get involved in non-viral HCC pathogenesis using MethyMix algorithm. Then we picked out 8 key genes out of 137 MDGs that could affect the overall survival (OS) of both methylation and expression level. Using PCA, Uni-variate, Multi-variate, and LASSO cox regression analyses, we confirmed the potential prognostic value of these eight epigenetic genes. Ultimately, combined with immunohistochemistry (IHC), ROC, OS, and GSEA analyses, fat storage-inducing transmembrane protein1 (FITM1) was identified as a novel tumor suppressor gene in non-viral HCC and an applicable FITM1-methylation-based signature was built in a training set and validated in a testing set. Briefly, our work provides several potential biomarkers, especially FITM1, as well as a new method for disease surveillance and treatment strategy development.

Keywords: methylation-driven genes, non-viral hepatocellular carcinoma, FITM1, signature, nomogram

INTRODUCTION

Hepatocellular carcinoma (HCC) is a highly malignant tumor with high mortality and brings a great burden to social economy (Siegel et al., 2017). Chronic virus infection, commonly hepatitis B virus (HBV) and hepatitis C virus (HCV), and long-term alcohol consumption are the major etiology of HCC development (Braillon, 2012). Thanks for the development of vaccine and anti-virus medication treatment, the morbidity of virus-related HCC shows a decreasing tendency. Although viral hepatitis infection is strongly responsible for liver cancer progression, various non-viral risk factors play important roles in promoting HCC development (Alzahrani et al., 2014).

The epidemiological studies show that the incidence of HCC has failed to decline over the past decades partly owing to the increasing of HCC without virus infection. Thus, there is an unmet need to understand the underlying molecular mechanism of non-viral HCC. Due to the high heterogeneity and molecular diversities (Bruix et al., 2014), the prognosis of non-viral HCC patients is widely divergent. Therefore, an effective and accurate model to predict the prognosis of non-viral HCC individually is important and helpful to inform future clinical-decision making.

DNA methylation, one of the predominant forms of pre-transcriptional modification, has been widely studied in carcinogenesis (Noguchi et al., 1994; Chuang and Chiang, 2014). To date, great attention has been paid to investigate the relationship between methylation-driven genes (MDGs) and tumorigenesis (Pu et al., 2017). Recent studies have proved that MDGs participates in various lethal diseases like lung adenocarcinoma (Gao et al., 2018), pancreatic cancer (Gevaert et al., 2015), renal carcinoma (Zhang et al. 2019; Wang et al., 2020), and colon cancer (Chen et al., 2016). Meanwhile, many studies revealed that numerous genes are abnormally hypermethylated or hypomethylated in HCC (Hlady et al., 2019). Therefore, whether MDGs take part in the initialization and progression of non-viral HCC or not remains to be verified and a comprehensive understanding of several potential targets or biomarkers urges to be made.

Since long, methylation has been proved to negatively regulate gene expression, and DNA methylation is accountable for multiple cancers, including HCC (Revill et al., 2013; Hlady et al., 2019). Recently, Sun et al. has revealed that a novel gene signature (CTHRC1 expression, ZIC4 expression, and OTX1expression) may be regulated by DNA hypermethylation and closely associated with HCC through weighted correlation network analysis (WGCNA) (Sun et al., 2018). A 21-gene pairs signature was established to predict HCC patients at their early stages through the C-index forward search method (Liu et al., 2018). In addition, six MDGs, including SNHG6, S100P, DCDC2, LIME1, FMO3, and GPR171, have been selected to construct a predictive signature for HCC patients and the contribution of virus infection has been highly emphasized in their work (Li et al., 2019). Wang et al. also constructed a risk score system consisting of BRCA1 expression, CAD expression, RBM8A expression and CDC20 expression by using four GSE data (Wang et al., 2019). However, a novel methylation prognostic signature for non-viral HCC still remains undeveloped and a systematic exploration of non-viral HCC signature is needed. To our knowledge, the eight MDGs studied in present work has never been recruited to be part of a score signature in HCC, indicating that they might exert important effect on the tumorigenesis and development of non-viral HCC specifically.

In current study, we utilized an integrative method, including MethyMix tool, principal component analysis (PCA), nomogram algorithm, and least absolute shrinkage and selection operator (LASSO) regression analysis, to explore prognosis related to MDGs in non-viral HCC and validate the efficacy of the built methylation-related risk signature, providing a novel direction for treatment and surveillance strategy and personalized follow-up for non-viral HCC patients.

METHODS AND MATERIALS

Data Processing and Analysis

The RNA-seq data, methylation data, and corresponding clinicopathological information were retrieved for 101 non-viral HCC patients from TCGA database. Clinicopathological features for the TCGA datasets were described in **Supplementary Table S1**. On the basis of the MethylMix algorithm (Gevaert, 2015; Cedoz et al., 2018), we analyzed the correlation between gene methylation and expression level in 121 non-viral HCC samples. Due to the strict constraints of MethylMix algorithm, we set the parameters as followed: Adjust P-value< 0.05; Log FC (Fold Change) > 0 or Log FC< 0; Pearson correlation threshold< -0.3. Then, we identified aberrantly hypomethylated or hypermethylated genes by constructing the β-mixed model. Finally, according to the overall survival analysis results, we filtered most MDGs and obtained several key genes for further study. The mRNA expression and methylation data of non-viral HCC provided by TCGA is open-access and the approval of a local ethics committee is unneeded.

Gene Ontology (GO), Disease Ontology (DO), and KEGG Pathway Enrichment Analyses

In the present study, the clusterProfiler package (version 3.12.0) was applied to conduct GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses (Yu et al., 2012). The GO analysis includes cellular composition (CC), molecular function (MF), and biological process (BP). Disease ontology (DO) annotates genes based on human disease. DO is vital annotation, translating obtained key genes to clinical relevance. And DOSE, an R package, is capable of analyzing semantic similarity computations of the DO terms and genes. Therefore, DOSE enables us to figure out the closeness between diseases and gene functions (Yu et al., 2015). To investigate the underlying mechanism of these MDGs, 137 MDGs were subjected to clusterProfiler and DOSE packages for GO, KEGG, and DO analyses; and P-value < 0.05 was set as the cutoff.

Kaplan-Meier Curves of MDGs and Methylated Sites

For the sake of studying the prognostic evaluation of MDGs, the survival R package (version3.5.1) was used to calculate the prognostic survival analysis of the gene expression, gene mean methylation level and the methylated sites, performed by integrating the clinical data and prognostic information of non-viral HCC in TCGA. Meantime, we conducted a joint survival analysis of gene methylation and expression levels to further determine key genes associated with prognosis in non-viral HCC patients.

PCA Analysis and Subgroup Analysis

To study the function of eight methylation-driven key genes in non-viral hepatocellular carcinoma, we separated 101 non-viral HCC patients into different subgroups by the approach of "ConsensusClusterPlus", an algorithm for determining clusters

by the unsupervised analysis based on gene expression (Wilkerson and Hayes, 2010). The consensus clustering tool presents measurable and visible evidence to estimate the counts of unsupervised classes in a dataset. The maximum evaluated k (max K) is 9 and other parameters of ConsensusClusterPlus are default. As a result, Subgroup 1 had 59 non-viral HCC patients while Subgroup 2 had 41 patients. And when the data were classified into three subgroups, there were 33, 36, and 31 patients in Subgroup 1, 2, and 3, respectively. The R package (R v3.5.1) of PCA analysis was adopted to explore the gene methylation patterns in subgroups of non-viral HCC. Gene Sets Enrichment Analysis (GSEA) was performed by GSEA 4.0.0 software to explore the specific KEGG pathways related to distinct subgroups of non-viral HCC and the underlying function of FITM1 (Subramanian et al., 2005). Regarding the GSEA results, |NES|> 1 and *P*-value < 0.05 were considered significant in our study.

Construction of MDGs Signature

Caret R package (Classification and Regression Training; Version:6.0-84) can provide a wide variety of predictive models by integrating more than 25 other relative packages and has various unique features such as data splitting, characterizing performance, pre-processing, parallel processing, and variable importance (Kuhn, 2008). Due to the lack of other datasets with integrative data of non-viral HCC patients (epigenomics, transcriptomics, and clinical pathologic data), we stochastically divided the 101 non-viral HCC patients into 2 sets, training set (52 patients), and testing set (49 patients). The classification was based on the caret R package. To confirm the prognostic value of 8 MDGs, the Uni-variate cox regression, LASSO cox regression, and Multi-variate cox regression algorithms were performed in the training set and a potential risk signature of non-viral HCC was developed (Bøvelstad et al., 2007; Qiu et al., 2017; Wang et al., 2019). The risk score for the signature (Lossos et al., 2004) was computed using the formula:

$$risk\ \ score = \sum_{i=1}^{n} Coef_i \star X_i$$

As the formula shown above, "X" represents the methylation level of each methylation-driven gene in the samples; The "Coef" means corresponding Multi-variate cox regression coefficient of each factor in the prognostic model. The value of "n" in our study is smaller than 8. On the other hand, a nomogram of non-viral HCC patients was constructed based upon the results of the LASSO cox regression analysis using rms package (version 3.5.1). The prognostic risk value of each patient was calculated using the formula, and the median of the score value was cut off. The non-viral HCC patients were classified into high and low risk groups. Then, we conducted ROC curve and Kaplan-Meier survival curve analyses to validate the signature in both the training set and testing set. Log-rank test was applied to figure out the difference of overall survival rate between the high-risk and low-risk groups. "$P < 0.05$" was considered statistically significant.

RESULTS

Identification and Functional Analyses of MDGs in Non-Viral Hepatocellular Carcinoma

The flow diagram for present study was exhibited in **Figure 1**. After downloading the comprehensive data of 101 non-viral hepatocellular carcinoma patients, the MethylMix algorithm mentioned above was adopted to figure out 137 MDGs in non-viral HCC (**Figure 2A** and **Table S2**). To elucidate the potential function of these genes, GO, KEGG, and DO analyses were carried out. As shown in **Figure 2B**, the GO top significant terms were various, and some of them were as followed: "lipid localization", "cholesterol homeostasis", "lipid homeostasis", "sterol homeostasis", "lipid storage regulation of lipoprotein", "lipoprotein particle", and "protein-lipid complex", which indicated that the aberrant methylation level of 137 MDGs may cause abnormal lipid metabolism, one of the most pivotal function of normal liver. In addition, KEGG analysis revealed that these 137 MDGs were significantly enriched in pathways in "Glutathione metabolism", "Aldosterone-regulated sodium reabsorption", "Fat digestion and absorption", and "Cholesterol metabolism", consistent with the result of GO analysis. "p53 signaling pathway", "HIF-1 signaling pathway", and "EGFR tyrosine kinase inhibitor resistance" were also enriched, suggesting the potential regulating signaling pathway of non-viral HCC by MDGs (**Figure 2C**). In addition, for the sake of investigating the relationship between137 MDGs and human diseases. DO analysis was applied. As shown in **Figure 2D**, these genes might be involved in the following DO terms: "lipodystrophy", "fatty liver disease", "liver cirrhosis", "obesity", and so on. Complete data of the enrichment analyses above were displayed in **Tables S3**, **S4**, **S5**. Taken together, these results indicate that 137 MDGs might participate in the carcinogenesis of non-viral HCC through the regulation of liver lipid metabolism and chronic liver injury.

Screening and Verifying Survival-Related Key Genes Among 137 MDGs

Figuring out some oncogenes, which act importantly on hepatocarcinogenesis as well as the progression of non-viral HCC, is of great significance. Therefore, we performed not only the overall survival analysis of gene expression but also the joint survival analysis, which analyzes the OS combining the level of expression and corresponding methylation together, across 137 genes in 101 non-viral HCC patients. Thereafter, discarding those results without significant difference (*P*-value > 0.05) from both expression OS and joint OS, eight key genes were selected for further study: FITM1, FES, ABCG5, GPX7, FURIN, BSCL2, B3GALNT1, and GPAM. As shown in **Figure 3A**, low expression of FITM1, ABCG5, BSCL2, and GPAM in non-viral HCC tumor specimens generally predicted worse survival status. However, the high expression of FES, GPX7, FURIN, and B3GALNT1 leaded to shorter survival time. On the other hand, as presented in **Figure 3B**, the hypomethylation and high expression of FES, GPX7, FURIN, and B3GALNT1 were

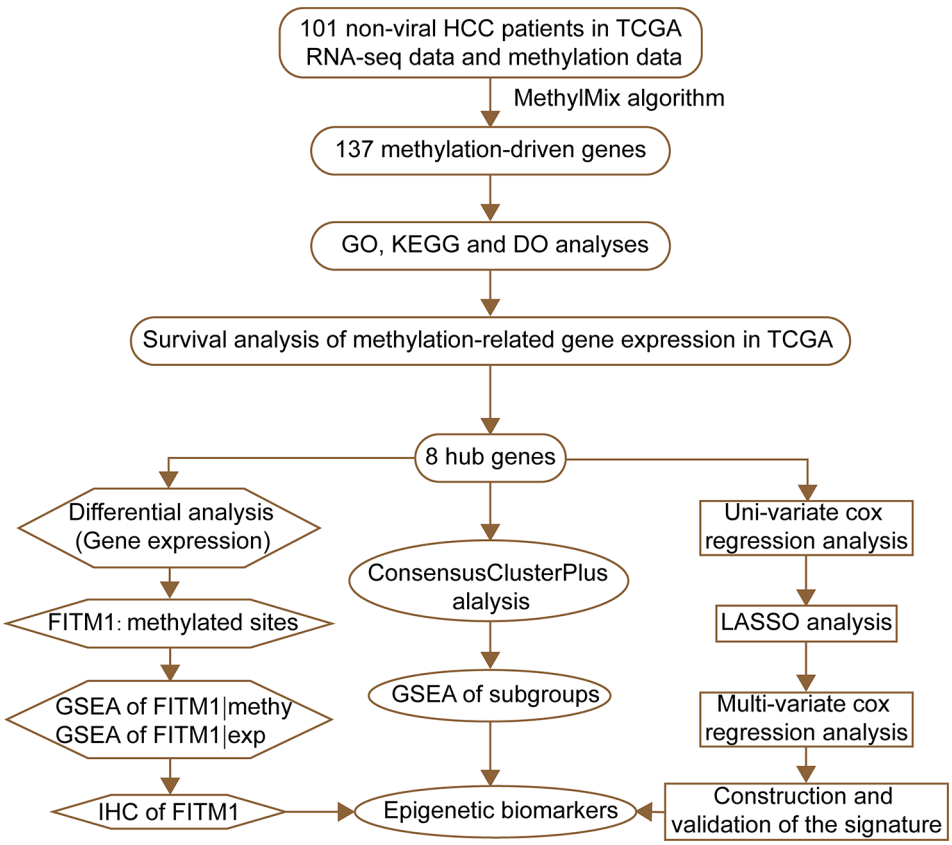


FIGURE 1 | The flowchart of this study.

related to poorer overall survival while the adverse results could be obtained from the survival curves of FITM1, ABCG5, BSCL2, and GPAM. In all, these results reveal that the methylation and expression level of these epigenetic genes could affect the prognosis of non-viral HCC patients.

Abnormal DNA Methylation of Eight Key Genes in Tumor Tissues Negatively Regulated Gene Expression

To obtain a deeper understanding of DNA methylation and mRNA expression of eight key driver genes, correlation analyses between methylation level and gene expression were employed. According to **Figure 4A**, along with the increase of the methylation degree, the key gene expression showed a downward trend, suggesting the negative correlation between DNA methylation and gene expression. Notably, we found that the aberrant methylation degree was much higher in tumor rather than normal tissues. Significantly, only 1 methylation curve of FITM1 or GPAM was gathered, indicating that the hypermethylated status of FITM1 and GPAM were centralized and common in tumor samples (**Figure 4B**). And FES, ABCG5, GPX7, FURIN, BSCL2, and B3GALNT1 had 2 methylation curves and the comparison of methylation level in tumors and normal tissues was ambiguous, driving us to elucidate the gene methylation and expression level between malignant samples and normal samples in non-viral HCC patients.

FITM1 Was Significantly Down-Regulated in Non-Viral HCC and Was a Potential Tumor Suppressor Gene

To illustrate the clear distribution and different expression of eight key genes in the normal and tumor samples of non-viral HCC, a thermal map was performed based on the gene methylation and corresponding expression. As vividly shown in **Figure 5A**, we found that FITM1, BSCL2, B3GALNT1, and GPX7 also had significantly different expression in comparison tumors with normal specimens. However, only FITM1 expression was significantly down-regulated in tumor specimens (Log FC = -1.74, *P*-value = 8.75E-09) while other three were over-expressed. Moreover, the FITM1 methylation was highly up-regulated in tumor specimens (Log FC = 0.49, *P*-value = 2.00E-09). This significantly negative correlation between methylation and expression in tumor as well as normal tissues triggered us to further explore the specific function of FITM1. For the sake of identifying the most worthwhile MDGs related to prognosis of non-viral HCC, we explored the characteristics of CpG methylated sites of FITM1. The methylation degree of cg20306574 methylated sites was negatively correlative with FITM1 expression level

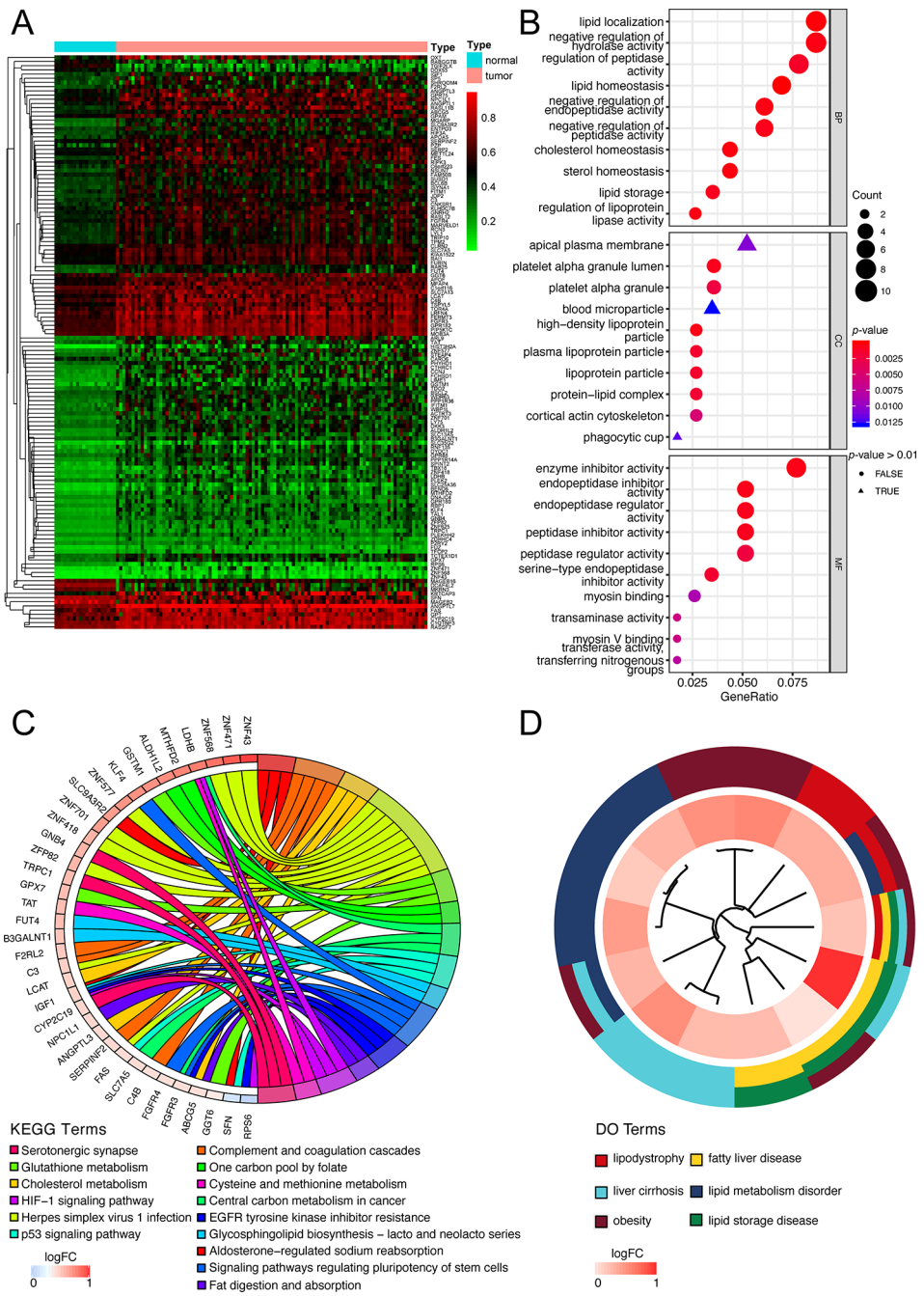


FIGURE 2 | Functional exploration of MDGs. **(A)** Heatmap of 137 aberrant MDGs in non-viral hepatocellular carcinoma. The green color stands for hypomethylation while the red shows hypermethylation. **(B)** Gene Ontology (GO) analysis of 137 MDGs. Only top 10 terms of BP, CC, and MF were shown and the complete data were in **Table S3**. **(C)** KEGG pathway analysis of 137 MDGs. The color of curves represents different KEGG terms; The left semi-circle color means different gene expression and the corresponding genes are labeled. The *P*-value of all terms is < 0.05. **(D)** Disease Ontology (DO) analysis of 137 MDGs. The inner circle is composed of different genes and their expression (LogFC) while the outer circle consists of different DO terms. The *P*-value of shown DO terms is < 0.05. Only liver-related terms were exhibited and the whole results were in **Table S5**.

(**Figure 5B**). Besides, high cg20306574 methylation predicted poorer prognosis in 101 non-viral HCC patients (**Figure 5C**). To investigate the anti-tumor mechanism of FITM1, the GSEA analysis was applied to analyze the methylation data and the mRNA expression data of 101 non-viral HCC patients in TCGA. As displayed in **Figure 5D**, we figured out that low FITM1 expression could activate cancer-related pathway. In addition, several metabolism-related signaling pathways might be able to

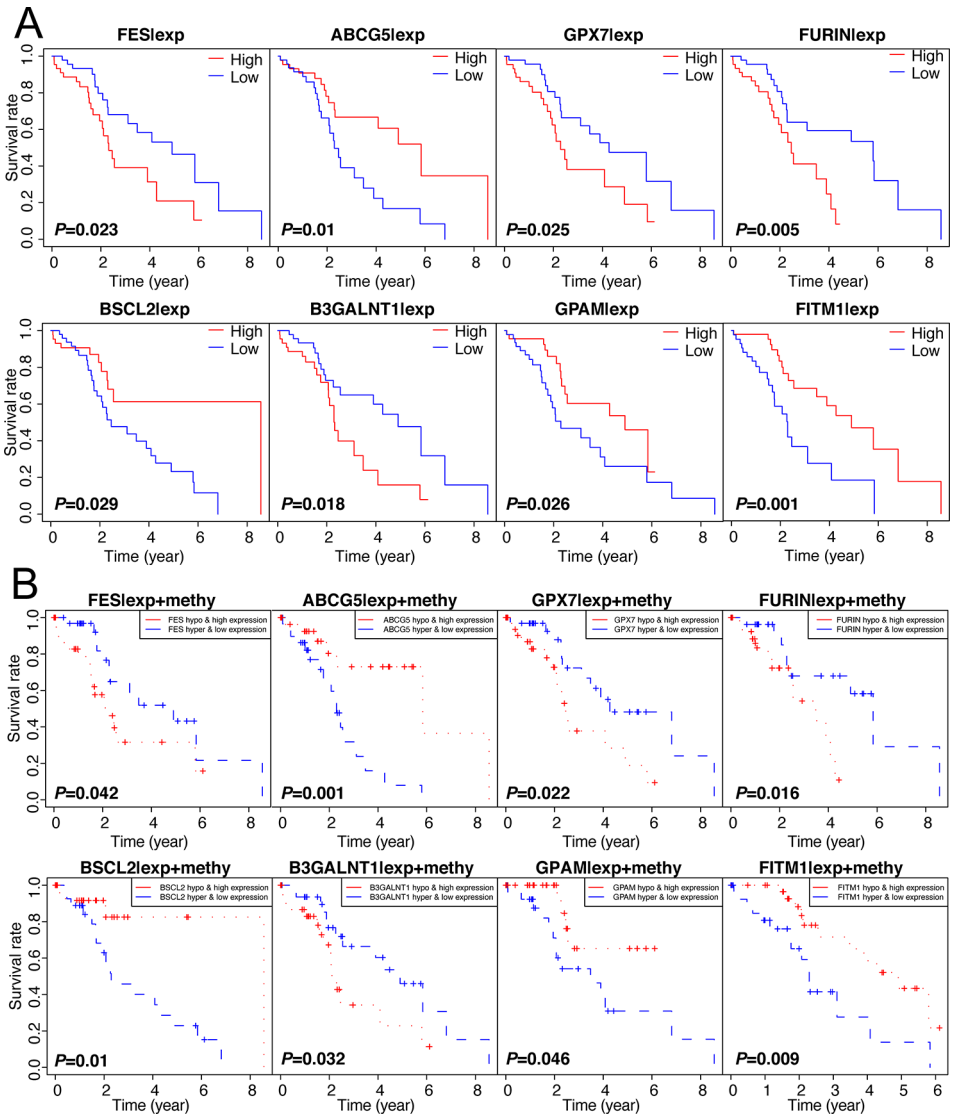


FIGURE 3 | Survival analysis of 137 genes and selection of 8 methylation-driven key genes. **(A)** 137 genes were analyzed by survival analysis in 101 non-viral HCC specimens and only eight key genes were obtained according to the specific cut-off (*P*-value < 0.05). **(B)** Joint survival analysis of 137 oncogenes. we performed survival curve analysis based on the combination of methylation level and expression level and the prognosis-related gene were selected (*P*-value < 0.05).

account for the anti-tumors effect of FITM1, indicating the underlying function of FITM1 in non-viral HCC (**Table S6**). Regarding the gene methylation level of FITM1, one of the enriched results pointed toward the NOTCH signaling pathway. Moreover, several lipid-related metabolism pathways were also enclosed (**Table S7**), partly consistent with the GO results of 137 MDGs in **Figure 2B**. As for the immunohistochemistry results (IHC) of FITM1 obtained from the Human Protein Atlas database (https://www.proteinatlas.org) (Uhlen et al., 2015), **Figure 5F** vividly shows that FITM1 expression was much higher in normal liver tissues rather than HCC tissues. Briefly, these aforementioned results indicate that FITM1 is closely related to non-viral HCC.

ConcensusCluster Analysis Revealed That Cluster 1 Might Regulate the NOTCH and TGF-Beta Signaling Pathways

Considering that the eight survival-related MDGs might exert its regulatory effect on non-viral HCC synergistically, we performed the correlation and subgroup analysis among the methylation level of eight key genes. As shown in **Figure 6A**, the methylation degrees of eight key genes were positively relative to each other, especially FITM1, BSCL2, and FES, which were significantly correlated with other seven key genes respectively. Furthermore, ConcensusCluster analysis was utilized to classify the tumor samples based upon the methylation level similarity of the eight MDGs. As revealed in **Figure 6B** and **Figure S1A**, k = 2

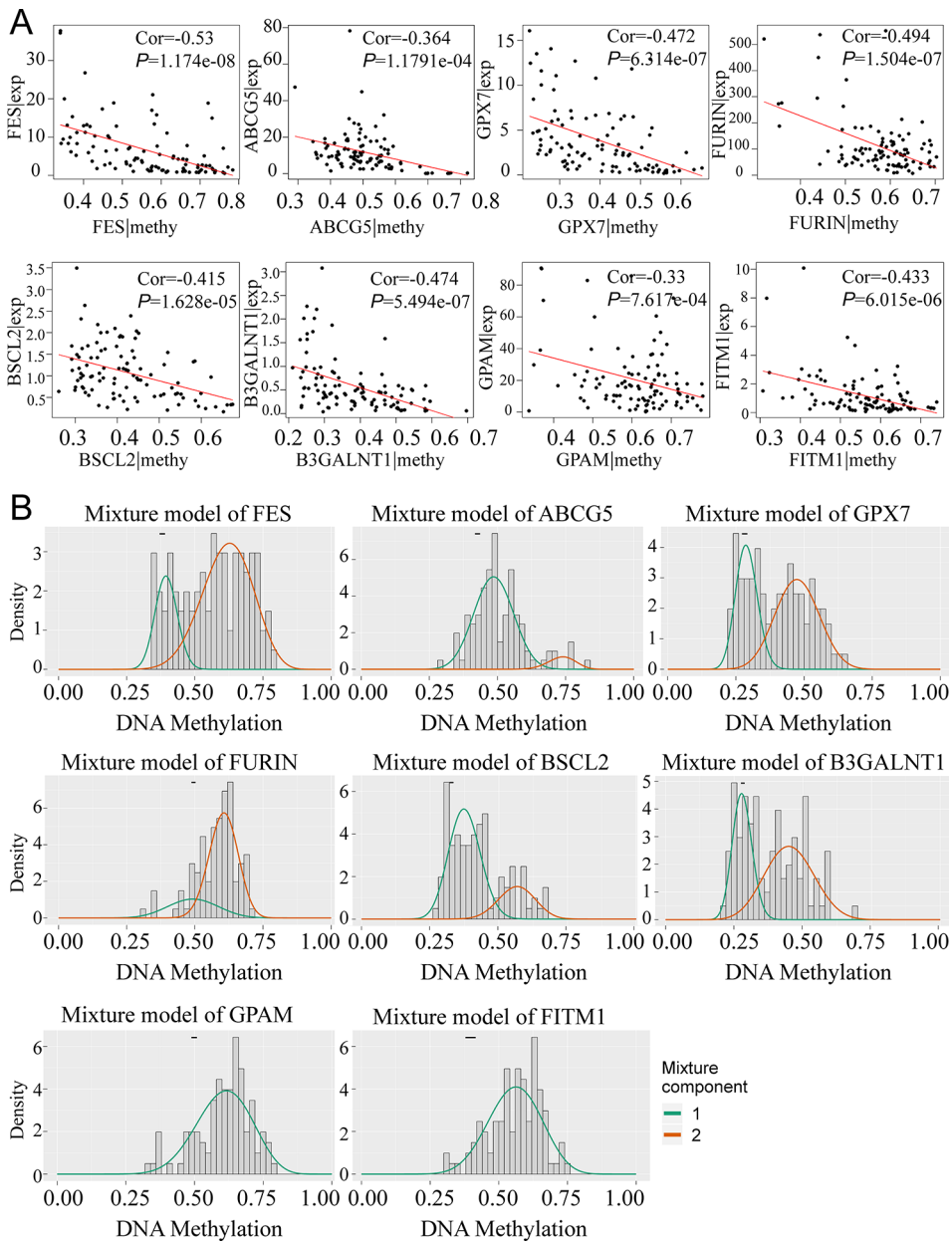


FIGURE 4 | The relationship between eight genes expression and methylation level in non-viral HCC patients. **(A)** The correlation between gene expression and methylation. **(B)** MethylMix model of eight DNA methylation-driven key genes. The abscissa is the methylation degree and the ordinate is the number of sample. The histogram exhibits the distribution of methylation in 101 non-viral HCC samples and the short black bar above the histogram is the methylation distribution of 20 normal tissues. The classification of different methylation degree of the malignancies relative to the normal tissues can be distinctly observed from the figures.

was selected as a reasonable choice with cluster stability rising from k = 2 to 10 in the non-viral HCC dataset. However, the CDF curve revealed that k = 3 was also preferable (**Figure 6C**). Therefore, we used PCA analysis to investigate the characteristic of methylation profile based on the classification of both k = 2 and k = 3 in 101 non-viral HCC patients. The results exhibited an evident difference between 101 non-viral HCC patients in both the 2D and 3D plotting of PCA results according to the two-subgroup classification (**Figures 6D, E**). And the two-subgroup classification could also well distinct tumor samples from normal samples (**Figures S1D, E**). Meantime, as illustrated in **Figures S1B, C**, 3-subgroup classification of 101 non-viral HCC patients was also capable of classifying non-viral HCC patients while the overlaid part was more than two-subgroup classification (**Figures S1B, C**). In all, the classification built on the methylation level of eight key genes might be more

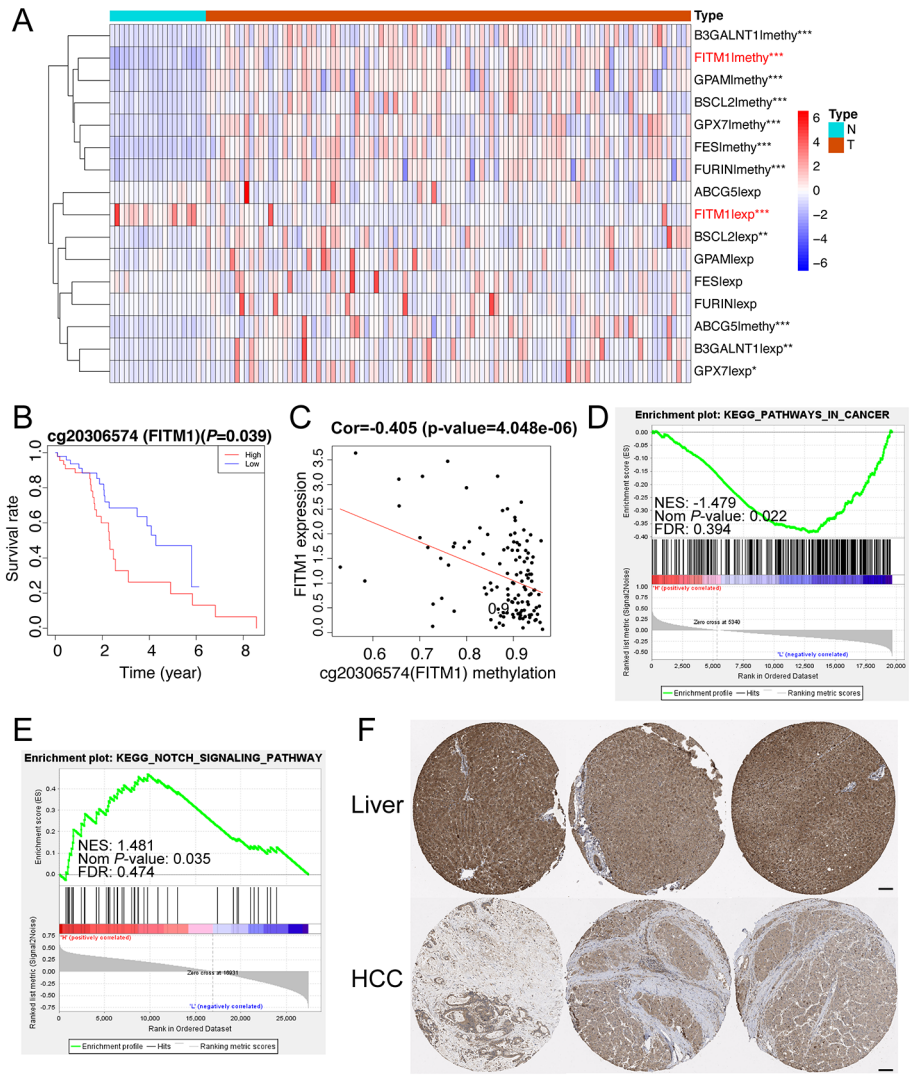


FIGURE 5 | Function and mechanism of FITM1 in non-viral HCC. **(A)** The thermal map of eight gene expression levels and eight gene methylation levels. The differential analysis between 20 normal samples and 101 tumor samples were conducted by Limma R package (version: 3.42.0). "*", "**", and "***" stands for "*P*-value < 0.05", "*P*-value < 0.01", and "*P*-value < 0.001", respectively. The thermal map was drawn with the pheatmap R package (version: 1.0.12) and the row scale (Z-score) was chosen to better visualize the related data. **(B)** The correlation of FITM1-related CpG methylated site and its gene expression. **(C)** The survival analysis of FITM1-related CpG methylation. **(D)** The GSEA result based on FITM1 expression level in 101 non-viral HCC patients. **(E)** The GSEA result based on FITM1 methylation level in 101 non-viral HCC patients. **(F)** The IHC of FITM1 retrieved from the Human Protein Atlas database (https://www.proteinatlas.org). Scale bars:100um.

distinguishable when the non-viral HCC patients were divided into two subgroups. In addition, the GSEA analysis was further conducted to explore the hallmark of two-subgroup classification. As presented in **Figure 6F**, we found that Cluster 1 was closely related to NOTCH and TGF-beta signaling pathways, both playing vital roles in malignancies. In all, we demonstrate that the carcinogenesis of patients in Cluster 1, rather than Cluster 2, might be involved in NOTCH and TGF-beta signaling pathways.

A Risk Signature Was Established in Training Set Using Three Selected Key Genes Related to DNA Methylation and Prognosis

In order to establish and validate a risk signature for non-viral HCC patients, Caret R package (Version:6.0-84) were conducted to randomly allocate 101 non-viral HCC patients into two sets: training set and testing set. As a result, 52 non-viral HCC

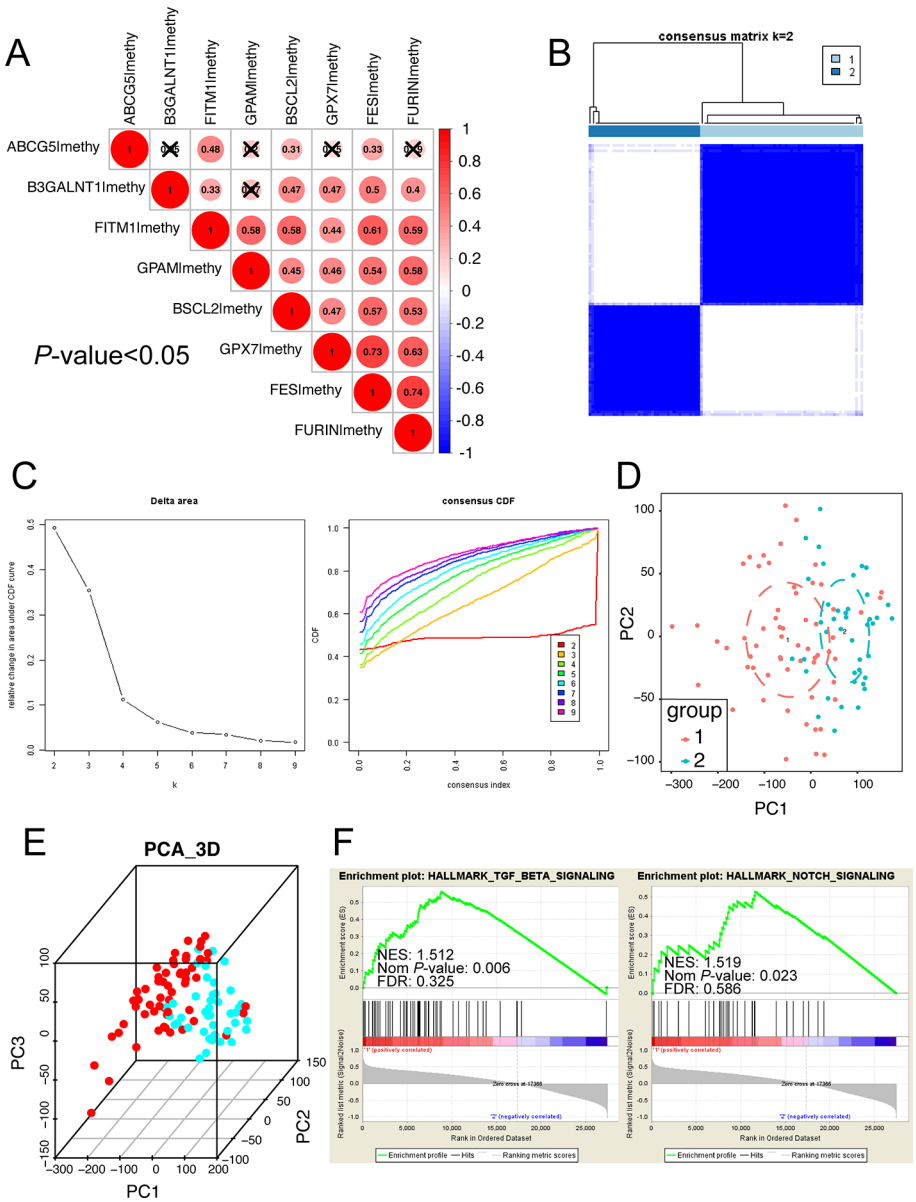


FIGURE 6 | Subgroup analysis developed on the methylation level of eight key MDGs. **(A)** The methylation relationship among eight key genes. The bigger the circle size, the more correlative two genes are. **(B)** Consensus matrix of two subgroups (k = 2). The k = 3 to 10 of the consensus matrix were shown in **Figure S1**. This study distinctly separated the whole methylation data into two subgroups: cluster 1 and cluster 2. **(C)** Classification of consensus clusters by 8 key MDGs. Consensus clustering cumulative distribution function (CDF) was set from k = 2 to 10. **(D)** Principal component analysis (PCA) of the total methylation level in 101 non-viral HCC patients based upon the consensus clustering. Non-viral HCC patients in different clusters are noted with different colors. **(E)** The 3D PCA of two subgroups according to 101 non-viral HCC patients. The 3D PCA of three subgroups were displayed in **Figure S1**. **(F)** The significant GSEA analysis results of cluster 1. The full GSEA data of clusters were included in **Table S8**.

patients were included in the training set while 49 patients in the testing set. To better predict the clinical outcomes of non-viral HCC with the eight MDGs, we used the Uni-variate, LASSO, and Multi-variate cox regression algorithm to build the risk signature according to the minimum criteria. As shown in **Figure 7A**, four out of eight key genes were chosen as risk factors using Uni-variate analysis (*P*-value< 0.05). In order to further confirm the result of Uni-variate analysis, these 4 genes were imported into the LASSO algorithm. **Figures 7B, C** show that both 4 or 3 key genes were reliable to construct a risk signature. Finally, Multi-variate analysis was used to construct the risk signature and only three risk factors were significantly chosen (**Figure 7D**). The coefficients of risk factors were retrieved from the Multi-variate analysis and the formula was as followed: risk score = 4.37 *

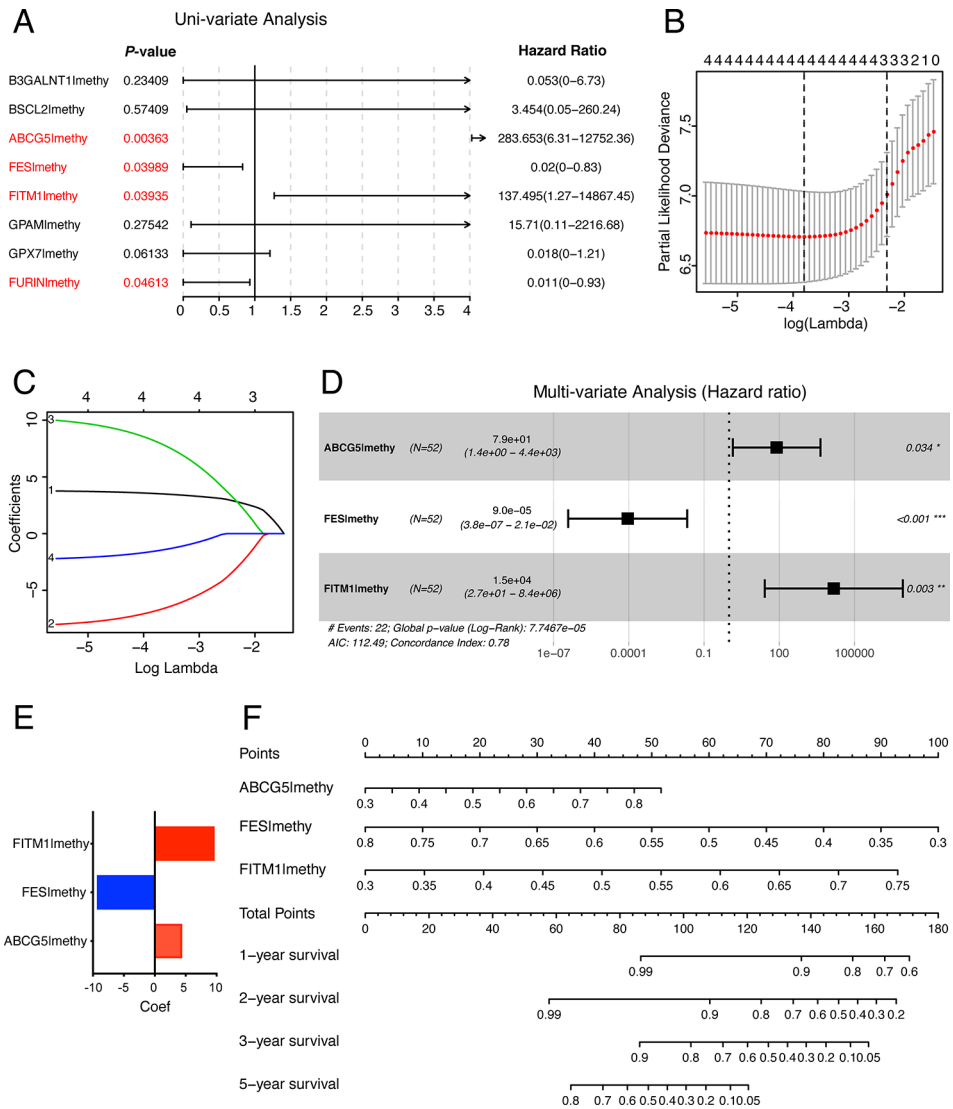


FIGURE 7 | The construction of the score model in training set**. (A)** Uni-variate analysis of eight key MDGs in the training set of non-viral HCC patients. The genes in red (*P*-value < 0.05) were selected and imported into the LASSO algorithm to conduct risk score model for non-viral HCC patients. **(B**, **C)** LASSO analysis of four selected genes in training set. **(D)** Multi-variate analysis of four selected genes in training set and only 3 out of 4 were significantly imported into the score model. **(E)** Our risk score formula obtained from discovery cohort was as followed: risk score = 4.37 * methylation of ABCG5- 9.31 * methylation of FES + 9.61 * methylation of FITM1. **(F)** Prediction of OS in non-viral HCC based upon nomogram. Three factors were included in this nomogram. The methylation level of these four genes could be used to create points according to the scale plotted upward. And the total points could point to the corresponding probability of 1-year, 2-year, 3-year, and 5-year OS rate drawn on the three lines below.

methylation of ABCG5- 9.31 * methylation of FES + 9.61 * methylation of FITM1 (**Figure 7E**). Notably, we found that FITM1 was also recruited in the risk signature (**Figure 7E**), indicating the pivotal role of FITM1 in non-viral HCC. To further explore a driver genes model that could serve as an independent prognosticator for non-viral HCC patients, a visualized and applicable nomogram was built based on three key genes selected by cox regression analyses applied above (**Figure 7F**).

Prognostic Risk Scores Exhibited Strong Predictive Power in the Prognosis of Non-Viral HCC Patients in Both Training Set and Testing Set

To better understand the function of this risk signature, **Figures 8A, D** were plotted and it could show the explicit relationship between the risk score, survival status, and methylation level in training set and testing set. Not only did the high-risk group in training set have significantly worse OS than the low risk group, the high-risk group

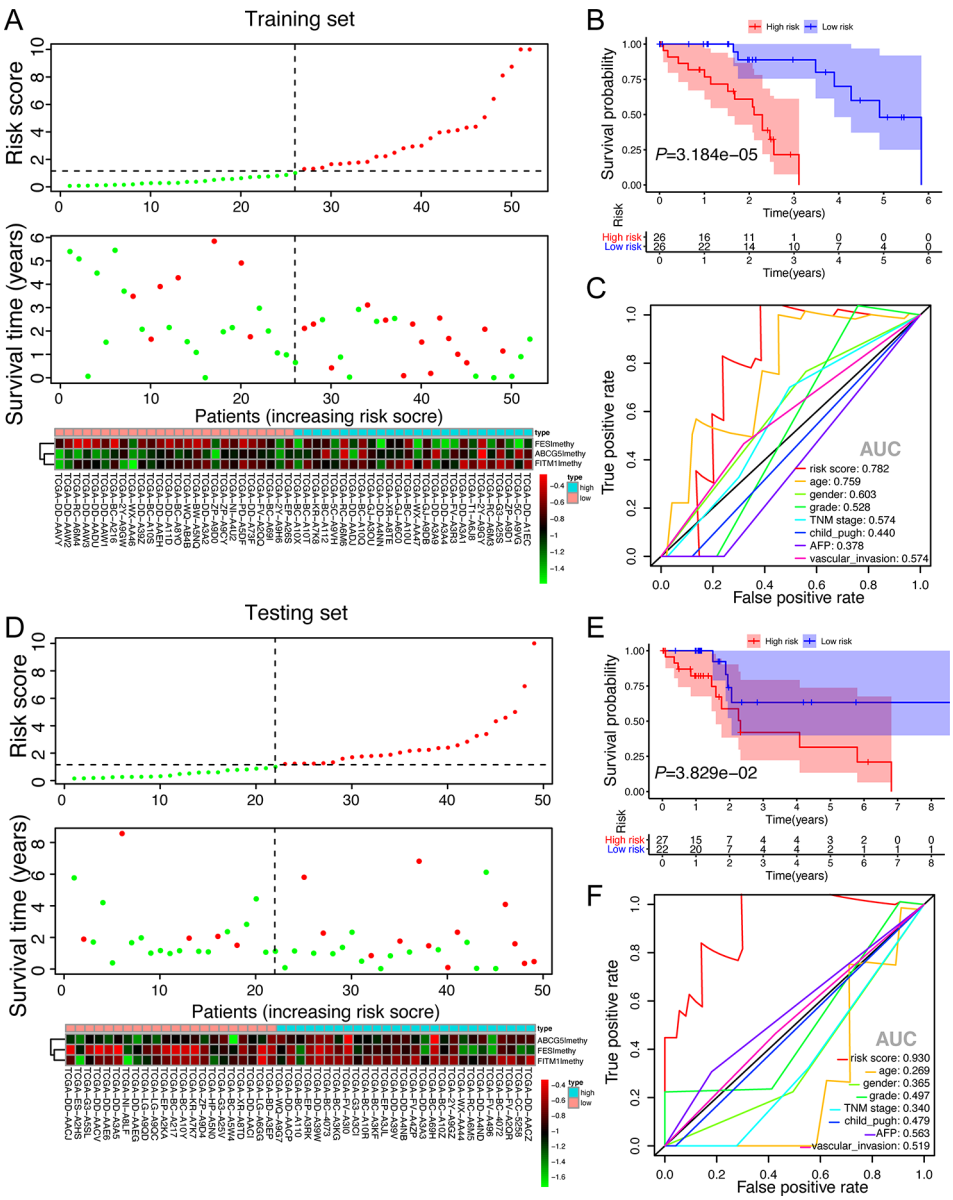


FIGURE 8 | The validation of the score model in training set and testing set. **(A, D)** Risk score distribution of non-viral HCC patients, survival curve and methylation heatmap of the three factors of the score model corresponding to each patient in training set (52 patients) and testing set (49 patients). The risk score was all calculated by the score model built on training set. Red color stands for high risk and hypermethylation; blue color means low risk and hypomethylation. **(B, E)** Kaplan-Meier survival analysis of the low and high-risk group. **(C, F)** ROC curves for a risk score model and several complete clinicopathological information of non-viral HCC patients retrieved from TCGA.

in testing set presented the similar phenomenon according to the survival curve in **Figures 8B, E**. In order to find out whether the risk signature was an effective prognostic indicator, receiver operating characteristic curve (ROC) was plotted. The ROC curves showed that the risk score model was able to predict 5-year survival rates for non-viral HCC patients both in training set (AUC = 78.2%) and in testing set (AUC = 93.0%). The predictive power of this model was better than other clinicopathological features included (**Figures 8C, F**). These results suggest that the risk score developed from three key genes could independently predict prognosis in non-viral HCC.

DISCUSSION

Virus infection is considered as one major cause of hepatocarcinogenesis (Lee et al., 2019). With the rapid development of anti-viral therapies, the virus-related HCC, especially HBV and HCV, is reducing all over the world. However, the incidence of hepatocellular carcinoma remains high and the recurrence rate still render it one of the most lethal malignancies. Other factors like diabetes, non-alcoholic steatohepatitis, non-viral hepatitis, tobacco smoking, obesity,

and dietary exposures are accountable for the etiology and progression of non-viral HCC (Dhanasekaran et al., 2016; Ogunwobi et al., 2019). And other potential gene-related causes of non-viral HCC, including TNFα, IL6, mTOR, MAPK, and NF-κB, are garnering close attention (Alzahrani et al., 2014). However, the specific mechanism of non-viral HCC is still unclear. It is of importance to figure out potential biomarkers, score signatures, and even the underlying specific targets to identify, monitor and treat non-viral HCC patients.

In our work, we downloaded methylation and corresponding expression data in TCGA database to explore and retrieve about 137 MDGs by MethyMix analysis. Enrichment analyses of the 137 MDGs indicated they participated in principal biological processes of liver, including lipid metabolism, cholesterol metabolic process, and lipid homeostasis. Given that disorder lipid metabolism is closely associated with non-alcoholic fatty liver disease (NAFLD) and has been growingly considered a hallmark of cancer cells (Cingolani and Czaja, 2016; Tian et al., 2019), It is reasonable that MDGs exerted great impact on non-viral HCC through regulation of lipid metabolism and homeostasis. In addition, KEGG analysis suggested that EGFR and P53 pathways were also significantly enriched, which were involved in the pathogenesis and tumorigenesis of HCC (Xie et al., 2018; Huang et al., 2019).

Based on gene expression overall survival and joint overall survival analysis, we figured out eight key genes. B3GALNT1 is involved in the tumorigenesis of lung adenocarcinoma (Aubry et al., 2015); BSCL2 is over-expressed in the better progression-free and overall survival group of high-grade serous ovarian cancer (Cuello et al., 2017) and may take part in regulating lipid storage in adipocytes and inhibiting ectopic lipid droplet formation in cancer cells (Salo et al., 2016); ABCG5, which could regulate the transport of hydrophobic mixtures, especially lipids, across cellular membranes, is hypermethylated in prostate cancer (Kerr et al., 2011; Devaney et al., 2015); FES hypermethylation and low protein expression were correlated with the PFS (progression-free survival) and OS in HCC (Zhang et al., 2019). As for FITM1, previous study showed that knocking out the FITM1, the lipid droplet accumulation reduces, suggesting that the expression of FITM1 has a connection with lipid droplet, which has a great impact on inflammation, metabolic disorders, and cell injury in liver (Goh and Silver, 2013). Moreover, FITM1 is a member of evolutionarily conserved gene family found in 2008, which plays an important role in fat storage (Kadereit et al., 2008). It closely relates to PPARα in an organ specific way and commonly express at a low level in liver compared with other organs like heart and skeletal muscle in mammals (Rodriguez and Kersten, 2017). Given that FITM1 belongs to a protein family with unique structure and involves in the key progress of lipid metabolism, the aberrant methylation state of FITM1 might result in disorder lipid homeostasis and NAFLD (Gross et al., 2011; Goh and Silver, 2013), triggering the carcinogenesis and progression of the non-viral HCC.

In our study, we revealed that FITM1 expression was much lower in tumor tissues compared with other seven key genes or corresponding normal samples. We also suggested that hypermethylation of FITM1 might account for the downregulation of FITM1 expression partly through modulating NOTCH signaling pathway. More interestingly, as displayed in **Figure S2A**, we found that FITM1 was also down-regulated in the whole HCC patients in TCGA with or without viral infection according to the GEPIA database (http://gepia.cancer-pku.cn) (Tang et al., 2017). And the expression of FITM1 was negatively correlated with the TNM stage (**Figure S2B**). Moreover, low expression of FITM1 predicted worse prognosis in HCC patients (**Figures S2C, D**). However, the specific molecular function of these key genes in HCC, principally FITM1, is still ambiguous. Though the exploration of FITM1 in silico strongly indicated that FITM1 hypermethylation participated in the progression of non-viral HCC by silencing FITM1 expression and it could act as a tumor suppressor gene, related experiments of FITM1 expression, and FITM1 methylation in non-viral HCC still need to be conducted *in vitro* and *in vivo* in the continued study.

While the efficacy of any single biomarker is inadequate, a multiple-risk signature might exert much greater prognostic value for non-viral HCC patients. Therefore, a FITM1-related signature was established in training set through Uni-variate, LASSO, and Multi-variate cox regression analyses and the validation was performed by survival curve and ROC curve analyses in training set and testing set. To make it suitable for the clinical context, we then constructed a nomogram to judge the prognosis of non-viral HCC patients directly and visually. The risk signature and nomogram could enable doctors to identify high and low risk non-viral HCC patients, delivering helpful evidences to make better individualized treatment.

CONCLUSION

In present research, we characterize FITM1 as both a methylation-driven gene and tumor suppressor gene. Based on the investigation of 101 non-viral HCC patients in TCGA, we demonstrate that the hypermethylated FITM1 down-regulates the corresponding FITM1 expression, thereby promoting the progression of non-viral HCC *via* cancer-related pathways. On the other hand, the results also suggest that the signature composed of three methylation-driven genes can function as the prognostic indicator for non-viral HCC patients. In all, not only are potential targets and epigenetic biomarkers discovered and illustrated in our work, a FITM1-related risk signature for non-viral HCC patients is built.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the TCGA database (http://www.tcga.org).

ETHICS STATEMENT

The mRNA expression and methylation data of non-viral HCC provided by TCGA is public and the approval of a local ethics committee is unneeded.

AUTHOR CONTRIBUTIONS

YC and TC designed the research. JC and XicW performed the specific data-processing procedures and wrote the manuscript. XinW and WL made the pictures and graphs. CS revised the final manuscript. All the authors have read and approved the manuscript. The authors declare no conflicts of interest.

FUNDING

This work was supported by the Science and Technology Program of Guangzhou, China (STPG; 2016201604030054) and the National Natural Science Foundation of China (Nos. 81800560).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00099/full#supplementary-material

TABLE S1 | Clinicopathological characteristics of 101 non-viral HCC patients obtained from TCGA.

TABLE S2 | Detailed information of 137 MDGs figured out by MethylMix algorithm.

TABLE S3 | Gene Ontology (GO) analysis results of 137 DEGs.

TABLE S4 | KEGG analysis results of 137 DEGs.

TABLE S5 | Disease Ontology (DO) analysis results of 137 DEGs.

TABLE S6 | GSEA results based on the mRNA level of FITM1.

TABLE S7 | GSEA results based on the methylation level of FITM1.

TABLE S8 | GSEA analysis of Clusters classified by ConsensusClusterPlus analysis.

FIGURE S1 | The additional results of ConsensusClusterPlus analysis and PCA analysis.

FIGURE S2 | The analysis of FITM1 in all HCC patients in TCGA based on the GEPIA database (http://gepia.cancer-pku.cn).

REFERENCES

Alzahrani, B., Iseli, T. J., and Hebbard, L. W. (2014). Non-viral causes of liver cancer: does obesity led inflammation play a role? *Cancer Lett.* 345, 223–229. doi: 10.1016/j.canlet.2013.08.036

Aubry, M.-C., Roden, A., Murphy, S. J., Vasmatzis, G., Johnson, S. H., Harris, F. R., et al. (2015). Chromosomal rearrangements and copy number abnormalities of TP63 correlate with p63 protein expression in lung adenocarcinoma. *Mod. Pathol* 28, 359–366. doi: 10.1038/modpathol.2014.118

Bøvelstad, H., Nygård, S., Størvold, H., Aldrin, M., Borgan, Ø., Frigessi, A., et al. (2007). Predicting survival from microarray data–a comparative study. *Bioinformatics* 23, 2080–2087. doi: 10.1093/bioinformatics/btm305

Braillon, A. (2012). Hepatocellular carcinoma. *Lancet (London England)* 380, 469; author reply 470–461. doi: 10.1016/s0140-6736(12)61282-3

Bruix, J., Gores, G., and Mazzaferro, V. (2014). Hepatocellular carcinoma: clinical frontiers and perspectives. *Gut* 63, 844–855. doi: 10.1136/gutjnl-2013-306627

Cedoz, P., Prunello, M., Brennan, K., and Gevaert, O.. (2018). MethylMix 2.0: an R package for identifying DNA methylation genes. *Bioinformatics* 34, 3044–3046. doi: 10.1093/bioinformatics/bty156

Chen, C., Sheng, S., Hsieh, L., Tang, R., Hsiung, C., Tsai, H., et al. (2016). DNA methylation identifies loci distinguishing hereditary nonpolyposis colorectal cancer without germ-line MLH1/MSH2 mutation from sporadic colorectal cancer. *Clin. Trans. Gastroenterol.* 7, e208. doi: 10.1038/ctg.2016.59

Chuang, T., and Chiang, T. (2014). Impacts of pretranscriptional DNA methylation, transcriptional transcription factor, and posttranscriptional microRNA regulations on protein evolutionary rate. *Genome Biol. Evol.* 6, 1530–1541. doi: 10.1093/gbe/evu124

Cingolani, F., and Czaja, M. (2016). Regulation and Functions of Autophagic Lipolysis. *Trends Endocrinol. Metab.* 27, 696–705. doi: 10.1016/j.tem.2016.06.003

Cuello, M. A., Kato, S., and Liberona, F. (2017). The impact on high-grade serous ovarian cancer of obesity and lipid metabolism-related gene expression patterns: the underestimated driving force affecting prognosis. *J. Cell. Mol. Med.* 22, 1805–1815. doi: 10.1111/jcmm.13463

Devaney, J., Wang, S., Furbert-Harris, P., Apprey, V., Ittmann, M., Wang, B. D., et al. (2015). Genome-wide differentially methylated genes in prostate cancer tissues from African-American and Caucasian men. *Epigenetics* 10, 319–328. doi: 10.1080/15592294.2015.1022019

Dhanasekaran, R., Bandoh, S., and Roberts, L. R. (2016). Molecular pathogenesis of hepatocellular carcinoma and impact of therapeutic advances. *F1000Res* 5. doi: 10.12688/f1000research.6946.1

Gao, C., Zhuang, J., Li, H., Liu, C., Zhou, C., Liu, L., et al. (2018). Exploration of methylation-driven genes for monitoring and prognosis of patients with lung adenocarcinoma. *Cancer Cell Int.* 18, 194. doi: 10.1186/s12935-018-0691-z

Gevaert, O., Tibshirani, R., and Plevritis, S. (2015). Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome Biol. Evol.* 16, 17. doi: 10.1186/s13059-014-0579-8

Gevaert, O. (2015). MethylMix: an R package for identifying DNA methylation-driven genes. *Bioinformatics* 31, 1839–1841. doi: 10.1093/bioinformatics/btv020

Goh, V., and Silver, D. (2013). The lipid droplet as a potential therapeutic target in NAFLD. *Semin. Liver Dis.* 33, 312–320. doi: 10.1055/s-0033-1358521

Gross, D., Zhan, C., and Silver, D. (2011). Direct binding of triglyceride to fat storage-inducing transmembrane proteins 1 and 2 is important for lipid droplet formation. *Proc. Natl. Acad. Sci. U. S. A.* 108, 19581–19586. doi: 10.1073/pnas.1110817108

Hlady, R., Sathyanarayan, A., Thompson, J., Zhou, D., Wu, Q., Pham, K., et al. (2019). Integrating the Epigenome to Identify Drivers of Hepatocellular Carcinoma. *Hepatology* 69, 639–652. doi: 10.1002/hep.30211

Huang, S. Z., Wei, M. N., Huang, J. R., Zhang, Z. J., Zhang, W. J., Jiang, Q. W., et al. (2019). Targeting TF-AKT/ERK-EGFR Pathway Suppresses the Growth of Hepatocellular Carcinoma. *Front. Oncol.* 9, 150. doi: 10.3389/fonc.2019.00150

Kadereit, B., Kumar, P., Wang, W., Miranda, D., Snapp, E. L., Severina, N., et al. (2008). Evolutionarily conserved gene family important for fat storage. *Proc. Natl. Acad. Sci. U. S. A.* 105, 94–99. doi: 10.1073/pnas.0708579105

Kerr, I., Haider, A., and Gelissen, I. (2011). The ABCG family of membrane-associated transporters: you don't have to be big to be mighty. *Br. J. Pharmacol.* 164, 1767–1779. doi: 10.1111/j.1476-5381.2010.01177.x

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28 (5), 1–26.

Lee, W., Bachtiar, M., Choo, C., and Lee, C. (2019). Comprehensive review of Hepatitis B Virus-associated hepatocellular carcinoma research through text mining and big data analytics. *Biol. Rev. Camb. Philos. Soc.* 94, 353–367. doi: 10.1111/brv.12457

Li, G., Ding, Z., Wang, Y., Liu, T., Chen, W., Wu, J., et al. (2019). Integrative analysis of DNA methylation and gene expression identify a six epigenetic driver signature for predicting prognosis in hepatocellular carcinoma. *J. Cell. Physiol.* 234, 11942–11950. doi: 10.1002/jcp.27882

Liu, S., Miao, C., Liu, J., Wang, C., and Lu, X. (2018). Four differentially methylated gene pairs to predict the prognosis for early stage hepatocellular carcinoma patients. *J. Cell. Physiol.* 233, 6583–6590. doi: 10.1002/jcp.26256

Lossos, I. S., Czerwinski, D. K., Alizadeh, A. A., Wechser, M. A., Tibshirani, R., Botstein, D., et al. (2004). Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N Engl. J. Med.* 350, 1828–1837. doi: 10.1056/NEJMoa032520

Noguchi, S., Motomura, K., Inaji, H., Imaoka, S., and Koyama, H.. (1994). Clonal analysis of predominantly intraductal carcinoma and precancerous lesions of the breast by means of polymerase chain reaction. *Cancer Res.* 54, 1849–1853. doi: 10.1016/0304-3835(94)90050-7

Ogunwobi, O. O., Harricharran, T., Huaman, J., Galuza, A., Odumuwagun, O., Tan, Y., et al. (2019). Mechanisms of hepatocellular carcinoma progression. *World J. Gastroenterol.* 25, 2279–2293. doi: 10.3748/wjg.v25.i19.2279

Pu, W., Wang, C., Chen, S., Zhao, D., Zhou, Y., Ma, Y., et al. (2017). Targeted bisulfite sequencing identified a panel of DNA methylation-based biomarkers for esophageal squamous cell carcinoma (ESCC). *Clin. Epigenet* 9, 129. doi: 10.1186/s13148-017-0430-7

Qiu, J., Peng, B., Tang, Y., Qian, Y., Guo, P., Li, M., et al. (2017). CpG Methylation Signature Predicts Recurrence in Early-Stage Hepatocellular Carcinoma: Results From a Multicenter Study. *J. Clin. Oncol.* 35, 734–742. doi: 10.1200/JCO.2016.68.2153

Revill, K., Wang, T., Lachenmayer, A., Kojima, K., Hrrington, A., Li, J., et al. (2013). Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. *Gastroenterology* 145, 1424–1435.e1421-1425. doi: 10.1053/j.gastro.2013.08.055

Rodriguez, MADLR, and Kersten, S. (2017). Regulation of Lipid Droplet-Associated Proteins by Peroxisome Proliferator-Activated Receptors. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* 1862. doi: 10.1016/j.bbalip.2017.07.007

Salo, V. T., Belevich, I., Li, S., Karhinen, L., Vihinen, H., Vigouroux, C., et al. (2016). Seipin regulates ER–lipid droplet contacts and cargo delivery. *Embo. J.* 35, 2699–2716. doi: 10.15252/embj.201695170

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer Statistics 2017. *CA Cancer J. Clin.* 60, 277–300. doi: 10.3322/caac.21387

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gellette, 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

Sun, X., Wang, M., Zhang, F., and Kong, X. (2018). An integrated analysis of genome-wide DNA methylation and gene expression data in hepatocellular carcinoma. *J. FEBS Open Bio.* 8, 1093–1103. doi: 10.1002/2211-5463.12433

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, W98–W102. doi: 10.1093/nar/gkx247

Tian, Y., Yang, B., Qiu, W., Hao, Y., Zhang, Z., Yang, B., et al. (2019). ER-residential Nogo-B accelerates NAFLD-associated HCC mediated by metabolic reprogramming of oxLDL lipophagy. *Nat. Commun.* 10, 3391. doi: 10.1038/s41467-019-11274-x

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419. doi: 10.1126/science.1260419

Wang, Y., Ruan, Z., Yu, S., Tian, T., Liang, X., Jing, L., et al. (2019). A four-methylated mRNA signature-based risk score system predicts survival in patients with hepatocellular carcinoma. *Aging* 11, 160–173. doi: 10.18632/aging.101738

Wang, J., Zhang, Q., Zhu, Q., Liu, C., Nan, X., Wang, F., et al. (2020). Identification of methylation-driven genes related to prognosis in clear-cell renal cell carcinoma. *J. Cell Physiol.* 235, 1296–1308. doi: 10.1002/jcp.29046

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

Xie, X., Wang, X., Liao, W., Fei, R., Wu, N., Cong, X., et al. (2018). MYL6B, a myosin light chain, promotes MDM2-mediated p53 degradation and drives HCC development. *J. Exp. Clin. Cancer Res.* 37. doi: 10.1186/s13046-018-0693-7

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, G., Wang, L. G., Yan, G. R., and He, Q. Y.. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 31, 608–609. doi: 10.1093/bioinformatics/btu684

Zhang, C., Ge, S., Wang, J., Jing, X., Li, H., Mei, S., et al. (2019). Epigenomic profiling of DNA methylation for hepatocellular carcinoma diagnosis and prognosis prediction. *J. Gastroenterol. Hepatol.* 34, 1869–1877. doi: 10.1111/jgh.14694

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 © 2020 Chen, Wang, Wang, Li, Shang, Chen 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 10 March 2020
doi: 10.3389/fbioe.2020.00176

MicroRNAs Associated With Colon Cancer: New Potential Prognostic Markers and Targets for Therapy

Junfeng Zhu[1†], *Ying Xu*[2†], *Shanshan Liu*[1], *Li Qiao*[3], *Jianqiang Sun*[4] *and Qi Zhao*[1,5*]

[1] Department of Clinical Laboratory, Affiliated Hospital of Guilin Medical University, Guilin, China, [2] Office of Drug Clinical Trials, Affiliated Hospital of Guilin Medical University, Guilin, China, [3] Department of Clinical Laboratory, General Hospital of Northern Theater Command, Shenyang, China, [4] School of Automation and Electrical Engineering, Linyi University, Linyi, China, [5] College of Computer Science, Shenyang Aerospace University, Shenyang, China

MicroRNAs (miRNAs) are a kind of non-coding RNA (ncRNA) that regulate the expression of target genes and play a role in the occurrence and development of cancers. Colon cancer (COAD) is the second most common cause of cancer-related mortality. However, the prognostic value of miRNAs in COAD is still confusing. In this study, we obtain miRNAs and messenger RNAs (mRNAs) expression profiles of COAD from the Cancer Genome Atlas (TCGA) database. After preliminary data screening and preprocessing, we acquire the expression data of 894 miRNAs and 17,019 mRNAs. Then, compared with the normal samples, 39 upregulated miRNAs and 54 downregulated miRNAs are identified by differential expression analysis. Furthermore, we obtain 1,487 upregulated mRNAs and 2,847 downregulated mRNAs. We confirm nine key miRNAs related to the survival rate of COAD patients. Moreover, by using bioinformatics methods, we get 461 common genes from both the target genes of these nine key miRNAs and differentially expressed mRNAs. Through analyzing the protein-protein interaction (PPI) network of these 461 common genes and survival analysis, we confirm five hub genes as promising biomarkers for COAD prognosis. It is worth mentioning that no previous reports have found that PGR and KCNB1 are related to COAD. We expect these key miRNAs and hub genes will provide a new way for the study of COAD.

Keywords: miRNA, hub gene, colon cancer, prognostic marker, bioinformatics

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Yuangen Yao,
Huazhong Agricultural University, China

***Correspondence:**
Qi Zhao
zhaoqi@lnu.edu.cn

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *31 December 2019*
Accepted: *20 February 2020*
Published: *10 March 2020*

Citation:
Zhu J, Xu Y, Liu S, Qiao L, Sun J and Zhao Q (2020) MicroRNAs Associated With Colon Cancer: New Potential Prognostic Markers and Targets for Therapy. Front. Bioeng. Biotechnol. 8:176. doi: 10.3389/fbioe.2020.00176

INTRODUCTION

Non-coding RNAs (ncRNAs) are a kind of RNA which cannot be translated into protein. NcRNAs were once considered junk RNA. During the last few decades, with the development of high-throughput sequencing technology, people have realized the critical function of ncRNAs (Bussotti et al., 2013). NcRNAs contain functional types of RNA such as transfer RNAs, ribosomal RNAs, microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and so on. Although most of the biological functions are performed by proteins in organisms, ncRNAs also play important roles in various biological processes (Chen et al., 2017, 2019a; Zhou et al., 2017, 2018b; Ferreira and Esteller, 2018). MiRNAs are endogenous, short ncRNAs which can regulate the expression of more than 30% of human genes (Lewis et al., 2005). Abnormal expression of miRNAs is related to various biological modifications, such as apoptosis, cell differentiation, and carcinogenesis (Su et al., 2017; Truong et al., 2017). More and more studies show that miRNAs play a necessary role in

the occurrence and development of a variety of cancers, including lung cancer, breast cancer, and colon cancer (Seo et al., 2019).

Colon cancer (COAD) is a common malignant tumor that happens at the junction of rectum and sigmoid colon. In 2018, COAD ranked second for mortality and fourth for incidence (Bray et al., 2018). The 5-year survival rate of colon cancer patients is ∼65% (Siegel et al., 2017). Until now, the histological feature is the only prognostic indicator for COAD. According to the histological feature, it is hard to decide to receive adjuvant chemotherapy after surgery or not. Therefore, it is imperative to excavate biomarkers for prognosis of COAD. In 2016, Caritg et al. confirmed three miRNAs (miR-103a-3p, miR-143-5p, and miR-215) as prognostic markers of COAD in patients (Caritg et al., 2016). Then Bobowicz et al. reported that five miRNAs (miR-1296, miR-135b, miR-539, miR-572, and miR-185) have prognostic values for colon cancer in patients (Bobowicz et al., 2016). Later, Maierthaler et al. showed that the miR-122 and miR-200 families have prognostic value in COAD (Maierthaler et al., 2017). However, experimentally determining the prognostic value of miRNAs is not only time-consuming but also costly. So far, many computational methods have been proposed in the field of bioinformatics, for example, circRNA-disease association (Ge et al., 2019; Zhao et al., 2019), drug-side effect (Ding et al., 2019), lncRNA-miRNA interaction (Liu et al., 2020), and lncRNA-protein interaction predictions (Hu et al., 2018; Zhao et al., 2018a,b; Bao et al., 2019; Shen et al., 2019). Therefore, there is an urgent need for developing effective bioinformatics analysis to identify molecular mechanisms related to COAD from the accumulated clinical and experimental data.

TCGA is a large database which contains epigenomic and standardized clinical data from massive samples of each kind of cancer. The data of TCGA includes exon expression data and miRNAs expression data, copy number segments, DNA methylation, phenotype, and so on. Gene chips laid on a high-throughput test can detect thousands of gene expressions in one experiment. Therefore, TCGA database can help us to obtain a large amount of genetic information about cancer in a short time and provide new targets for diagnosis and treatment of cancer (Rajendran et al., 2012, 2018; Rajendran, 2016). Several lines of evidence suggest that some modifications in miRNAs, such as ectopia, mutation, and overexpression, can cause severe pathological alterations (Xiao and Rajewsky, 2009). These effects are attributable to the translation of messenger RNA (mRNA) into protein regulated by miRNA (Bartel, 2018). Besides, different kinds of miRNA present particular expression levels in specific sorts of cancers (Hou et al., 2017). Recently, Xu et al. demonstrated that four miRNAs were significantly associated with the overall survival of COAD patients (Xu et al., 2016). Nevertheless, this study only analyzed the relationship between the expression of miRNAs and COAD patient survival.

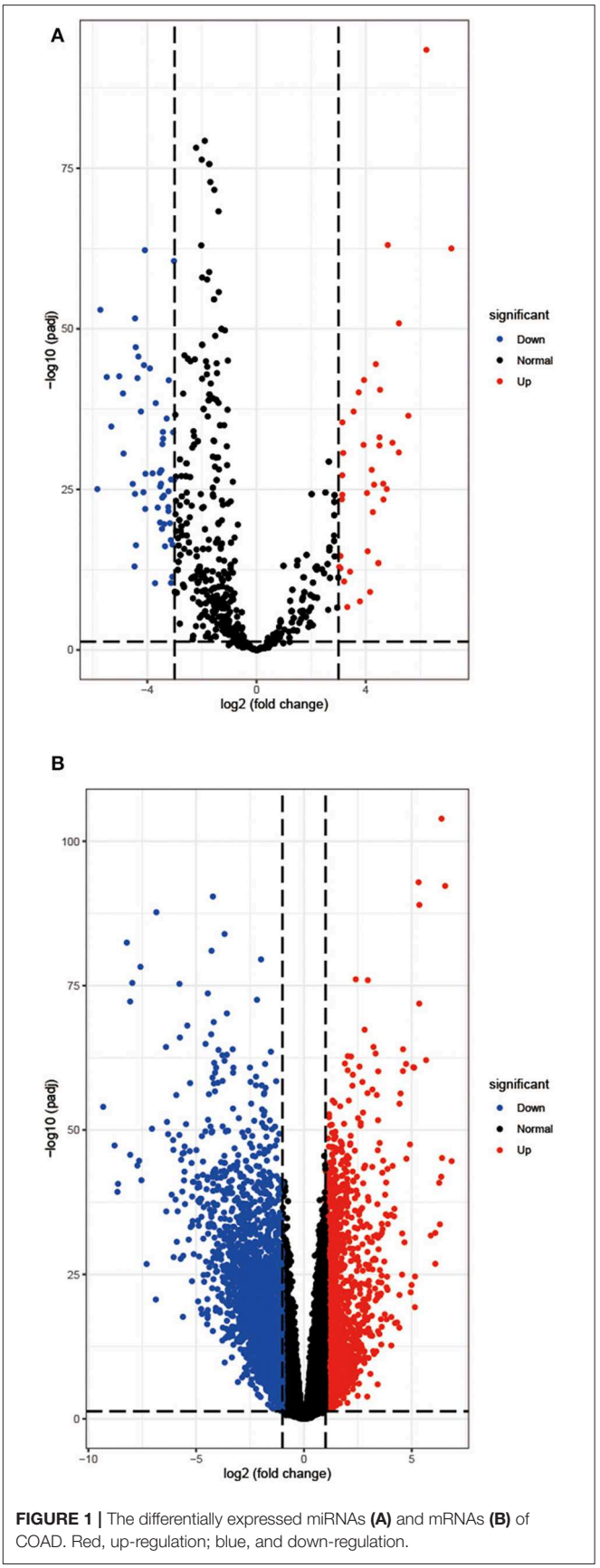


FIGURE 1 | The differentially expressed miRNAs **(A)** and mRNAs **(B)** of COAD. Red, up-regulation; blue, and down-regulation.

Abbreviations: miRNAs, microRNAs; COAD, colon cancer; TCGA, the Cancer Genome Atlas; PPI, protein-protein interaction; ncRNAs, Non-coding RNAs; HMGB1, High mobility group box protein 1; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; CAMs, cell adhesion molecules; CCL-1, cell adhesion molecule 1; CC, cellular component; BP, biological process; MF, molecular function; MAPK, Mitogen-activated protein kinase; HIF-1, hypoxia-inducible factor-1.

It did not explore the role of the target genes of these miRNAs in COAD. Hence, the mRNA and miRNA data of COAD from TCGA should be considered together to detect new biomarkers.

In this study, we download miRNA and mRNA expression data of COAD from TCGA database. We do a series of analyses of miRNA and mRNA, such as expression difference analysis, Gene Ontology (GO) analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis, survival analysis, and PPI network analysis. Then nine key miRNAs (miR-217, miR-144, miR-129, miR-125a, miR-125b, miR-375, miR-328, miR-486, and miR-194) and five hub genes (PPARGC1A, COL1A1, SYT1, PGR, and KCNB1) are confirmed to have prognostic value in COAD. The novel promising prognostic miRNA and mRNA identified in our study will provide

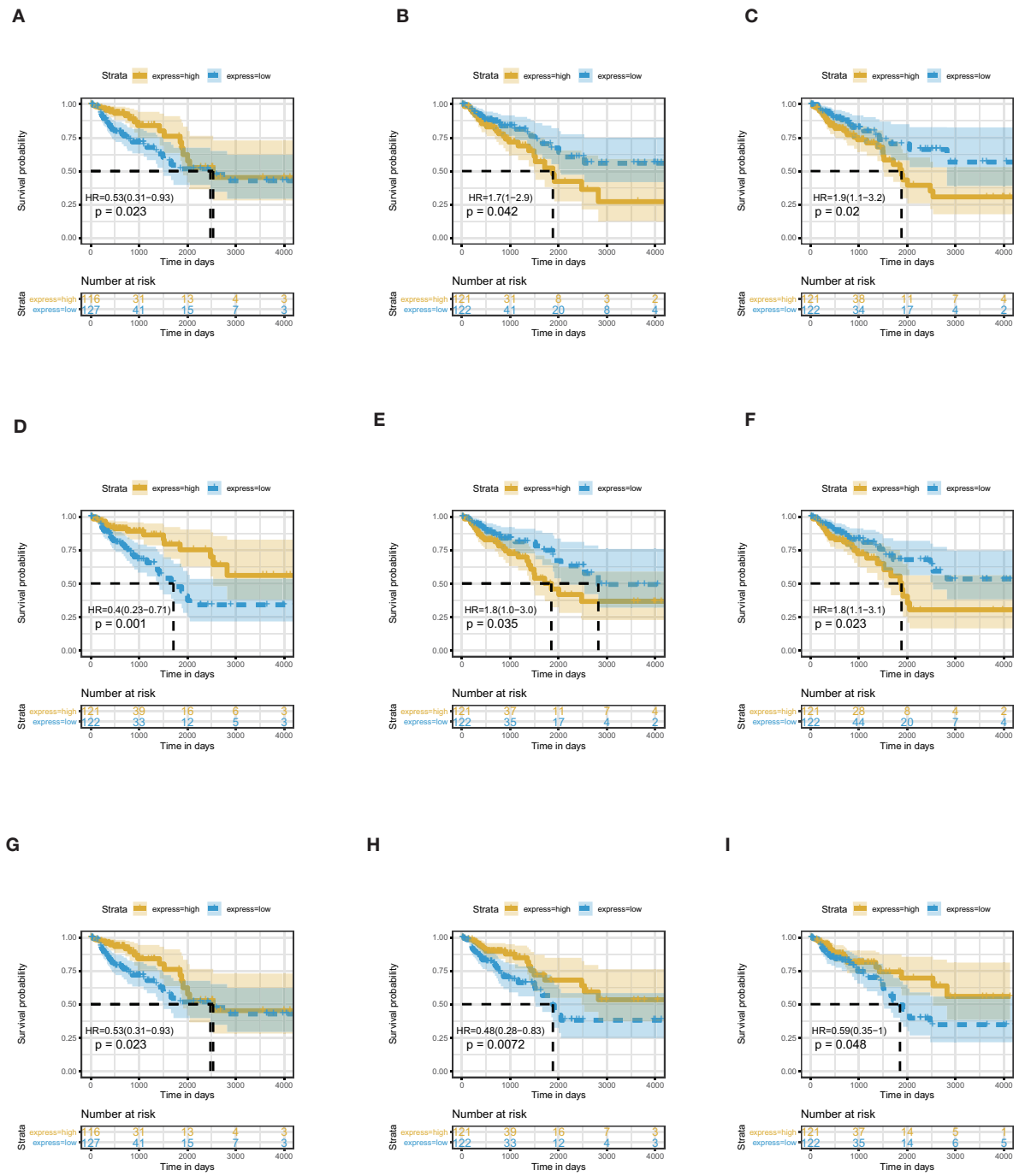


FIGURE 2 | Nine differentially expressed miRNAs are associated with overall survival in COAD patients by using Kaplan–Meier curve, Log-rank test. **(A)** miR-129; **(B)** miR-217; **(C)** miR-125a; **(D)** miR-375; **(E)** miR-328; **(F)** miR-125b; **(G)** miR-144; **(H)** miR-194; **(I)** miR-486.

a new approach for clinical and experimental research in COAD.

MATERIALS AND METHODS

Data Set, Differential Expression Analysis, and Survival Analysis

The "miRNA mature strand expression RNAseq by IlluminaHiseq," "gene expression RNAseq," and "phenotype" dataset of COAD are downloaded from TCGA. The "miRNA mature strand expression RNAseq by IlluminaHiseq" dataset contains the miRNA expression data from 261 samples, including 8 normal samples and 253 COAD samples. Then the "gene expression RNAseq" dataset consists of mRNA expression data from 329 samples, including 43 normal samples and 286 COAD samples. Furthermore, the "phenotype" dataset contains the clinical-pathological data of 545 samples, including 85 normal samples and 460 COAD samples.

We analyze the downloaded dataset as follows: above all, we separate COAD tissue and adjacent non-tumor colon tissue according to the sample number. Then we remove the data of miRNAs and mRNAs with reported expression data for <50% of the patients. Next, a CancerSubtypes package is employed to analyze the expression data of mRNA from the "gene expression RNAseq" dataset. The differentially expressed mRNAs are identified by using the thresholds which are |log2fold-change (FC)| > 1.0 and adjusted $p < 0.05$ for COAD samples compared with the normal samples. Then we identify the differentially expressed miRNA by analyzing the expression data of miRNAs from the "miRNA mature strand expression RNAseq by IlluminaHiseq" dataset in the similar way but using a different threshold which are |log2FC| > 3.0 and adjusted $p < 0.05$. In addition, a volcano map is drawn by ggplot2 package.

We use the Cox regression analysis to investigate the relationship between each miRNA/mRNA expression level and the overall survival rate of COAD patients in the "phenotype" dataset. Log-rank $P < 0.05$ is considered statistically significant for survival differences. Moreover, Kaplan–Meier curves of nine key miRNAs and five hub genes are drawn by the survminer package.

Prediction of Target Genes of miRNAs and Functional Enrichment Analysis

The target genes of nine key miRNAs are predicted by three kinds of online analysis software including miRDB (http://www.mirdb.org/miRDB/), TargetScanHuman (version 7.2, http://www.targetscan.org/), and mirDIP (http://ophid.utoronto.ca/mirDIP/). Then the Venn diagram is applied to confirm the common genes both in the target genes of miRNA and differentially expressed mRNA. To further understand the biological functions of the common genes, we perform GO and KEGG pathway enrichment analyses by using KOBAS (version 3.0; https://kobas.cbi.pku.edu.cn/anno_iden.php) online tool. $P < 0.05$ is regarded as statistically significant.

PPI Network Analysis

The STRING (version 11.0, http://string-db.org) is used for searching PPI of the common genes. At the start, a Venn diagram is used to identify the common genes both in the target genes of the nine key miRNAs and the differential expression mRNAs. After importing the official gene symbols of the common genes into STRING, we get the PPI network of the common genes. Then, Cytoscape (version 3.7.1) is applied for the visualization of PPI networks. The confidence score 0.4 is used as the cut-off criterion.

Confirmation of Hub Genes

CytoHubba, an app of Cytoscape, is applied to confirm hub genes. We employ a Venn diagram to extract the overlapping genes of the top 50 genes by six different algorithms, including MCC, Degree, Closeness, Radiality, Betweenness, and Stress. These overlapping genes are confirmed as the hub genes. Subsequently, we utilize Cox regression analysis to determine the prognostic role of the hub genes.

RESULT

Identification of Differentially Expressed miRNAs and mRNAs in COAD

Based on the analysis of the CancerSubtypes package, 93 differentially expressed miRNAs are acquired, including 39 upregulated miRNAs and 54 downregulated miRNAs (**Figure 1A**). Then in a similar way, 4,334 differentially expressed mRNAs containing 1,487 upregulated miRNAs and 2,847 downregulated mRNAs are extracted (**Figure 1B**).

Identification of miRNA With Prognostic Value in COAD

Through survival analysis, we find nine miRNAs that are significantly associated with the overall survival of COAD patients (**Figure 2**). The name, Log2FC, *p*-value and adjusted *p*-value of these key miRNAs are displayed in **Table 1**. In these miRNAs, miR-217 and miR-144 are upregulated, miR-129, miR-125a, miR-125b, miR-375, miR-328, miR-486, and miR-194 are downregulated. In COAD, miR-217 specifically inhibits

TABLE 1 | The name, Log2FC, *P*-value, and Adjusted *p*-value of nine key miRNAs in COAD.

MiRNA name	Log2FC	*p*-value	Adjusted *p*-value
miR-217	3.21	1.10E-11	2.15E-11
miR-144	4.06	1.69E-16	4.25E-16
miR-129	−4.87	4.62E-32	2.61E-31
miR-125b	−3.08	1.94E-12	4.01E-12
miR-125a	−4.36	4.02E-44	4.65E-43
miR-375	−3.71	2.11E-11	4.04E-11
miR-328	−5.72	4.17E-55	1.09E-53
miR-486	−5.83	2.16E-26	8.98E-26
miR-194	−4.44	1.26E-25	4.99E-25

DKK1, which is an important antagonist of the Wnt signaling pathway to promote apoptosis of colon cells (Jia et al., 2019). By controlling the expression of SMAD4, miR-144 inhibits invasion and migration of colon cancer cells (Sheng et al., 2019). High mobility group box protein 1 (HMGB1) plays a part in immune escape in COAD cells (Zheng and Zhu, 2018). MiR-129, which targets the 3′UTR of HMGB1, is able to repress the development of COAD (Wu et al., 2018). By inhibiting cell proliferation and inducing cell apoptosis, miR-125a acts as a suppressor of COAD (Tong et al., 2015). The level of miR-375 is lower in colorectal cancer tissues than normal human colon tissues. In addition, miR-375 exerts an inhibitory effect on the proliferation of colorectal cancer by targeting the 3′UTR of KLF4 (Mao et al., 2016). Recently, a new study reported that miR-375 has a prognostic value in COAD (Huang and Pan, 2019). By regulating SLC2A1/GLUT1, miR-328 participates in the Warburg effect in COAD (Santasusagna et al., 2018). MiR-486 is related to the molecular mechanisms of several cancers, including cervical cancer (Li et al., 2018a), breast cancer (Li et al., 2019), lung cancer (Tian et al., 2019), esophageal cancer (Lang and Zhao, 2018), ovarian cancer (Ma et al., 2016), and pancreatic cancer (Xia et al., 2019). Kelley et al. nearly demonstrated that the expression of miR-486 associates with early-stage of COAD (Kelley et al., 2018). Moreover, Ren et al. reported that miR-486 plays a prognostic role in COAD (Ren et al., 2016). MiR-194, whose expression is upregulated by p53, inhibits THBS1 expression to promote angiogenesis and facilitate tissue repair in COAD (Sundaram et al., 2011). In a previous study, miR-125b is confirmed as a prognostic biomarker of colorectal cancer (Zhou et al., 2018c). Up to now, the prognostic value of four miRNAs (miR-217, miR-125, miR-129, and miR-194) in COAD has not been previously reported. These miRNAs will become new potential prognostic biomarkers of COAD.

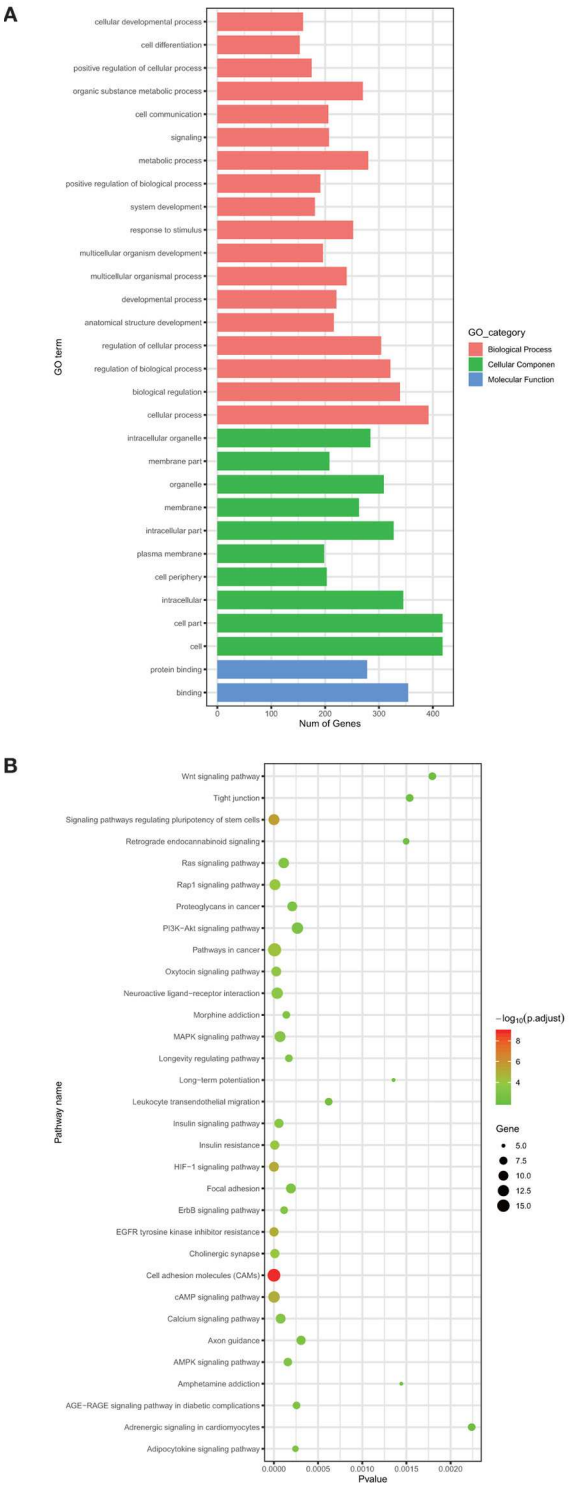


FIGURE 3 | The GO analysis **(A)** and KEGG pathway analysis **(B)** of common genes.

Prediction of miRNA-mRNA Interaction and Functional Enrichment Analysis

The target gene predictive tools, including miRDB, mirDIP, and TargetScanHuman, are employed to find miRNAs-targeted genes. During this process, 7,592 genes are identified as the target genes of nine miRNAs with prognostic value in COAD. To improve the accuracy of subsequent analysis, a Venn diagram is used to find common genes both in the differentially expressed mRNAs and in the 7,592 target genes. At last, we get 461 common genes. To further understand the biological roles of these 461 common genes, we perform GO analysis and KEGG pathways enrichment analysis of them. The GO analysis shows that 418 genes are involved in cell and cell parts in the cellular component (CC) category. Regarding the biological process (BP) category, this result shows that most of the 461 common genes enrich several functions, including cellular process (392 genes), biological regulation (339 genes), regulation of biological process (321 genes), and regulation of cellular process (304 genes). Then for the molecular function (MF) group, a large proportion of the common genes are mainly enriched in binding and protein binding (**Figure 3A**). In addition, the KEGG Pathway analysis demonstrates that the 461 common genes enrich cancer-associated signaling pathways, containing pathways in cancer (17 genes, *p*-Value: 7.41E-06), cell adhesion molecules (CAMs) (16

genes, *p*-Value: 7.43E-11), cAMP signaling pathway (13 genes, *p*-Value: 1.22E-06), signaling pathways regulating the pluripotency of stem cells (12 genes, *p*-Value: 2.55E-06), MAPK signaling pathway (12 genes, *p*-Value: 6.88E-05), HIF-1 signaling pathway (10 genes, *p*-Value: 7.70E-07) and proteoglycans in cancer (10 genes, *p*-Value: 2.08E-04) (**Figure 3B**). Pathways in cancer include Wnt signaling pathway, P53 signaling pathway, and so on. These signaling pathways involve the development of COAD. A previous study showed that CAMs can regulate cancer cell invasion and tumor metastasis (Geletu et al., 2018). The cAMP signaling pathway takes part in regulating various biological processes including cell proliferation, secretion, metabolism and apoptosis. Targeted modulation of cAMP signaling pathway can induce proliferation and apoptosis of a variety of malignant lymphoma cells (Mehta and Patel, 2019).

Confirmation of Hub Genes

The 461 common genes are used to construct a PPI network which contains 351 nodes and 881 edges (**Figure 4**). We utilize the cytoHubba app of Cytoscape software to calculate the connectivity between the genes. Because the genes with high connectivity can play an important role in the cancers, we

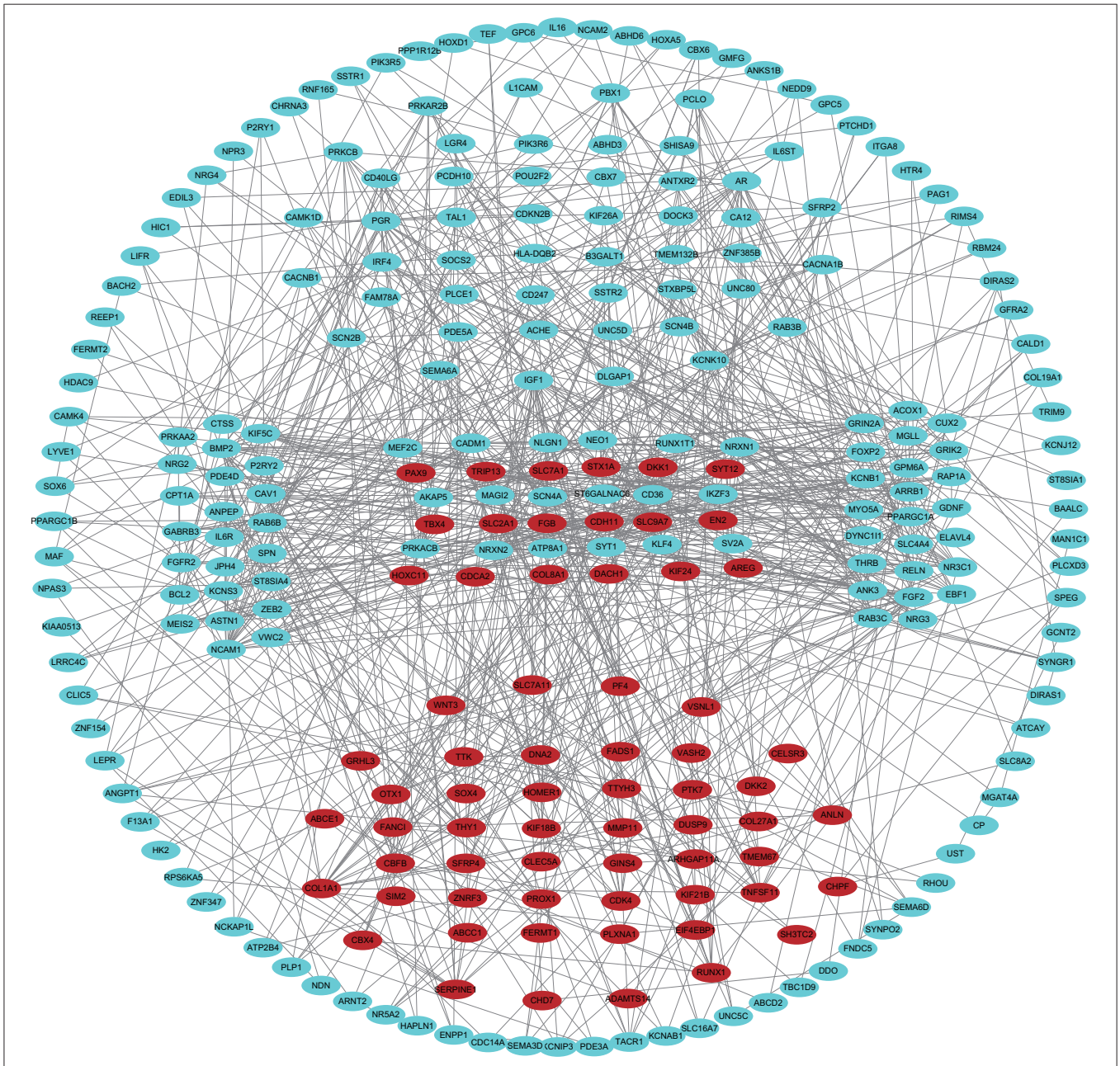


FIGURE 4 | PPI network of common genes. The blue and red dots represent down-regulated and up-regulated genes, respectively.

identify five hub genes. The gene symbol, Log2FC, *p*-value, and adjusted *p*-value of these hub genes are showed in **Table 2**. Furthermore, through survival analysis, we find the five hub genes that have prognostic value in COAD (**Figure 5**). Among them, when the level of PPARGC1A expression rises, the survival rate of COAD patients improves. Conversely, when the expression levels of COL1A1, SYT1, PGR, and KCNB1 decline, the survival rate of COAD patients improves. As an activator of p53, PPARGC1A can suppress cancer cell apoptosis (Sen et al., 2011). COL1A1 participates in the process of focal adhesion and may influence the metastatic ability of cells (Tian et al., 2015). COL1A1, which is upregulated in COAD, may be a biomarker for colon cancer progression (Yang et al., 2019). Because of down regulation in left-sided colon carcinoma compared with right-sided colon carcinoma, STY1 possibly plays a role in their genetic susceptibilities to neoplastic transformation (Zhu et al., 2013). As the receptor of hormones, PGR is very important for breast growth. Moreover, it is related to the development of breast cancer and endometrial cancer (Kurozumi et al., 2017; He et al., 2019). As a prognostic biomarker for gliomas, KCNB1 suppresses tumor growth by inducing autophagy (Wang et al., 2017). So far, there is no study showing the effect of PGR and KCNB1 in COAD. Therefore, PGR and KCNB1 may be new potential prognostic biomarkers in COAD.

TABLE 2 | The Gene symbol, Log2FC, *P*-value, and Adjusted *p*-value of five hub genes in COAD.

Gene symbol	Log2FC	*P*-value	Adjusted *p*-value
COL1A1	2.21	6.17E-15	2.71E-14
SYT1	−1.06	1.64E-03	2.55E-03
PGR	−2.56	6.35E-24	5.93E-23
KCNB1	−2.56	1.47E-29	2.21E-28
PPARGC1A	−2.11	4.53E-16	2.18E-15

DISCUSSION

COAD is a common gastrointestinal tumor. High mortality and morbidity are the main clinical features of this disease. At present, the diagnosis of COAD has many problems, such as poor specificity and low time efficiency (Benson et al., 2017; Sun et al., 2019). Previous studies show that some genetic changes are closely associated with the occurrence of COAD

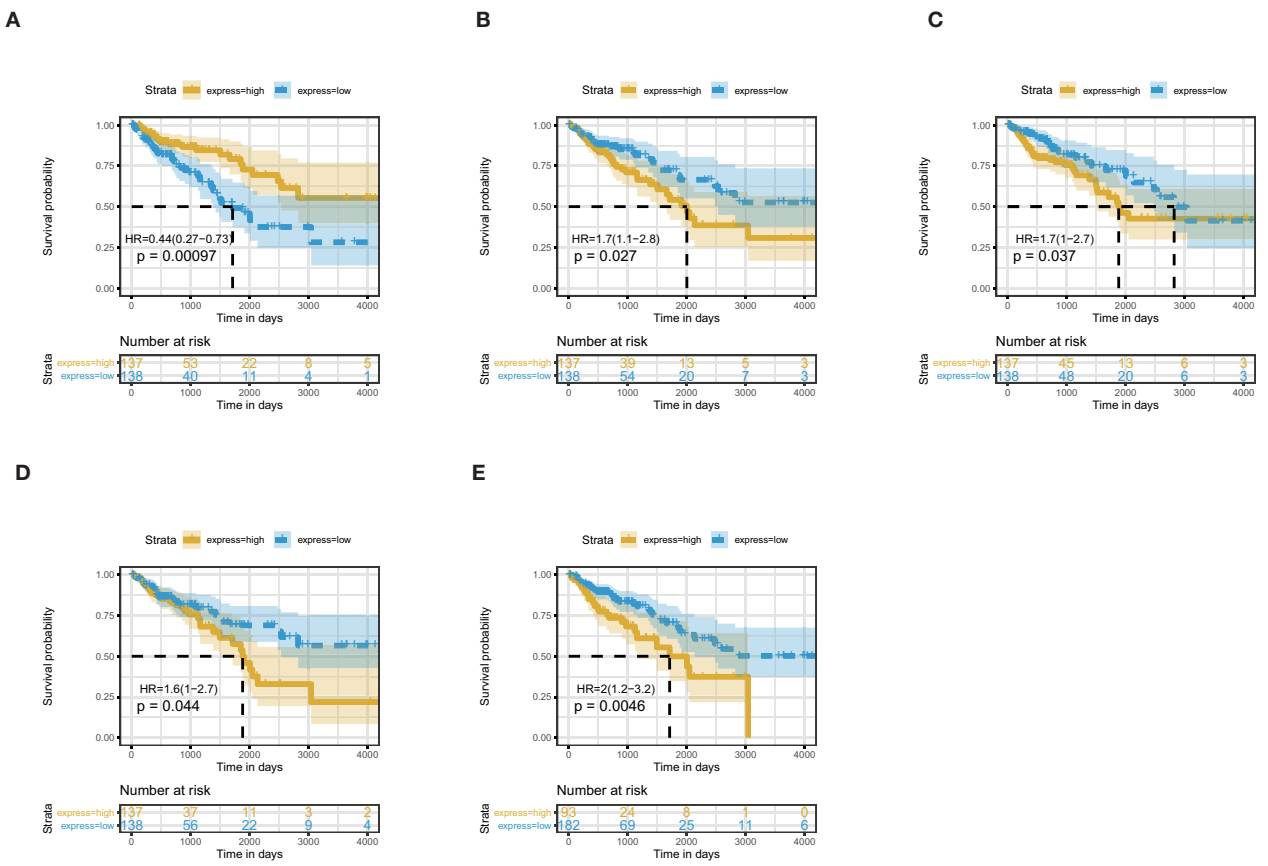


FIGURE 5 | Five hub genes are associated with overall survival in COAD patients by using Kaplan–Meier curve, Log-rank test. **(A)** PPARGC1A; **(B)** COL1A1; **(C)** SYT1; **(D)** PGR; **(E)** KCNB1.

(Li et al., 2018b; Wei et al., 2018). For this reason, it is critical to find prognostic biomarkers for the treatment of COAD. It is well-known that the experimental research on the role of miRNAs needs many materials and costs huge time. Compared with traditional experimental methods, the bioinformatics method is cost-effective and timesaving when studying the role of miRNAs. Therefore, in this study we research the prognostic role of miRNAs in COAD using the bioinformatics methods. Furthermore, the role of target genes in COAD can further prove the accuracy of miRNAs predictive results. Hence, we assess not only four miRNAs but also two target genes of nine miRNAs with prognostic value as new potential prognostic markers for COAD. In addition, the target genes with prognostic value can also be therapeutic targets for COAD. These findings will provide a basis for further research on diagnoses of COAD, as well as a new target of monoclonal antibody drugs for treatment of COAD. In future, we plan to use several previous effective computational models to identify colon cancer-related miRNAs (Chen et al., 2018a,b, 2019b).

Despite some promising results, there are several limitations in the current study. First, all data analyzed in the study come from TCGA database, and there is no actual clinical patient data involved in the study. Next, we confirm several novel key miRNAs and hub genes that have never been reported to own prognostic value in COAD. But their molecular mechanisms should be further explored by the experimental way. In the end, more and more studies reported on the role of lncRNA in COAD (Wang et al., 2018; Zhou et al., 2018a). However, lncRNAs are not involved in this study. In the following study, we will collect enough clinical data for further verifying the results of the study. Then we will design a series of experiments to investigate the molecular mechanisms of novel key miRNAs and hub genes. Moreover, lncRNAs should be introduced into the prognostic prediction system for improving accuracy.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

AUTHOR CONTRIBUTIONS

QZ contributed to the design of the study protocol. JZ and YX performed statistical analysis and drawn the pictures. LQ and JS contributed to the writing of the study protocol. SL downloaded data. All authors approved the final version of the manuscript.

FUNDING

This work was supported by Doctoral Scientific Research Starting Foundation of Guilin Medical University, the Education Department of Liaoning Province under Grant No. LQN201711 and the National Natural Science Foundation of China under Grant No. 11805091.

REFERENCES

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief. Bioinform*. bbz118. doi: 10.1093/bib/bbz118

Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell* 173, 20–51. doi: 10.1016/j.cell.2018.03.006

Benson, A. B., Venook, A. P., Cederquist, L., Chan, E., Chen, Y.-J., Cooper, H. S., et al. (2017). Colon cancer, version 1.2017, NCCN clinical practice guidelines in oncology. *J. Natl. Compr. Cancer Netw.* 15, 370–398. doi: 10.6004/jnccn.2017.0036

Bobowicz, M., Skrzypski, M., Czapiewski, P., Marczyk, M., Maciejewska, A., Jankowski, M., et al. (2016). Prognostic value of 5-microRNA based signature in T2-T3N0 colon cancer. *Clin. Exp. Metas.* 33, 765–773. doi: 10.1007/s10585-016-9810-1

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

Bussotti, G., Notredame, C., and Enright, J. A. (2013). Detecting and comparing non-coding RNAs in the high-throughput era. *Int. J. Mol. Sci.* 14, 15423–15458. doi: 10.3390/ijms140815423

Caritg, O., Navarro, A., Moreno, I., Martínez-Rodenas, F., Cordeiro, A., Muñoz, C., et al. (2016). Identifying high-risk stage II colon cancer patients: a three-MicroRNA-based score as a prognostic biomarker. *Clin. Colorect. Cancer* 15, e175–e182. doi: 10.1016/j.clcc.2016.04.008

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018a). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Xie, D., Zhao, Q., and You, Z.-H. (2019a). MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 20, 515–539. doi: 10.1093/bib/bbx130

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2017). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., Yin, J., Qu, J., and Huang, L. (2018b). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, X., Zhu, C.-C., and Yin, J. (2019b). Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput. Biol.* 15:e1007209. doi: 10.1371/journal.pcbi.1007209

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

Ferreira, H. J., and Esteller, M. (2018). Non-coding RNAs, epigenetics, and cancer: tying it all together. *Cancer Metas. Rev.* 37, 55–73. doi: 10.1007/s10555-017-9715-8

Ge, E., Yang, Y., Gang, M., Fan, C., and Zhao, Q. (2019). Predicting human disease-associated circRNAs based on locality-constrained linear coding. *Genomics* 112, 1335–1342. doi: 10.1016/j.ygeno.2019.08.001

Geletu, M., Mohan, R., Arulanandam, R., Berger-Becvar, A., Nabi, I. R., Gunning, P. T., et al. (2018). Reciprocal regulation of the cadherin-11/Stat3 axis by caveolin-1 in mouse fibroblasts and lung carcinoma cells. *Biochim. Biophys. Acta Mol. Cell Res.* 1865, 794–802. doi: 10.1016/j.bbamcr.2018.02.004

He, D., Wang, X., Zhang, Y., Zhao, J., Han, R., and Dong, Y. (2019). DNMT3A/3B overexpression might be correlated with poor patient survival, hypermethylation and low expression of ESR1/PGR in endometrioid carcinoma: an analysis of the cancer genome atlas. *Chin. Med. J.* 132, 161–170. doi: 10.1097/CM9.0000000000000054

Hou, L.-K., Ma, Y.-S., Han, Y., Lu, G.-X., Luo, P., Chang, Z.-Y., et al. (2017). Association of microRNA-33a molecular signature with non-small cell lung cancer diagnosis and prognosis after chemotherapy. *PLoS ONE* 12:e0170431. doi: 10.1371/journal.pone.0170431

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Huang, Q. R., and Pan, X. B. (2019). Prognostic lncRNAs, miRNAs, and mRNAs form a competing endogenous RNA network in colon cancer. *Front. Oncol.* 9:712. doi: 10.3389/fonc.2019.00712

Jia, Y., Chen, L., Guo, S., and Li, Y. (2019). Baicalin induced colon cancer cells apoptosis through miR-217/DKK1-mediated inhibition of Wnt signaling pathway. *Mol. Biol. Rep.* 46, 1693–1700. doi: 10.1007/s11033-019-04618-9

Kelley, K. A., Wieghard, N., Chin, Y., Potter, A., Mori, M., Wong, M. H., et al. (2018). MiR-486-5p downregulation marks an early event in colorectal carcinogenesis. *Dis. Colon. Rectum.* 61, 1290–1296. doi: 10.1097/DCR.0000000000001192

Kurozumi, S., Matsumoto, H., Hayashi, Y., Tozuka, K., Inoue, K., Horiguchi, J., et al. (2017). Power of PgR expression as a prognostic factor for ER-positive/HER2-negative breast cancer patients at intermediate risk classified by the Ki67 labeling index. *BMC Cancer* 17:354. doi: 10.1186/s12885-017-3331-4

Lang, B., and Zhao, S. (2018). miR-486 functions as a tumor suppressor in esophageal cancer by targeting CDK4/BCAS2. *Oncol. Rep.* 39, 71–80. doi: 10.3892/or.2017.6064

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

Li, C., Zheng, X., Li, W., Bai, F., Lyu, J., and Meng, Q. H. (2018a). Serum miR-486-5p as a diagnostic marker in cervical cancer: with investigation of potential mechanisms. *BMC Cancer* 18:61. doi: 10.1186/s12885-017-3753-z

Li, F., Li, Q., and Wu, X. (2018b). Construction and analysis for differentially expressed long non-coding RNAs and MicroRNAs mediated competing endogenous RNA network in colon cancer. *PLoS ONE* 13:e0192494. doi: 10.1371/journal.pone.0192494

Li, H., Mou, Q., Li, P., Yang, Z., Wang, Z., Niu, J., et al. (2019). MiR-486-5p inhibits IL-22-induced epithelial-mesenchymal transition of breast cancer cell by repressing Dock1. *J. Cancer* 10, 4695–4706. doi: 10.7150/jca.30596

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. *Knowled. Based Syst.* 191:105261. doi: 10.1016/j.knosys.2019.105261

Ma, H., Tian, T., Liang, S., Liu, X., Shen, H., Xia, M., et al. (2016). Estrogen receptor-mediated miR-486-5p regulation of OLFM4 expression in ovarian cancer. *Oncotarget* 7, 10594–10605. doi: 10.18632/oncotarget.7236

Maierthaler, M., Benner, A., Hoffmeister, M., Surowy, H., Jansen, L., Knebel, P., et al. (2017). Plasma miR-122 and miR-200 family are prognostic markers in colorectal cancer. *Int. J. Cancer* 140, 176–187. doi: 10.1002/ijc.30433

Mao, Q., Quan, T., Luo, B., Guo, X., Liu, L., and Zheng, Q. (2016). MiR-375 targets KLF4 and impacts the proliferation of colorectal carcinoma. *Tumor Biol.* 37, 463–471. doi: 10.1007/s13277-015-3809-0

Mehta, A., and Patel, B. M. (2019). Therapeutic opportunities in colon cancer: focus on phosphodiesterase inhibitors. *Life Sci.* 230, 150–161. doi: 10.1016/j.lfs.2019.05.043

Rajendran, V. (2016). Structural analysis of oncogenic mutation of isocitrate dehydrogenase 1. *Mol. BioSyst.* 12, 2276–2287. doi: 10.1039/C6MB00182C

Rajendran, V., Gopalakrishnan, C., and Sethumadhavan, R. (2018). Pathological role of a point mutation (T315I) in BCR-ABL1 protein—a computational insight. *J. Cell. Biochem.* 119, 918–925. doi: 10.1002/jcb.26257

Rajendran, V., Purohit, R., and Sethumadhavan, R. (2012). *In silico* investigation of molecular mechanism of laminopathy caused by a point mutation (R482W) in lamin A/C protein. *Amino Acids* 43, 603–615. doi: 10.1007/s00726-011-1108-7

Ren, C., Chen, H., Han, C., Fu, D., Zhou, L., Jin, G., et al. (2016). miR-486-5p expression pattern in esophageal squamous cell carcinoma, gastric cancer and its prognostic value. *Oncotarget* 7, 15840–15853. doi: 10.18632/oncotarget.7417

Santasusagna, S., Moreno, I., Navarro, A., Muñoz, C., Martinez, F., Hernández, R., et al. (2018). miR-328 mediates a metabolic shift in colon cancer cells by targeting SLC2A1/GLUT1. *Clin. Transl. Oncol.* 20, 1161–1167. doi: 10.1007/s12094-018-1836-1

Sen, N., Satija, Yatendra K., and Das, S. (2011). PGC-1α, a key modulator of p53, promotes cell survival upon metabolic stress. *Mol. Cell* 44, 621–634. doi: 10.1016/j.molcel.2011.08.044

Seo, A. H., Moeng, S., Sim, S., Kuh, J. H., Choi, Y. S., and Park, K. J. (2019). MicroRNA-based combinatorial cancer therapy: effects of MicroRNAs on the efficacy of anti-cancer therapies. *Cells* 9:E29. doi: 10.3390/cells9010029

Shen, C., Ding, Y., Tang, J., Jiang, L., and Guo, F. (2019). LPI-KTASLP: Prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* 7, 13486–13496. doi: 10.1109/ACCESS.2019.2894225

Sheng, S., Xie, L., Wu, Y., Ding, M., Zhang, T., and Wang, X. (2019). MiR-144 inhibits growth and metastasis in colon cancer by down-regulating SMAD4. *Biosci. Rep.* 39:BSR20181895. doi: 10.1042/BSR20181895

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G. S., Barzi, A., et al. (2017). Colorectal cancer statistics, 2017. 67, 177–193. doi: 10.3322/caac.21395

Su, J., Lu, E., Lu, L., and Zhang, C. (2017). MiR-29a-3p suppresses cell proliferation in laryngocarcinoma by targeting prominin 1. *FEBS Open Bio.* 7, 645–651. doi: 10.1002/2211-5463.12199

Sun, J., Zhao, H., Lin, S., Bao, S., Zhang, Y., Su, J., et al. (2019). Integrative analysis from multi-centre studies identifies a function-derived personalized multi-gene signature of outcome in colorectal cancer. *J. Cell. Mol. Med.* 23, 5270–5281. doi: 10.1111/jcmm.14403

Sundaram, P., Hultine, S., Smith, L. M., Dews, M., Fox, J. L., Biyashev, D., et al. (2011). p53-responsive miR-194 inhibits thrombospondin-1 and promotes angiogenesis in colon cancers. 71, 7490–7501. doi: 10.1158/0008-5472.CAN-11-1124

Tian, F., Wang, J., Ouyang, T., Lu, N., Lu, J., Shen, Y., et al. (2019). MiR-486-5p serves as a good biomarker in nonsmall cell lung cancer and suppresses cell growth with the involvement of a target PIK3R1. *Front. Genet.* 10:688. doi: 10.3389/fgene.2019.00688

Tian, Z.-Q., Li, Z.-H., Wen, S.-W., Zhang, Y.-F., Li, Y., Cheng, J.-G., et al. (2015). Identification of commonly dysregulated genes in non-small-cell lung cancer by integrated analysis of microarray data and qRT-PCR validation. *Lung* 193, 583–592. doi: 10.1007/s00408-015-9726-6

Tong, Z., Liu, N., Lin, L., Guo, X., Yang, D., and Zhang, Q. (2015). miR-125a-5p inhibits cell proliferation and induces apoptosis in colon cancer via targeting BCL2, BCL2L12 and MCL1. *Biomed. Pharmacother.* 75, 129–136. doi: 10.1016/j.biopha.2015.07.036

Truong, A. D., Rengaraj, D., Hong, Y., Hoang, C. T., Hong, Y. H., and Lillehoj, H. S. (2017). Differentially expressed JAK-STAT signaling pathway genes and target microRNAs in the spleen of necrotic enteritis-afflicted chicken lines. *Res. Vet. Sci.* 115, 235–243. doi: 10.1016/j.rvsc.2017.05.018

Wang, H.-Y., Wang, W., Liu, Y.-W., Li, M.-Y., Liang, T.-Y., Li, J.-Y., et al. (2017). Role of KCNB1 in the prognosis of gliomas and autophagy modulation. *Sci. Rep.* 7:14. doi: 10.1038/s41598-017-00045-7

Wang, L., Wei, Z., Wu, K., Dai, W., Zhang, C., Peng, J., et al. (2018). Long noncoding RNA B3GALT5-AS1 suppresses colon cancer liver metastasis via repressing microRNA-203. *Aging* 10, 3662–3682. doi: 10.18632/aging.101628

Wei, H. T., Guo, E. N., Liao, X. W., Chen, L. S., Wang, J. L., Ni, M., et al. (2018). Genomescale analysis to identify potential prognostic microRNA biomarkers for predicting overall survival in patients with colon adenocarcinoma. *Oncol. Rep.* 40, 1947–1958. doi: 10.3892/or.2018.6607

Wu, Q., Meng, W.-Y., Jie, Y., and Zhao, H. (2018). LncRNA MALAT1 induces colon cancer development by regulating miR-129-5p/HMGB1 axis. *J. Cell. Physiol.* 233, 6750–6757. doi: 10.1002/jcp.26383

Xia, L., Song, M., Sun, M., Chen, W., and Yang, C. (2019). miR-486 promotes capan-2 pancreatic cancer cell proliferation by targeting phosphatase and tensin homolog deleted on chromosome 10 (PTEN). *Front. Genet.* 10:541. doi: 10.3389/fgene.2019.00541

Xiao, C., and Rajewsky, K. (2009). MicroRNA control in the immune system: basic principles. *Cell* 136, 26–36. doi: 10.1016/j.cell.2008.12.027

Xu, J., Zhao, J., and Zhang, R. (2016). Four microRNAs signature for survival prognosis in colon cancer using TCGA data. *Sci. Rep.* 6:38306. doi: 10.1038/srep38306

Yang, W., Ma, J., Zhou, W., Li, Z., Zhou, X., Cao, B., et al. (2019). Identification of hub genes and outcome in colon cancer based on bioinformatics analysis. *Cancer Manag. Res.* 11, 323–338. doi: 10.2147/CMAR.S173240

Zhao, Q., Yang, Y., Ren, G., Ge, E., and Fan, C. (2019). Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations. *IEEE Trans. NanoBiosci.* 18, 578–584. doi: 10.1109/TNB.2019.2922214

Zhao, Q., Yu, H., Ming, Z., Hu, H., Ren, G., and Liu, H. (2018a). The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. *Mol. Ther. Nucleic Acids* 13, 464–471. doi: 10.1016/j.omtn.2018.09.020

Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018b). IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zheng, Y., and Zhu, G. (2018). HMGB1 suppresses colon carcinoma cell apoptosis triggered by coculture with dendritic cells via an ER stressassociated autophagy pathway. *Mol. Med. Rep.* 17, 3123–3132. doi: 10.3892/mmr.2017.8202

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018a). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018b). An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. *Mol. Neurobiol.* 55, 3684–3697. doi: 10.1007/s12035-017-0572-9

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

Zhou, X. G., Huang, X. L., Liang, S. Y., Tang, S. M., Wu, S. K., Huang, T. T., et al. (2018c). Identifying miRNA and gene modules of colon cancer associated with pathological stage by weighted gene co-expression network analysis. *Onco. Targets Ther.* 11, 2815–2830. doi: 10.2147/OTT.S163891

Zhu, H., Wu, T.-C., Chen, W.-Q., Zhou, L.-J., Wu, Y., Zeng, L., et al. (2013). Screening for differentially expressed genes between left- and right-sided colon carcinoma by microarray analysis. *Oncol. Lett.* 6, 353–358. doi: 10.3892/ol.2013.1414

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 © 2020 Zhu, Xu, Liu, Qiao, Sun 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 Genetics

ORIGINAL RESEARCH
published: 17 March 2020
doi: 10.3389/fgene.2020.00237

Combination of Four Serum Exosomal MiRNAs as Novel Diagnostic Biomarkers for Early-Stage Gastric Cancer

Shuli Tang[1,2†], Jianan Cheng[1,2†], Yuanfei Yao[1,2], Changjie Lou[1], Liang Wang[3], Xiaoyi Huang[4] and Yanqiao Zhang[1,2]**

[1] Department of Gastrointestinal Medical Oncology, Harbin Medical University Cancer Hospital, Harbin, China, [2] Translational Medicine Research and Cooperation Center of Northern China, Heilongjiang Academy of Medical Sciences, Harbin, China, [3] Department of Tumor Biology, H. Lee Moffitt Cancer Center, Tampa, FL, United States, [4] Biotherapy Center, Harbin Medical University Cancer Hospital, Harbin, China

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Ya Cui,
University of California, Irvine, United States
Guangchao Yang,
City of Hope National Medical Center, United States

***Correspondence:**
Yanqiao Zhang
yanqiaozhang@ems.hrbmu.edu.cn
Xiaoyi Huang
xyhuang@hrbmu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 07 January 2020
Accepted: 27 February 2020
Published: 17 March 2020

Citation:
Tang S, Cheng J, Yao Y, Lou C, Wang L, Huang X and Zhang Y (2020) Combination of Four Serum Exosomal MiRNAs as Novel Diagnostic Biomarkers for Early-Stage Gastric Cancer. Front. Genet. 11:237. doi: 10.3389/fgene.2020.00237

Gastric cancer (GC) remains a leading cause of cancer-related mortality in the United States and China, there is an urgent need to discover novel non-invasive biomarkers for the early diagnosis of GC to improve the prognosis of GC patients. Exosomal miRNAs are considered promising biomarkers for cancer diagnosis. Using next-generation sequencing (NGS), bioinformatics and further validation, we identified and evaluated exosomal miRNAs in serum as early diagnostic markers for GC. NGS revealed that the average mappable reads in the RNA libraries were about 6.5 million per patient including miRNAs (73.38%), rRNAs (17.10%), snRNAs (8.83%), snoRNAs (0.65%), and tRNAs (0.04%). A total of 66 up and 13 down-regulated exosomal miRNAs were found in the screened cohort. In the validation cohort, by comparing with healthy individuals, higher levels of serum exosomal miR-92b-3p, let-7g-5p, miR-146b-5p, and miR-9-5p were found to be significantly associated with early-stage GC ($p < 0.05$). Diagnostic power of the combined panels of the exosomal miRNAs or the combination of exosomal miRNAs and CEA outperformed that of single exosomal miRNA marker for establishing a diagnosis of early-stage GC. The combined diagnosis of exosomal miR-92b-3p + let-7g-5p + miR-146b-5p + miR-9-5p with CEA had the most powerful efficiency with an AUC up to 0.786. In addition, serum levels of exosomal miR-92b-3p were significantly associated with poor cohesiveness ($p = 0.0021$), let-7g-5p and miR-146b-5p were significantly correlated with nerve infiltration ($p = 0.0234$ and $p = 0.0126$, respectively), and miR146b-5p was statistically correlated with tumor invasion depth in early-stage GC ($p = 0.0089$). In conclusion, serum exosomal miR-92b-3p, -146b-5p, -9-5p, and let-7g-5p may serve as potential non-invasive biomarkers for early diagnosis of GC.

Keywords: exosomes, microRNAs, gastric cancer, early diagnosis, liquid biopsy

INTRODUCTION

Gastric cancer (GC) remains a leading cause of cancer-related mortality in the United States (Bray et al., 2018) and China (Chen et al., 2018). Because GC is mostly asymptomatic until it progresses to advanced stages, and lacking of an efficient biomarker with high sensitivity and specificity, the prognosis of GC patients remains poor. Hence, there is an urgent need to discover novel non-invasive biomarkers for the early diagnosis of GC.

Exosomes are 30–100 nm diameter, membrane-enclosed vesicles that are secreted by numerous cell types and present in various body fluids, such as blood, urine, saliva, etc. (Ohno et al., 2013). Exosomes act as mediators in intercellular communications by transporting proteins, lipids and nucleic acids to recipient cells, resulting in the modulation of different processes such as tumor invasion, angiogenesis, metastasis, and chemoresistance (Kosaka et al., 2016; Hu et al., 2019). The features of quick detection, convenient collection and minimal pain make exosomes serve as an ideal liquid biopsy tool for clinical application (Chi, 2016; Batth et al., 2017; Siravegna et al., 2017). MicroRNAs (miRNAs) are small non-coding RNAs (18–24 nt) that regulate protein translation. Extracellular miRNAs can be packaged into exosomes, which protects them from digestion by RNases (Manterola et al., 2014). High stability and enrichment of circulating exosomal miRNAs offer an attractive option for cancer diagnosis and prognosis. For example, serum exosomal miR-1246 and miR-21 could be used as promising diagnostic biomarkers for pancreatic cancer (Madhavan et al., 2015; Lai et al., 2017), prostate cancer (Bhagirath et al., 2018), and hepatocellular carcinoma (Sugimachi et al., 2015). To date, the majority of efforts are focused on breast cancer, prostate cancer, colorectal cancer, pancreatic cancer, hepatocellular carcinoma, and lung cancer (Jin et al., 2017). However, the relevance of serum exosomal miRNAs in early-stage GC has not been clearly elucidated.

We previously established an optimized procedure for screening and validating serum exosomal RNA biomarkers in prostate cancer and colorectal cancer (Huang et al., 2015; Wang et al., 2017). In this study, we used next-generation sequencing (NGS) to identify differential miRNA signatures in exosomes isolated from serum of early-stage GC patients, and provide a non-invasive method for early-stage GC detection. Finally, real-time qRT-PCR verified that the expression levels of exosomal miR-92b-3p, -146b-5p, -9-5p, and let-7g-5p were significantly higher in early-stage GC by comparison with matched healthy individuals, which established their utility as potential early-stage GC biomarkers.

MATERIALS AND METHODS

Patients and Samples

Serum samples were obtained from early-stage GC patients and healthy individuals at the Harbin Medical University Cancer Hospital between 2015 and 2016. The screening cohort included 36 early-stage GC patients, and 12 age and gender-matched healthy individuals. Fifty pairs of newly recruited early-stage GC patients and matched healthy individuals were included in the validation cohort. All serum samples were taken before treatment. All patients were histopathologically confirmed as gastric non-cardia adenocarcinoma patients (stages I and II) after operation. Clinical information including age, gender, pathological differentiation, pathological tumor stage, tumor invasion depth, presence of lymph node metastases, WHO cohesiveness, carcinoembryonic antigen (CEA) and carbohydrate antigen 199 (CA199) levels, and nerve infiltration were collected from medical records. The pathological tumor stage was identified according to the seventh edition AJCC TNM classification. CEA and CA199 levels were obtained from routine blood test before surgery, and CEA $>$ 5 ng/ml and CA199 $>$ 37 U/ml were defined as high levels. Inclusion criteria for healthy individuals, who were used as normal controls, included no malignancy, autoimmune disorders, endocrine disease, hepatitis, or HIV infection. The studies have been performed in accordance with the Declaration of Helsinki. All participants gave written informed consent, and the Harbin Medical University Cancer Hospital Ethics Committee approved the study. The overview flowchart of this study is shown in **Figure 1A**, and patient data are summarized in **Table 1**. Total 5 ml venous blood samples were collected before initial treatment, and was centrifugated at 3000 *g* for 10 min at 4°C within 2 h after collection. The supernatant (serum) was then transferred to RNase/DNase-free tubes and stored at −80°C until further processing.

Exosome Isolation

Exosomes were isolated using the ExoQuick exosome precipitation solution (System Biosciences, Mountain View, CA, United States) according to the manufacturer's instructions with slight modifications. Briefly, 300 μl of cell-free serum samples were mixed with 75 μl of ExoQuick solution and RNase A (Sigma, St. Louis, MO, United States) to a final concentration of 10 μg/ml. The mixture was kept at 4°C overnight. Afterward, murine RNase inhibitor (NEB, Ipswich, MA, United States) was added to the mixture (150 units/ml) before centrifugation at 1500 *g* for 30 min. Finally, the supernatants were aspirated and the pelleted exosomes were re-suspended in 50 μl 1 × PBS, and immediately used for RNA extraction.

Exosomal RNA Extraction and RNA Library Preparation

RNA of exosomes was isolated using miRNeasy micro kit (Qiagen, Valencia, CA, United States) according to the manufacturer's protocol. The extracted RNA was eluted with 14 μl of RNase-free water. The library preparation was based on the protocols of Multiplex Small RNA Library Prep Set for Illumina (NEB, Ipswich, MA, United States) as previously described (Huang et al., 2013, 2015). Total 2 ng of isolated RNA was reverse-transcribed into cDNA sequencing libraries. Twelve sequencing libraries with different indices were pooled at a final amount of 2 nM and subjected to DNA sequencing.

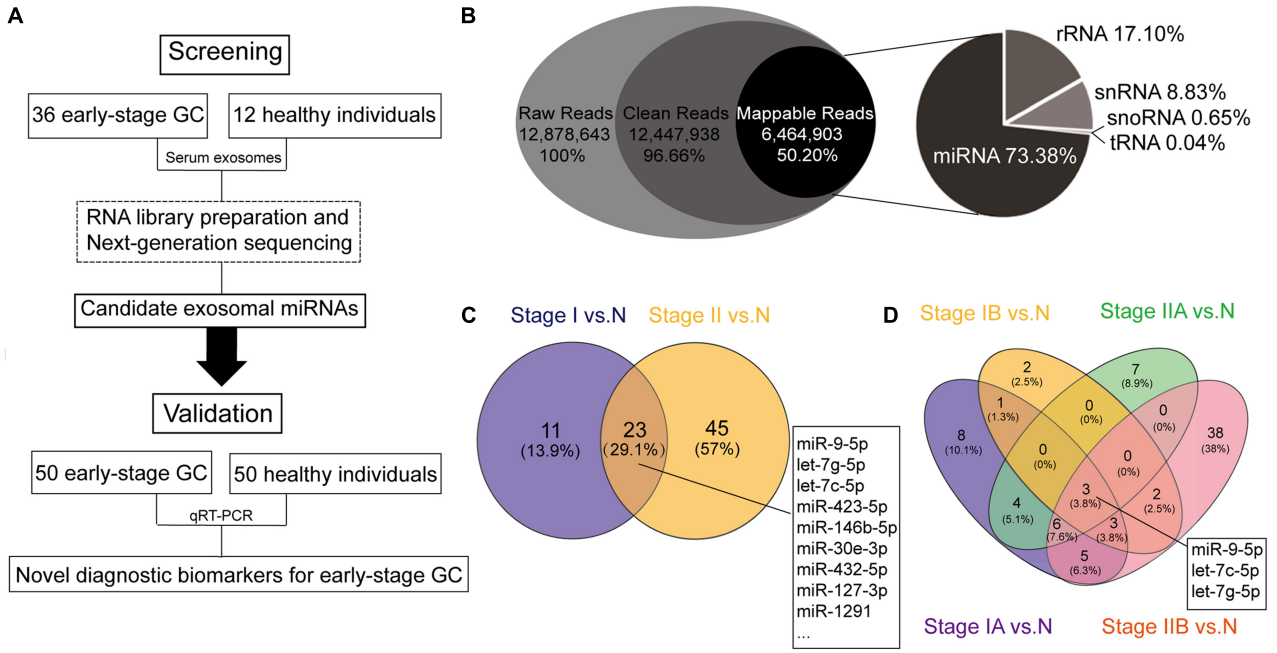


FIGURE 1 | Schematic flowchart of the analytical pipeline and the results based on the next-generation sequencing (NGS). **(A)** Schematic flowchart of identification of early-stage gastric cancer (GC)-specific exosomal miRNAs. **(B)** RNA biotypes in the small RNA library prepared from human serum exosomes. Pie chart showing the mean percentage reads of serum exosomal small RNA library. Raw reads were the sequences obtained by RNA sequencing (RNA-seq). Clean reads were generated after read filtering and adapter trimming. Mappable reads were the RNA-seq reads mapped to known human RNAs, and were sorted into the following small ncRNAs: micro RNA (miRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and transfer RNA (tRNA). **(C)** Venn diagram showing overlap of the differentially expressed miRNAs between stage I gastric cancer vs. healthy individuals (stage I vs. N) and stage II gastric cancer vs. healthy individuals (stage II vs. N). **(D)** Venn diagram showing overlap of the differentially expressed miRNAs in all four stages (stages IA, IB, IIA, and IIB) vs. healthy individuals (N).

Sequencing Data Analysis

Next-generation sequencing was carried out on an Illumina HiSeq2000 platform by Novogen, Inc. (Beijing, China). The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq SR Cluster kit v3-cBot-HS (Illumina). After cluster generation, the libraries were sequenced and 50 bp single-end reads were generated. Clean data were obtained by processing raw data in FASTQ format through custom perl and python scripts. Clean reads with certain range of length were mapped to reference sequence using Bowtie from miRBase (Release 20) and NCBI human genome reference sequences. The softwares miREvo and mirdeep2 were integrated to predict novel miRNA from the clean data. The DESeq R package (version 1.8.3) was used to identify differential expression of miRNAs between two groups. The *p*-values was adjusted using the Benjamini & Hochberg method. Corrected *p*-value of 0.05 was set as the threshold for significantly differential expression by default. To make the sequencing profiles comparable, we normalized RNA profiles as read count of a target RNA per million mapped reads (RPM).

Real-Time Quantitative RT-PCR and Data Normalization

Some of the differentially expressed miRNAs identified by sequencing were validated using miScript SYBR Green PCR kit (Qiagen, Valencia, CA, United States). MiR-9-5p, let-7g-5p, let-7c-5p, miR-146b-5p, miR-101-3p, miR-92b-3p, miR-21-5p, and miR-26a-5p (MS00010752, MS00008337, MS00003129, MS00003542, MS00008372, MS00032144, MS00009079, MS00029239; Qiagen, Valencia, CA, United States) were selected. MiR-30e-5p (MS00009401; Qiagen, Valencia, CA, United States) was selected as the endogenous normalization control based on our previous study (Huang et al., 2015). Briefly, 4 ng of total serum exosomal RNA was reverse-transcribed into cDNA in a 20 μl reaction using the miScript II RT kit (Qiagen, Valencia, CA, United States), followed by dilution of the cDNA products for subsequent real-time PCR reactions. Then, the quantitative PCR was carried out on LightCycler® 480 Real-Time PCR System (Roche Diagnostics, Mannheim, Germany) in a 384-well plate. The PCR reactions were set as follows: 95°C for 15 min, 40 cycles of 94°C for 15 s, 55°C for 30 s, and 72°C for 30 s. All quantitative experiments were independently repeated at least three times to remove any outliers.

Data Processing and Statistical Analysis

Cycle threshold (Ct) values, obtained from real-time qPCR, were calculated as the expression data for miRNAs. ΔCt values were analyzed directly to compare different exosomal miRNA transcription levels (Silver et al., 2006). $\Delta Ct = Ct_{targetmiRNA} - Ct_{miR-30e-5p}$, and lower ΔCt values indicated higher expression

TABLE 1 | Clinicopathological features of early-stage gastric cancer patients in screening and validating cohort.

Variables	Screening cohort (%)	Validating cohort (%)	*p*-Value
Total number	36	50	
Age (mean ± SD)	56.50 ± 9.87	58.10 ± 10.41	0.557
Gender			
Female	11 (30.5)	12 (24.0)	0.622
Male	25 (69.5)	38 (76.0)	
Pathological differentiation			0.125
Poor	22 (61.1)	21 (42.0)	
Moderate + Well	14 (38.9)	29 (58.0)	
Invasion depth			0.986
T1	11 (30.5)	16 (32.0)	
T2	10 (27.8)	14 (28.0)	
T3	15 (41.7)	20 (40.0)	
Lymph nodes metastasis			0.374
Negative	29 (80.5)	44 (88.0)	
Positive	7 (19.5)	6 (12.0)	
TNM stage			0.782
IA	11 (30.5)	15 (30.0)	
IB	10 (27.8)	15 (30.0)	
IIA	8 (22.2)	14 (28.0)	
IIB	7 (19.5)	6 (12.0)	
Poorly cohesive			0.045
Negative	17 (47.2)	35 (70.0)	
Positive	19 (52.8)	15 (30.0)	
Nerve infiltrate			1
Negative	21 (58.3)	30 (60.0)	
Positive	15 (41.7)	20 (40.0)	
CEA			0.694
Normal	34 (94.4)	45 (90.0)	
High	2 (5.6)	5 (10.0)	
CA199			0.419
Normal	35 (97.2)	50 (100.0)	
High	1 (2.8)	0 (0.0)	

levels of exosomal miRNAs. Fold change (FC) was based on FC = $2^{-\Delta\Delta Ct}$, and represented as follows: if FC > 1, true FC = FC and if FC < 1, true FC = −1/FC. The subsequent statistical analyses were performed using GraphPad Prism version 6.0 (GraphPad Software, La Jolla, CA, United States) and SPSS software version 20.0 (IBM, Corp., Armonk, NY, United States). Data were presented as the mean ± SD. Unpaired Student's *t*-test was used to analyze the differences in miRNA expression between early-stage GC patients and healthy individuals. Receiver operating characteristic (ROC) curve was used to evaluate the diagnostic power of the candidate exosomal miRNAs for early-stage GC. The cut-off value was determined using the Youden Index. Logistic regression adjusted for sex and age was utilized to establish the combination of exosomal miRNAs for GC diagnosis, so did the combined models based on exosomal miRNAs and traditional biomarkers. The performance of the combined models was evaluated by the area under the ROC curves (AUC). A two-sided $p < 0.05$ was considered statistically significant.

RESULTS

Mapping of RNA Sequencing

The average raw reads from the RNA sequencing libraries were about 13 million, from which approximately 12.5 million (96.66%) reads with certain range of length (clean reads) were received after trimming. Among these, about 6.5 million reads (50.20%) were mapped to known RNA species. We annotated the mapped reads into the following biotypes of small ncRNAs: miRNAs, ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), and transfer RNAs (tRNAs). Analysis of the mapped reads revealed miRNAs were the most common, accounting for 73.38% of all mappable RNAs, followed by 17.10% rRNAs, 8.83% snRNAs, 0.65% snoRNAs, and 0.04% tRNAs (**Figure 1B**). The raw sequencing data have been deposited in the Gene Expression Omnibus database (accession number: GSE130654).

Exosomal MiRNA Profiling in the Screening Stage

Next, to identify candidate serum exosomal miRNAs for early-stage GC diagnosis, miRNAs with log2 transformed read counts < 5 and $p > 0.05$ were removed from the RNA sequencing data. A total of 79 known miRNAs remained after the initial screening step, including 66 up-regulated and 13 down-regulated miRNAs compared to normal control. Further systematic evaluation on the differential expression of the remaining miRNAs using Venn diagrams (**Figure 1C**) revealed that 23 miRNAs were significantly aberrant in both stage I and stage II GC when compared with normal control (**Supplementary Table S1**), 11 were uniquely found between stage I GC and normal control, and 45 miRNAs between stage II GC and normal control. When the early stage GC was further stratified into stages IA, IB, IIA, and IIB, miR-9-5p, let-7g-5p, and let-7c-5p were the only three miRNAs significantly aberrant in all early stages of GC, and miR-146b-5p was significantly aberrant in stages IA, IIA and IIB, but not in stage IB (**Figure 1D**). Meanwhile, many serum exosomal miRNAs were differentially expressed in a unique stage, suggesting the complexity of the miRNA transcriptome in tumorigenesis of GC. For example, miR-21-5p and miR-26-5p were significantly aberrant in stage IIB.

To assess the diagnostic value of the aberrantly expressed miRNAs, we performed ROC analysis based on the expression levels of the miRNAs. The miRNAs with $p < 0.05$, AUC ≥ 0.75, and both sensitivity and specificity at least ≥ 60% were selected as potential candidates for diagnostic markers. After screening, we selected eight serum exosomal miRNAs for downstream validation (**Table 2**).

Evaluation of the Diagnostic Potential of Exosomal MiRNAs for Early-Stage GC

To investigate whether candidate exosomal miRNAs in serum could be used as a potential biomarker, an independent patient cohort was validated by real-time qRT-PCR. Among the eight candidates, only four miRNAs were found to be potential diagnostic biomarkers for early-stage GC patients. As shown in

TABLE 2 | Diagnostic efficiencies of serum exosomal miRNA candidates in early-stage gastric cancer.

Column ID	AUC	*p*-Value	Cut-off	Sensitivity %	Specificity %
miR-9-5p	0.920	<0.0001	>11.36	80	90.91
let-7g-5p	0.850	0.0005	>11.42	74.29	90.91
let-7c-5p	0.886	0.0001	>12.03	80	90.91
miR-146b-5p	0.784	0.0048	>8.705	71.43	81.82
miR-92b-3p	0.788	0.0304	>13.04	66.67	90.91
miR-101-3p	0.805	0.0057	>10.28	75	90.91
miR-21-5p	1.000	0.0005	>12.35	100	90.91
miR-26a-5p	0.961	0.0013	>12.96	85.71	90.91

AUC, area under the receiver operating characteristic curve.

Figures 2A–D, the ΔCt values of miR-92b-3p, let-7g-5p, miR-146b-5p, and miR-9-5p were significantly lower in early-stage GC patients as compared to healthy individuals ($p < 0.05$). Meanwhile, serum levels of CEA failed to distinguish early-stage GC patients from healthy individuals ($p = 0.7329$) (**Figure 2E**). Compared to healthy individuals, the expression levels of miR-92b-3p, let-7g-5p, miR-146b-5p, and miR-9-5p were 2.018, 3.926, 1.784, and 1.266-fold higher, respectively, in early-stage GC patients (**Supplementary Table S2**). We assessed the diagnostic capacity of each miRNA by computing their ROC curve. The results are illustrated in **Figures 3A–D**. Among the four candidates, let-7g-5p possessed the highest diagnostic power in discriminating early-stage GC patients and healthy individuals with AUC at 0.756 (95% CI, 0.659–0.892, cut-off < 4.184, sensitivity = 54%, specificity = 88%), followed by miR-92b-3p at 0.714 (95% CI, 0.613–0.815, cut-off < 1.690, sensitivity = 58%, specificity = 80%), miR-146b-5p at 0.674 (95% CI, 0.569–0.779, cut-off < 3.784, sensitivity = 46%, specificity = 82%), and miR-9-5p at 0.626 (95% CI, 0.515–0.738, cut-off < 4.932, sensitivity = 50%, specificity = 84%). In contrast, the AUC value of CEA was 0.520 (95% CI, 0.405–0.635, cut-off < 1.440, sensitivity = 38%, specificity = 70%) (**Figure 3E**).

Combining Multiple Biomarkers Enhances the Diagnostic Power for Early-Stage GC

We next explored whether combination of the four exosomal miRNAs could be used to better discriminate early-stage GC patients from healthy individuals. As shown in **Table 3**, totally seven combined panels hold stronger diagnostic power than any single exosomal miRNA. The most powerful diagnostic panel consisting of miR-92b-3p and let-7g-5p received a AUC of 0.775, with sensitivity of 64% and specificity of 78%. The most sensitive diagnostic panel was the combination of miR-92b-3p + miR-146b-5p + miR-9-5p (68%), and the combination of miR-146b-5p + miR-9-5p was the most specific panel (88%). Moreover, significant synergistic effect was also found in the combinations of exosomal miRNA panels and traditional biomarkers (**Table 3**). The combined diagnosis of exosomal miR-92b-3p + let-7g-5p + miR-146b-5p + miR-9-5p with CEA had the most powerful efficiency with an AUC up to 0.786. Interestingly, compared with CEA and the exosomal miRNA panels, the combinations of CEA and any exosomal miRNA panel led to a mild loss of sensitivities but dramatically improved specificities. Among the different combined panels, the best specificity in the panel consisting of miR-92b-3p + miR-146b-5p + miR-9-5p + CEA reached 96% with an AUC of 0.758 (**Table 3**). Our results indicated that the exosomal miRNA combination panels and the combination of exosomal miRNAs with CEA could serve as potential predictive markers for the diagnosis of early-stage GC.

Correlation of Serum Exosomal MiRNAs With Clinicopathological Features of Early-Stage GC

Next, we focused on the relationship between the miRNAs and the clinicopathological features of early-stage GC patients. There was no statistical association between miR-92b-3p, let-7g-5p, miR-146b-5p, and miR-9-5p with age, gender, pathological differentiation, and lymph node metastases. However, the level of miR-92b-3p was significantly ($p = 0.0021$) lower in non/poorly cohesive GC patients (**Figure 4A**), the levels of let-7g-5p and miR-146b-5p were significantly ($p = 0.0234$ and $p = 0.0126$, respectively) lower in non/low nerve infiltration GC patients (**Figures 4B,C**), and the level of miR146b-5p was significantly ($p = 0.0089$) increased with the increase in tumor invasion depth (**Figure 4D**). In this dataset, we did not find any statistical correlations between miR-9-5p and nerve infiltration, poorly cohesive, and invasion depth ($p > 0.05$).

Expression of MiRNAs in GC Tissues

To further clarify the relationship between the identified exosomal miRNA biomarkers and GC, the miRNA expression data derived from 446 GC tissues and 45 normal gastric tissues were downloaded from The Cancer Genome Atlas (TCGA) database.[1] We used DESeq R package (version 1.8.3) to identify differentially expressed miRNAs between the two groups. Compared to normal gastric tissues, the expression levels of miR-92b-3p and miR-146b-5p were significantly upregulated in GC tissues ($p = 0.0042$ and $p < 0.0001$, respectively), and the expression levels of let-7g-5p and miR-9-5p were significantly downregulated in GC tissues ($p = 0.0003$ and $p = 0.0026$, respectively) (**Supplementary Figure S1**).

DISCUSSION

Although treatments for GC have advanced in recent decades, the early diagnosis relies on traditional imaging examinations, biopsy, or serum biomarkers. Due to inherent clinical heterogeneity, early diagnosis continues to be one of the key challenges in GC. As we known, reliable molecular markers play vital roles in detecting cancers, monitoring recurrence, assessing prognosis, and improving effective personalized cancer therapy (Zhou et al., 2018; Sun et al., 2019). Thus, development of a new non-invasive diagnostic method has become critical to improve therapeutic effect. In this study, we demonstrated that a panel of the circulating exosomal miRNAs had higher diagnostic

[1] cancergenome.nih.gov/

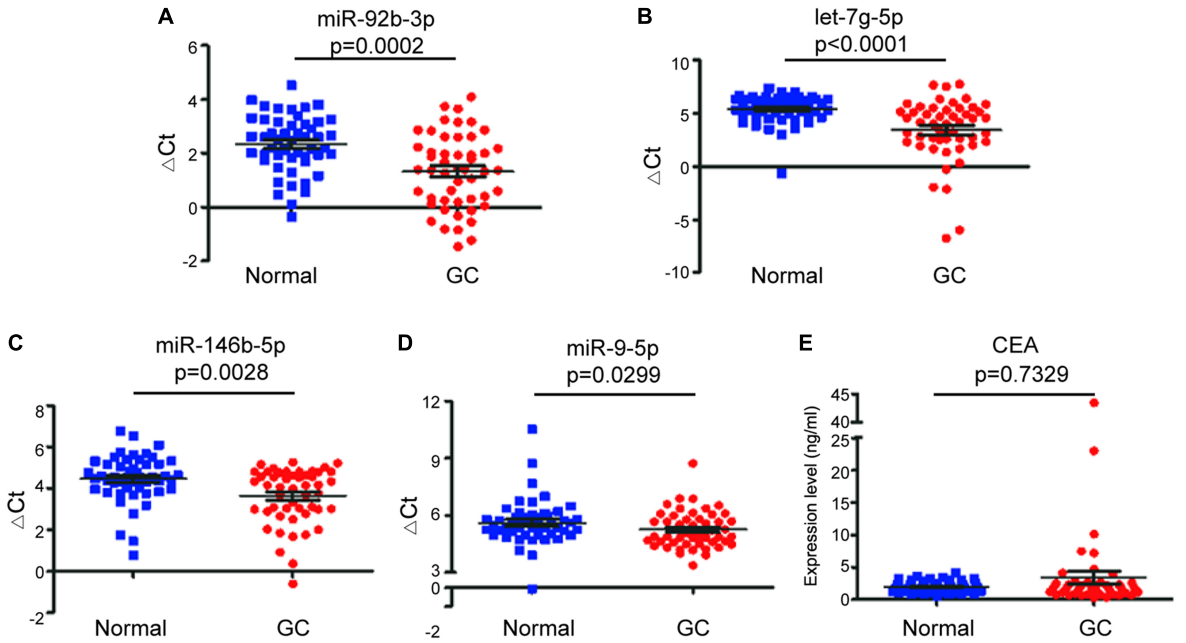


FIGURE 2 | Comparison of expression of exosomal miRNAs in early-stage GC patients (GC) and healthy individuals (Normal). The scatter plots of the mean ± SD of ΔCt values of miR-92b-3p **(A)**, let-7g-5p **(B)**, miR-146b-5p **(C)**, and miR-9-5p **(D)**, and the mean ± SD of expression levels of serum CEA **(E)** between early-stage GC patients and healthy individuals.

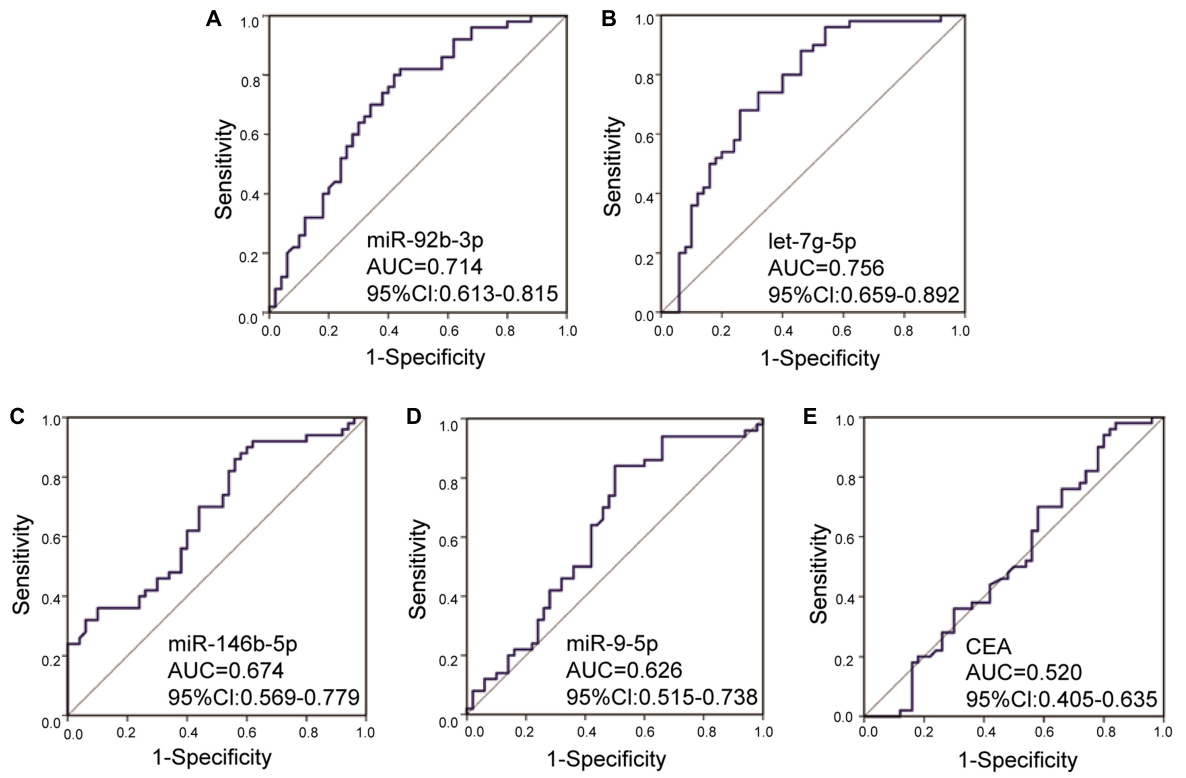


FIGURE 3 | Receiver operating characteristic curves for the serum exosomal miRNAs and CEA to discriminate early-stage gastric cancer patients from healthy individuals. The AUC values of miR-92b-3p **(A)**, let-7g-5p **(B),** miR-146b-5p **(C)**, miR-9-5p **(D)**, and CEA **(E)** were 0.714 (p = 0.0002), 0.756 (p < 0.0001), 0.674 (p = 0.0028), 0.626 (p = 0.0299), and 0.520 (p = 0.7329), respectively. AUC, area under the receiver operating characteristic curve; CI, confidence interval.

TABLE 3 | ROC curve analysis on combinations of serum exosomal miRNAs for early-stage gastric cancer.

miRNA combination	AUC	*p*-Value	95% CI	Cut-off	Sensitivity %	Specificity %
miR-92b-3p + let-7g-5p	0.775	<0.0001	0.684–0.866	>0.484	64	78
miR-92b-3p + miR-146b-5p	0.736	<0.0001	0.638–0.834	>0.536	60	82
miR-146b-5p + miR-9-5p	0.705	0.0004	0.604–0.806	>0.607	44	88
miR-92b-3p + let-7g-5p + miR-146b-5p	0.774	<0.0001	0.683–0.865	>0.534	58	86
miR-92b-3p + let-7g-5p + miR-9-5p	0.774	<0.0001	0.684–0.865	>0.525	60	82
miR-92b-3p + miR-146b-5p + miR-9-5p	0.75	<0.0001	0.654–0.845	>0.492	68	74
miR-92b-3p + let-7g-5p + miR-146b-5p + miR-9-5p	0.773	<0.0001	0.682–0.864	>0.526	60	84
miR-92b-3p + CEA	0.722	0.0001	0.622–0.823	>0.527	60	80
miR-92b-3p + let-7g-5p + CEA	0.784	<0.0001	0.694–0.875	>0.599	58	90
miR-92b-3p + miR-146b-5p + CEA	0.743	<0.0001	0.645–0.841	>0.596	56	88
miR-92b-3p + miR-9-5p + CEA	0.724	0.0001	0.624–0.824	>0.533	58	80
miR-146b-5p + miR-9-5p + CEA	0.709	0.0003	0.609–0.810	>0.584	48	90
miR-92b-3p + let-7g-5p + miR-146b-5p + CEA	0.784	<0.0001	0.693–0.874	>0.640	54	94
miR-92b-3p + let-7g-5p + miR-9-5p + CEA	0.784	<0.0001	0.694–0.874	>0.599	58	90
miR-92b-3p + miR-146b-5p + miR-9-5p + CEA	0.758	<0.0001	0.663–0.853	>0.701	48	96
miR-92b-3p + let-7g-5p + miR-146b-5p + miR-9-5p + CEA	0.786	<0.0001	0.695–0.876	>0.613	58	90

AUC, area under the receiver operating characteristic curve; CI, confidence interval; ROC, receiver operating characteristic.

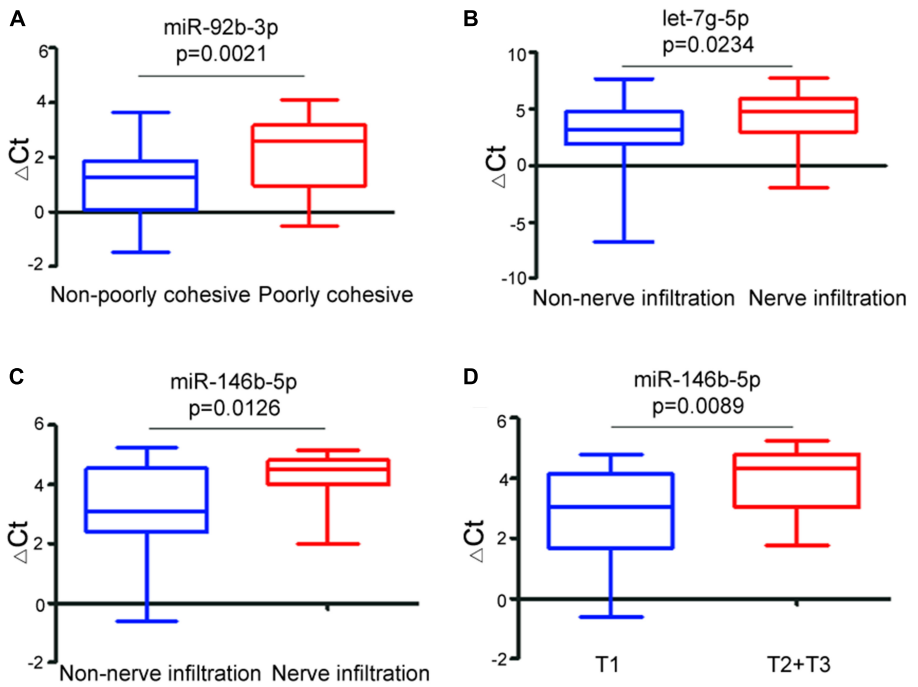


FIGURE 4 | Relationship between the ΔCt values of exosomal miRNA biomarkers and clinicopathological features of early-stage gastric cancer. Boxplot showing differential ΔCt values of miR-92b-3p between non-poorly cohesive and poorly cohesive GC patients **(A)**, differential ΔCt values of let-7g-5p **(B)** and miR-146b-5p **(C)** between non-nerve infiltration and nerve infiltration gastric cancer patients, and differential ΔCt values of miR-146b-5p between T1 stage and T2 + T3 stage gastric cancer patients **(D)**.

efficacies than traditional biomarker CEA in early-stage GC. Meanwhile, synergistic effects were also found in combining CEA with the exosomal miRNA markers. These results provided evidence for exosomal miRNAs as a new non-invasive diagnostic option for early-stage GC.

Emerging evidence suggests that exosomal miRNAs can serve as potential biomarkers of GC, and are eligible for diagnosis, predicting recurrences, and providing prognostic information. Serum exosomal miR-10b-5p, miR-195-5p, miR-20a-3p, miR-296-5p were identified as novel potential biomarkers for detecting GC (Huang et al., 2017b). Low levels of plasma exosomal miRNA-101 and miR-23b have been confirmed to be related to poor prognosis in GC patients, and plasma exosomal miR-23b can be used as a minimally invasive predictive biomarker for the

recurrence of GC (Kumata et al., 2018). Additionally, exosomal miR-21 and miR-1225-5p derived from peritoneal lavage fluid could be novel biomarkers for peritoneal dissemination after curative resection of GC (Tokuhisa et al., 2015). Even though promising advances have been made in using exosomal miRNAs as diagnostic biomarkers for GC in recent years, the GCs involved in the previous studies are composed of both early and advanced stages (stage I to stage IV). Early diagnosis of GC remains a challenge and non-invasive evidence on diagnosis of early-stage GC is lacking. Hence, reliable biomarkers with high sensitivity and specificity need to be developed. To the best of our knowledge, this is the first study involved only early-stage patients (stage I and stage II) to analysis the early diagnostic value of serum exosomal miRNAs in GC.

MiR-92b-3p has been confirmed to play an important role in cancer proliferation, invasion, and migration. As an oncogenic miRNA, miR-92b-3p promotes carcinogenesis and metastasis by down-regulating F-box and WD-40 domain protein 7 (FBXW7/hCdc4) in colorectal cancer (Gong et al., 2018). In accordance with this evidence, we found that up-regulated miR-92b-3p was associated with early-stage GC. Besides, let-7g-5p has been investigated in melanoma, lung adenocarcinoma, ovarian cancer, glioblastoma, and renal cell carcinoma (RCC) (Gowrishankar et al., 2014; Petrillo et al., 2016). Overexpression of let-7g-5p significantly inhibits cell proliferation and migration, and could be a biomarker of post-operative recurrence and prognosis (Petrillo et al., 2016). Let-7g-5p was clustered into the let-7g family. Down-regulated let-7g has been associated with poor survival and lymph node metastasis in GC (Ueda et al., 2010). However, the role of let-7g-5p in GC had not been explored. This is the first study to confirm that aberrant expression of serum exosomal let-7g-5p was associated with early-stage GC. Consistent with our data, miR-146b-5p was observed to be up-regulated in colorectal cancer and renal cell carcinoma (Zhu et al., 2017), but frequently down-regulated in pancreatic cancer and glioblastoma (Li et al., 2013). Recent studies showed that deregulation of miR-146b-5p in GC tissue was associated with the tumor stage, and could be used as a prognosis biomarker (Zhou et al., 2013). In this study, we further verified the diagnostic role of serum exosomal miR-146b-5p in early-stage GC. Moreover, we found that serum exosomal miR-9-5p was up-regulated in early-stage GC. Likewise, up-regulation of miR-9-5p was observed in both breast cancer and lung cancer, which promoted tumor proliferation, invasion, and metastasis (Barbano et al., 2017). However, functional roles of miR-146b-5p and miR-9-5p in GC remain unknown.

It is of note that the molecular profile of tumors dynamically changes among patients as well as over time. In terms of miRNA profile, the expression signatures are also significantly varies as the medical condition of the specific donor differs. Among the biomarkers for early diagnosis of GC, the levels of miR-92b in plasma and miR-146b-5p in tissue have also been found significantly different between early- and late-stage GC (Chen et al., 2013; Zhou et al., 2015). Additionally, differential expression of let-7g-5p in tissue has been found between early- and late-stage clear cell renal carcinoma (Gowrishankar et al., 2014). However, the evidences for the differential expression of miR-9-5p and let-7g-5p between early- and late-stage GC are still lacking. These phenomena suggest the exosomal miRNA biomarkers for early diagnosis of cancer are not necessarily aberrantly expressed throughout the process of GC.

It has been reported that different expression levels of miRNAs could be frequently observed between circulating and tissue samples (Valadi et al., 2007; Huang et al., 2017a). In our study, based on TCGA data, we found that miR-92b-3p and miR-146b-5p were significantly upregulated in GC tissues, while let-7g-5p and miR-9-5p were significantly downregulated in GC tissues. We suspected that the discrepancies of the expression signatures between tissues and serum exosomes might be due to the following explanations. First, the absorption and degradation efficiency of circulating exosomal miRNAs and tissue miRNAs are quite different, and varieties in techniques and sequencing platform might also contribute to the discrepancies (Rabinowits et al., 2017). Second, compared with circulating exosomal miRNA reflecting the systematic disease status of GC, tissue miRNAs just presented the landscape of local changes. Currently, although significant advances have been achieved in exosome study, the understanding of the packaging and releasing mechanism of circulating exosomal miRNAs is still far from complete. Intensive work is still required to clarify the relationship of miRNAs between circulating exosomes and tissues.

In this study, combination of serum CEA and a panel of serum exosomal miRNAs (miR-92b-3p + let-7g-5p + miR-146b-5p + miR-9-5p) reached the highest diagnostic power for early stage GC, which could relieve a substantial proportion of patients from invasive biopsy. Nonetheless, this study had several potential limitations. First, although the results successfully showed four GC-related exosomal miRNAs, the number of early-stage GC patients enrolled in the study was limited. Further studies are necessary to evaluate these exosomal miRNA signatures in a larger cohort of early-stage GC patients. Second, the heterogeneous patient population could impact overall accuracy of the exosomal miRNAs, and our future study will compare the exosomal miRNAs with diverse tumor types. Third, we cannot elucidate the origin of the diagnostic miRNA markers. Whether they are derived from GC cells requires further investigation.

CONCLUSION

This study demonstrated the diagnostic values of serum exosomal miR-92b-3p, -146b-5p, -9-5p, and let-7g-5p in early-stage GC, and suggested the possibility of non-invasive circulating exosomal miRNAs as an alternative to conventional invasive approaches in detection of GC in the future. Moreover, large-scale validation is still required to confirm the potential applicability of these markers in GC diagnosis.

DATA AVAILABILITY STATEMENT

The raw sequencing data for this study can be found in the Gene Expression Omnibus database (accession number: GSE130654).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Harbin Medical University Cancer Hospital Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ST, YZ, and XH conceived and designed the study. ST and JC performed the experiments. YY and CL collected the samples and organized the clinical information. LW was responsible for the data processing and analysis of next-generation sequencing. ST performed statistical analysis and wrote the manuscript. YY, YZ, and XH reviewed and edited the manuscript. All the authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant numbers 81872427, 81672428, 81703000, and 81572528); Outstanding Youth Fund of Heilongjiang Province (grant number JC2018024); Sister Institution Network Fund (SINF) of MD Anderson (grant number 201402); and the Haiyan Fund Project of Harbin Medical University Cancer Hospital (grant number JJQN2019-02).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00237/full#supplementary-material

REFERENCES

Barbano, R., Pasculli, B., Rendina, M., Fontana, A., Fusilli, C., Copetti, M., et al. (2017). Stepwise analysis of MIR9 loci identifies miR-9-5p to be involved in Oestrogen regulated pathways in breast cancer patients. *Sci. Rep.* 7:45283. doi: 10.1038/srep45283

Batth, I. S., Mitra, A., Manier, S., Ghobrial, I. M., Menter, D., Kopetz, S., et al. (2017). Circulating tumor markers: harmonizing the yin and yang of CTCs and ctDNA for precision medicine. *Ann. Oncol.* 28, 468–477. doi: 10.1093/annonc/mdw619

Bhagirath, D., Yang, T. L., Bucay, N., Sekhon, K., Majid, S., Shahryari, V., et al. (2018). microRNA-1246 is an exosomal biomarker for aggressive prostate cancer. *Cancer Res.* 78, 1833–1844. doi: 10.1158/0008-5472.CAN-17-2069

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, W., Sun, K., Zheng, R., Zeng, H., Zhang, S., Xia, C., et al. (2018). Cancer incidence and mortality in China, 2014. *Chin. J. Cancer Res.* 30, 1–12. doi: 10.21147/j.issn.1000-9604.2018.01.01

Chen, Z., Saad, R., Jia, P., Peng, D., Zhu, S., Washington, M. K., et al. (2013). Gastric adenocarcinoma has a unique microRNA signature not present in esophageal adenocarcinoma. *Cancer* 119, 1985–1993. doi: 10.1002/cncr.28002

Chi, K. R. (2016). The tumour trail left in blood. *Nature* 532, 269–271. doi: 10.1038/532269a

Gong, L., Ren, M., Lv, Z., Yang, Y., and Wang, Z. (2018). miR-92b-3p promotes colorectal carcinoma cell proliferation, invasion, and migration by inhibiting FBXW7 *in vitro* and *in vivo*. *DNA Cell Biol.* 37, 501–511. doi: 10.1089/dna.2017.4080

Gowrishankar, B., Ibragimova, I., Zhou, Y., Slifker, M. J., Devarajan, K., Al-Saleem, T., et al. (2014). MicroRNA expression signatures of stage, grade, and progression in clear cell RCC. *Cancer Biol. Ther.* 15, 329–341. doi: 10.4161/cbt.27314

Hu, Y., Qi, C., Liu, X., Zhang, C., Gao, J., Wu, Y., et al. (2019). Malignant ascites-derived exosomes promote peritoneal tumor cell dissemination and reveal a distinct miRNA signature in advanced gastric cancer. *Cancer Lett.* 457, 142–150. doi: 10.1016/j.canlet.2019.04.034

Huang, X., Yuan, T., Liang, M., Du, M., Xia, S., Dittmar, R., et al. (2015). Exosomal miR-1290 and miR-375 as prognostic markers in castration-resistant prostate cancer. *Eur. Urol.* 67, 33–41. doi: 10.1016/j.eururo.2014.07.035

Huang, X., Yuan, T., Tschannen, M., Sun, Z., Jacob, H., Du, M., et al. (2013). Characterization of human plasma-derived exosomal RNAs by deep sequencing. *BMC Genomics* 14:319. doi: 10.1186/1471-2164-14-319

Huang, Z., Zhang, L., Zhu, D., Shan, X., Zhou, X., Qi, L. W., et al. (2017a). A novel serum microRNA signature to screen esophageal squamous cell carcinoma. *Cancer Med.* 6, 109–119. doi: 10.1002/cam4.973

Huang, Z., Zhu, D., Wu, L., He, M., Zhou, X., Zhang, L., et al. (2017b). Six serum-based miRNAs as potential diagnostic biomarkers for gastric cancer. *Cancer Epidemiol. Biomarkers Prev.* 26, 188–196. doi: 10.1158/1055-9965.EPI-16-0607

Jin, X., Chen, Y., Chen, H., Fei, S., Chen, D., Cai, X., et al. (2017). Evaluation of tumor-derived exosomal miRNA as potential diagnostic biomarkers for early-stage non-small cell lung cancer using next-generation sequencing. *Clin. Cancer Res.* 23, 5311–5319. doi: 10.1158/1078-0432.CCR-17-0577

Kosaka, N., Yoshioka, Y., Fujita, Y., and Ochiya, T. (2016). Versatile roles of extracellular vesicles in cancer. *J. Clin. Invest.* 126, 1163–1172. doi: 10.1172/JCI81130

Kumata, Y., Iinuma, H., Suzuki, Y., Tsukahara, D., Midorikawa, H., Igarashi, Y., et al. (2018). Exosomeencapsulated microRNA23b as a minimally invasive liquid biomarker for the prediction of recurrence and prognosis of gastric cancer patients in each tumor stage. *Oncol. Rep.* 40, 319–330. doi: 10.3892/or.2018.6418

Lai, X., Wang, M., McElyea, S. D., Sherman, S., House, M., and Korc, M. (2017). A microRNA signature in circulating exosomes is superior to exosomal glypican-1 levels for diagnosing pancreatic cancer. *Cancer Lett.* 393, 86–93. doi: 10.1016/j.canlet.2017.02.019

Li, Y., Wang, Y., Yu, L., Sun, C., Cheng, D., Yu, S., et al. (2013). miR-146b-5p inhibits glioma migration and invasion by targeting MMP16. *Cancer Lett.* 339, 260–269. doi: 10.1016/j.canlet.2013.06.018

Madhavan, B., Yue, S., Galli, U., Rana, S., Gross, W., Muller, M., et al. (2015). Combined evaluation of a panel of protein and miRNA serum-exosome biomarkers for pancreatic cancer diagnosis increases sensitivity and specificity. *Int. J. Cancer* 136, 2616–2627. doi: 10.1002/ijc.29324

Manterola, L., Guruceaga, E., Gallego Perez-Larraya, J., Gonzalez-Huarriz, M., Jauregui, P., Tejada, S., et al. (2014). A small noncoding RNA signature found in exosomes of GBM patient serum as a diagnostic tool. *Neuro Oncol.* 16, 520–527. doi: 10.1093/neuonc/not218

Ohno, S., Ishikawa, A., and Kuroda, M. (2013). Roles of exosomes and microvesicles in disease pathogenesis. *Adv. Drug Deliv. Rev.* 65, 398–401. doi: 10.1016/j.addr.2012.07.019

Petrillo, M., Zannoni, G. F., Beltrame, L., Martinelli, E., DiFeo, A., Paracchini, L., et al. (2016). Identification of high-grade serous ovarian cancer miRNA species associated with survival and drug response in patients receiving neoadjuvant chemotherapy: a retrospective longitudinal analysis using matched tumor biopsies. *Ann. Oncol.* 27, 625–634. doi: 10.1093/annonc/mdw007

Rabinowits, G., Bowden, M., Flores, L. M., Verselis, S., Vergara, V., Jo, V. Y., et al. (2017). Comparative analysis of MicroRNA expression among benign and malignant tongue tissue and plasma of patients with tongue cancer. *Front. Oncol.* 7:191. doi: 10.3389/fonc.2017.00191

Silver, N., Best, S., Jiang, J., and Thein, S. L. (2006). Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* 7:33. doi: 10.1186/1471-2199-7-33

Siravegna, G., Marsoni, S., Siena, S., and Bardelli, A. (2017). Integrating liquid biopsies into the management of cancer. *Nat. Rev. Clin. Oncol.* 14, 531–548. doi: 10.1038/nrclinonc.2017.14

Sugimachi, K., Matsumura, T., Hirata, H., Uchi, R., Ueda, M., Ueo, H., et al. (2015). Identification of a bona fide microRNA biomarker in serum exosomes that predicts hepatocellular carcinoma recurrence after liver transplantation. *Br. J. Cancer* 112, 532–538. doi: 10.1038/bjc.2014.621

Sun, J., Zhao, H., Lin, S., Bao, S., Zhang, Y., Su, J., et al. (2019). Integrative analysis from multi-centre studies identifies a function-derived personalized multi-gene signature of outcome in colorectal cancer. *J. Cell Mol. Med.* 23, 5270–5281. doi: 10.1111/jcmm.14403

Tokuhisa, M., Ichikawa, Y., Kosaka, N., Ochiya, T., Yashiro, M., Hirakawa, K., et al. (2015). Exosomal miRNAs from peritoneum lavage fluid as potential prognostic biomarkers of peritoneal metastasis in gastric cancer. *PLoS One* 10:e0130472. doi: 10.1371/journal.pone.0130472

Ueda, T., Volinia, S., Okumura, H., Shimizu, M., Taccioli, C., Rossi, S., et al. (2010). Relation between microRNA expression and progression and prognosis of gastric cancer: a microRNA expression analysis. *Lancet Oncol.* 11, 136–146. doi: 10.1016/S1470-2045(09)70343-2

Valadi, H., Ekstrom, K., Bossios, A., Sjostrand, M., Lee, J. J., and Lotvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat. Cell Biol.* 9, 654–659. doi: 10.1038/ncb1596

Wang, J., Yan, F., Zhao, Q., Zhan, F., Wang, R., Wang, L., et al. (2017). Circulating exosomal miR-125a-3p as a novel biomarker for early-stage colon cancer. *Sci. Rep.* 7:4150. doi: 10.1038/s41598-017-04386-1

Zhou, L., Zhao, X., Han, Y., Lu, Y., Shang, Y., Liu, C., et al. (2013). Regulation of UHRF1 by miR-146a/b modulates gastric cancer invasion and metastasis. *FASEB J.* 27, 4929–4939. doi: 10.1096/fj.13-233387

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, X., Zhu, W., Li, H., Wen, W., Cheng, W., Wang, F., et al. (2015). Diagnostic value of a plasma microRNA signature in gastric cancer: a microRNA expression analysis. *Sci. Rep.* 5:11251. doi: 10.1038/srep11251

Zhu, Y., Wu, G., Yan, W., Zhan, H., and Sun, P. (2017). miR-146b-5p regulates cell growth, invasion, and metabolism by targeting PDHB in colorectal cancer. *Am. J. Cancer Res.* 7, 1136–1150.

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 © 2020 Tang, Cheng, Yao, Lou, Wang, Huang 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 20 March 2020
doi: 10.3389/fbioe.2020.00201

A Mendelian Randomization Analysis to Expose the Causal Effect of IL-18 on Osteoporosis Based on Genome-Wide Association Study Data

Ni Kou[1†], Wenyang Zhou[2†], Yuzhu He[1], Xiaoxia Ying[1], Songling Chai[1], Tao Fei[1], Wenqi Fu[1], Jiaqian Huang[1] and Huiying Liu[1]*

[1] Department of Oral Prosthodontics, School of Stomatology, Dalian Medical University, Dalian, China, [2] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Lei Deng,
Central South University, China
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Huiying Liu
lhy04512000@dmu.edu.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 29 January 2020
Accepted: 28 February 2020
Published: 20 March 2020

Citation:
Kou N, Zhou W, He Y, Ying X, Chai S, Fei T, Fu W, Huang J and Liu H (2020) A Mendelian Randomization Analysis to Expose the Causal Effect of IL-18 on Osteoporosis Based on Genome-Wide Association Study Data. Front. Bioeng. Biotechnol. 8:201. doi: 10.3389/fbioe.2020.00201

Accumulating evidence showed that Interleukin (IL) level is associated with Osteoporosis. Whereas, most of these associations are based on observational studies. Thus, their causality was still unclear. Mendelian randomization (MR) is a widely used statistical framework that uses genetic instrumental variables (IVs) to explore the causality of intermediate phenotype with disease. To classify their causality, we conducted a MR analysis to investigate the effect of IL-18 level on the risk of Osteoporosis. First, based on summarized genome-wide association study (GWAS) data, 8 independent IL-18 SNPs reaching genome-wide significance were deemed as IVs. Next, Simple median method was used to calculate the pooled odds ratio (OR) of these 8 SNPs for the assessment of IL-8 on the risk of Osteoporosis. Then, MR-Egger regression was utilized to detect potential bias due to the horizontal pleiotropy of these IVs. As a result of simple median method, we get the SE (−0.001; 95% CI −0.002 to 0; $P = 0.042$), which means low IL-18 level could increases the risk of the development of Osteoporosis. The low intercept (0; 95% CI −0.001 to 0; $P = 0.59$) shows there is no bias due to the horizontal pleiotropy of the IVs.

Keywords: genome-wide association studies, Interleukin-18, Osteoporosis, casual effect, Mendelian randomization

INTRODUCTION

Osteoporosis is a chronic disease with a variety of causals to bone mineral density and bone loss of quality (Sambrook, 2006). Since the bone weakening in Osteoporosis patients, it increases the risk of a broken bone and other diseases among the elderly (Tu et al., 2018). Bone is a type of living tissue, which is constantly being broken down and replaced. When the creation of new bone doesn't keep up with the loss of old one, Osteoporosis occurs (Seo et al., 2018; Papaleontiou et al., 2019; Wang et al., 2019). Osteoporosis affects all the countries including men and women. For example, about 90,000,000 Osteoporosis patients in china, which covers the 7.1% of the total population. Since lots of complications affected by Osteoporosis and its incurability, it is very important for preventing the Osteoporosis. Whereas, there are no typical symptoms in the early stages of bone loss according to the current knowledge (Hennemann, 2002). This raises the difficulties for prevention. To solve this problem, it is urgent to expose the causal clinical phenotypes of Osteoporosis.

Interleukin (IL) is a class of cytokines, which is produced by a variety of cells and also functions on a variety of cells (Kato and Perl, 2018; Zhang et al., 2018). Current, about forty types of IL was discoveries in human body. IL-18 is one of them, which locates at 11q22.2-22.3. IL-18 is powerful inflammatory cytokines, the most characteristic feature of which is the regulation of cellular proliferation and differentiation (Weiss et al., 2018; Youssef et al., 2018; Prencipe et al., 2019; Valero et al., 2019). Recent studies show that IL-18 plays important roles in immune regulation, resistance to infection and anti-tumor. Furthermore, IL-18 has been identified as the causal of multiple chronic diseases, such as type 2 diabetes (Zou et al., 2018).

The relationship between IL and Osteoporosis has been investigated in observational studies decades of years. Early in 1993, Lewis et al. investigated a transgenic mice with disorder in bone homeostasis that inappropriately express the cytokine IL-4 (Lewis et al., 1993). And then they observed that Osteoporosis was associated with the IL-4. In 2005, Rusinska et al. evaluated the relationship between multiple ILs and the etiopathogenesis of idiopathic osteoporosis in children (Rusinska and Chlebnasokół, 2005). In 2010, Edwards et al. exposed the relationship between IL-6 and rheumatoid arthritis-associated osteoporosis (Edwards and Williams, 2010).

Although current advantages on investigating the relationship between IL-18 and Osteoporosis, it is still not clear whether IL-18 is the consequent or causal effect of the Osteoporosis. This is the common problem for many associations between phenotypes and diseases. With the development of Genome-Wide Association Studies (GWAS) and identification of molecular mechanism in recent years (Li et al., 2015; Zhou et al., 2017a,b; Tang et al., 2018; Tan et al., 2019), Mendelian randomization (MR) analysis is widely used to expose the causal effect of phenotypes on the development of diseases. For example, body mass index and C-reaction protein are identified as the causal effect of the development of type 2 diabetes (Cheng et al., 2018b, 2019c). Meanwhile, some negative associations are also exposed, such as associations between infant length and type 2 diabetes (Zhuang et al., 2019a). As other statistical analysis and machine learning methods (Du et al., 2018; Liao et al., 2018; Wang B. et al., 2018; Wang L. et al., 2018; Cheng et al., 2019b; Han et al., 2019; Lv et al., 2019; Yang et al., 2019; Zeng et al., 2019; Zou and Ma, 2019; Zhao et al., 2020), MR is an instrumental variable (IV) based method for inferring associations between phenotypes and diseases. As shown in **Figure 1**, genetic variants as SNPs are often used as IVs. This is because that SNPs are genetic characterize and occurred before phenotypes and diseases, which can avoid reverse causality. Here Z (e.g., SNPs) represents IVs, X is phenotype (e.g., IL-18), and Y is the disease (Osteoporosis). To conduct MR analysis, the IVs should meet two assumptions. One is that SNPs should be robustly associated with phenotype (IL-18), and the other is that SNPs can influence the disease only through the phenotype.

To explore the causal effect of IL-18 on the development of Osteoporosis, we conducted a MR analysis in this study. First, we defined a framework for processing data to establish IVs for MR analysis. Next, simple median method was used for calculating the pooled result based on IVs. Then, to avoid bias and analysis the limitation of our method, MR Egger analysis and leave-one-out validation was conducted.

MATERIALS AND METHODS

Summary-level data of GWAS dataset was the fundamental for MR analysis. IVs should be extracted from IL-8 related GWAS dataset. And the further analysis need Osteoporosis related GWAS dataset. To meet the MR assumptions and reduce the bias, the summarized GWAS data was processed. Subsequently, MR analysis involving simple median method, leave-one-out validation, MR-Egger analysis was conducted to comprehensively assess the causal effect of IL-8 on the risk of the development of Osteoporosis.

Summarized GWAS Data for IL-18

In 2014, Matteini et al. (2014) performed a genome-wide association study (GWAS) on Cardiovascular Health Study and InCHIANTI cohorts. Totally, it contains 1200 InCHIANTI cohorts and 3200 CHS population. After genotyping, they used GWAPower software to assess the difference in power of the combined InCHIANTI-CHS meta-analysis compared to single study analyses. Then, significant SNPs of IL-18 were identified and used for analyzing causal relationship of IL-18 on the risk of Type 2 Diabetes (Zhuang et al., 2019a). Here, we extracted SNPs, effect allele (EA), allele frequencies, beta coefficients, and standard errors (SEs) as summarized data.

Summarized GWAS Data for Osteoporosis

In 2018, Bycroft et al. published their prospective cohort study (Bycroft et al., 2018), which contains approximately 500,000 individuals from across the United Kingdom, aged between 40 and 69 at recruitment. It also provided detailed description and summarized data of deep genetic and phenotypic data. We downloaded summarized GWAS data of Osteoporosis from UK Biobank, which involves susceptibility loci together with other SNPs, beta coefficient, EA, SEs and their *P*-values, and etc. Totally, it contains 933 cases and 360,261 controls.

Data Processing

We process the summarized GWAS data for constructing IVs of MR analysis. Here IVs are genetic variants. Each SNP should be significantly associated with IL-18, and should be not associated with Osteoporosis. Thus, we extracted SNPs significant associated with IL-18, and then removed these SNPs associated with Osteoporosis. We defined $P < 5^{*}10^{-8}$ as significant associated SNPs of IL-18, and we defined *P*-value more than 0.05 as not associated SNPs of Osteoporosis. To avoid over-precise estimates due to genetic pleiotropy, we should remove these SNPs with potential linkage disequilibrium (LD) relationships. The analogous method has been used in the MR analysis of causal effect of phenotype on T2DM (Cheng et al., 2019c; Zhuang et al., 2019b). To remove SNPs with LDs, we ranked significant SNPs of IL-18 based on *P*-values. For each SNP, we removed those SNPs in LD with it (R^2 threshold of 0.001) or within 10,000 kb physical distance based on a reference dataset

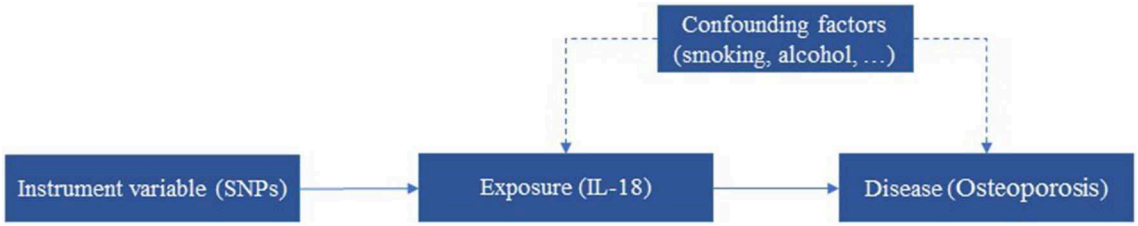


FIGURE 1 | MR analysis using SNPs as instrumental variables for estimating the influence of IL-18 on the risk of Osteoporosis.

(Devuyst, 2015). This process was iterated for each of significant SNPs of IL-18.

MR Analysis

MR analysis contains simple median method, leave-one-out validation, MR-Egger regression analysis (Bowden et al., 2015). Simple median method was used for assessing the pooled influence of IL-18 on the risk of Osteoporosis. Leave-one-out validation was performed for assessing sensitivity of each of IVs. Egger regression analysis was used to evaluate pleiotropy bias of IVs.

- Simple median method

Simple median method was described in the previous study (Burgess et al., 2016) for evaluating the influence of clinical phenotype on the risk of disease, which is defined as following equation.

$$beta_{XY} = beta_{ZY}/beta_{ZX}, \quad (1)$$

where *X, Y,* and *Z* are IL-18, Osteoporosis, and IVs, respectively, Wald ratio ($beta_{XY}$) of IL-18 to Osteoporosis through specified IV, $beta_{ZY}$ is the per-allele *log(OR)* of Osteoporosis from summarized GWAS data of Osteoporosis. $beta_{ZX}$ is the per-allele *log(OR)* of IL-18 from summarized GWAS data of IL-18. Then, we calculated SE of association between IL-18 and Osteoporosis for each Wald ratio, which is defined as Equation 2.

$$SE_{XY} = SE_{ZY}/SE_{ZX}, \quad (2)$$

where SE_{ZY} and SE_{ZX} represent the *SE* of the IV-Osteoporosis and IV-IL-18 associations from corresponding summarized GWAS data, respectively. Next, we calculated 95% confidence intervals (CIs) from the *SE* of each Wald ratio. To get the pooled influence of these IVs, simple median method was used as meta-analysis for estimating comprehensive influence of IVs.

- Leave-one-out validation

We used leave-one-out validation for evaluating the sensitivity of each of IVs as following. To assess the influence of a SNP of IVs to the pooled result, we remove this SNP from IVs to get the result using simple median method. Thus, the corresponding result is obtained without considering this SNP. The fluctuation of the pooled results before and after removing the SNP could reflect the sensitivity of this SNP. This process was iterated on each of these IVs to get the influence for each of IVs.

- MR-Egger analysis

We conducted a MR-Egger regression analysis about asymmetry test to measure bias based on potential pleiotropic effect of IVs (Bowden et al., 2015). The MR-Egger regression is source from Egger regression, which is designed for detecting bias due to small study and pleiotropy in meta-analysis. Here, MR-Egger used intercept as an estimated value for evaluating the average pleiotropic effect of IVs. For example, the larger or smaller an intercept, the more of pleiotropy effect should be. All statistical tests for this study were undertaken using the R Package of MRBase (Hemani et al., 2018).

RESULTS

Genetic Variants as IVs

Totally, 8 SNPs were extracted as significant associated SNPs of IL-18. Those SNPs were not associated with Osteoporosis and have no LD associations. As a result, those 8 significant SNPs of IL-18 were eventually selected as IVs for the MR analysis, which were shown in the **Table 1**. Each line of the table documents 10 items involving the SNP, EA, chromosome position, beta coefficients and SE of the SNP on the risk of IL-18 and Osteoporosis, and so on.

The Causal Effect of IL-18 on the Risk of Osteoporosis

After using 8 individual SNPs as IVs for MR analysis based on two summary-level GWAS data, we used simple median method for pooled analysis. **Figure 2** shows that there is no evidence of heterogeneity between variants of the summarized data. As a result, we get the SE (-0.001; 95% CI -0.002 to 0; $P = 0.042$), which means low IL-18 level could increase the risk of the development of Osteoporosis.

Sensitivity Analysis for Individual SNPs

Figure 3 shows estimate result of the leave-one-out analysis. After removing rs6760105, rs6748621, rs7577696, or rs2250417 from 8 IVs, the estimate value shows small fluctuation. And the result is consistent with using all the IVs. This means that this four SNPs activate weak influence to the

TABLE 1 | Associations of genetic variants with IL-18 and Osteoporosis.

SNP	EA (frequency)	Gene ID/Symbol	chr	Position	IL-18_beta	IL-18_SE	Beta	se	*p*-value
rs7577696	G (0.49381)	NA	2	32278782	0.08	0.01	8.71E-05	6.15E-05	1.57E-01
rs6760105	G (0.491214)	6683 (SPAST)	2	32307386	0.06	0.01	8.63E-05	6.15E-05	1.61E-01
rs6748621	C (0.495208)	84661 (DPY30)	2	32262201	0.08	0.01	1.01E-04	6.17E-05	1.03E-01
rs2300702	C (0.545527)	6716 (SRD5A2)	2	31788018	0.07	0.01	−9.89E-05	6.11E-05	1.06E-01
rs2268797	C (0.552716)	6716 (SRD5A2)	2	31783752	0.07	0.01	−9.53E-05	6.11E-05	1.18E-01
rs2250417	T (0.304113)	83875 (BCO2)	2	32412832	0.1	0.01	−3.32E-05	6.01E-05	5.81E-01
rs212745	C (0.480232)	55676 (SLC30A6)	2	32457537	0.07	0.01	−9.38E-05	6.16E-05	1.28E-01
rs212713	C (0.494409)	58484 (NLRC4)	11	112085316	0.06	0.01	−9.82E-05	6.00E-05	1.02E-01

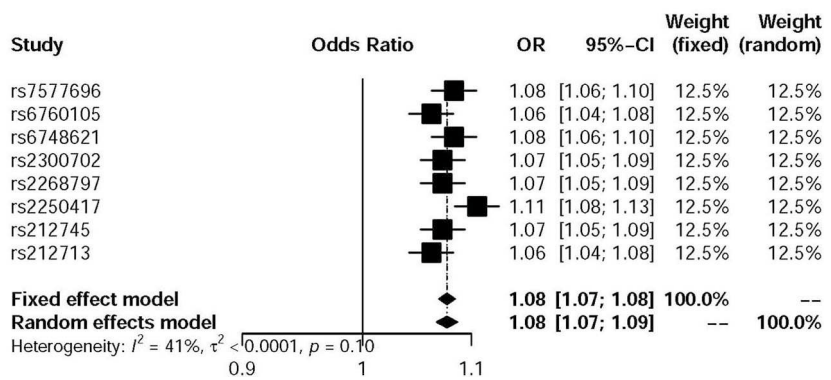


FIGURE 2 | Forest plot of Wald ratios and 95% CIs of IVs.

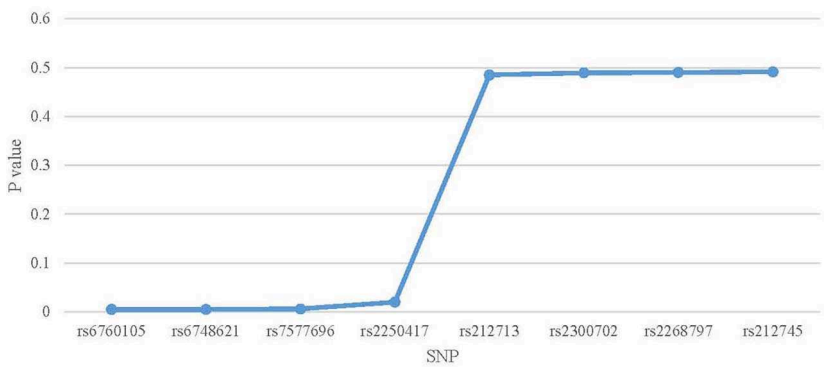


FIGURE 3 | Scatter plot of the *P*-values in leave-one-out analysis.

estimate result. In comparison, after removing rs212713, rs2300702, rs2268797, or rs212745, estimate value shows large fluctuation. This means that this four SNPs activate strong influence to the estimate result. The detailed information about leave-one-out validation result is shown in **Table 2**.

Pleiotropic Effect Analysis for IVs

Figure 4 shows the effect estimate based on MR-Egger regression. The low intercept (0; 95% CI −0.001 to 0; $P = 0.59$) shows there is no bias due to the horizontal pleiotropy of the IVs.

DISCUSSION

Till now, it is not clear that the level of IL-18 is the causal or consequence of the development of Osteoporosis. To expose the relationship, we conducted an MR analysis based on two summarized GWAS datasets in this study. As a result of simple median method, we get the SE (−0.001; 95% CI −0.002 to 0; $P = 0.042$). This means that low IL-18 level could increase the risk of the development of Osteoporosis.

Observational study is a widely used way to reveal the associations between phenotypes and diseases. Whereas, it couldn't be used for exposing the causal effect. With the increase

of GWAS data and abundance of molecular characterize (Cheng, 2019; Cheng et al., 2019a; Dao et al., 2020; Zhang et al., 2020), more and more researchers choose MR analysis for this purpose. MR analysis is an IV-based framework, which requires summarized GWAS data. In recent years, MR analysis has helped us to identify lots of causal effects, such as body mass index and C-reactive protein increase the risk of type 2 diabetes (Cheng et al., 2019c; Zhuang et al., 2019b). Here, the number of the case and control for GWAS data is very important for the estimation. In the previous study, the summarized data of IL-18 GWAS data has been applied in exposing the relationship between IL-18 and T2DM (Zhuang et al., 2019a). And the number of case and control is over 5,000. In addition, the UK biobank provided 933 Osteoporosis and 360,261 controls. The number of the samples for IL-18 and Osteoporosis GWAS data is the baseline for our MR analysis.

To make the MR analysis reliably, the summarized GWAS data was processed strictly to choose suitable IVs. First, significant associated SNPs of IL-18 ($P < 5^{*}10^{-8}$) was extracted. Then, these Osteoporosis ($P < 5^{*}10^{-2}$) associated SNPs were removed. This process is to make sure the IVs meet the requirement of MR's assumption. In addition, to reduce the bias of IVs due to the pleiotropic effect, SNPs in IVs with LDs were removed. In actually, the comprehensive effect of IVs are used for estimating the causal effect. The SNPs with LDs should be deemed as a SNP to reduce the bias based on the replication. To detect the potential bias due to the horizontal pleiotropy of IVs, MR egger method was conducted. As a result, we got the low intercept (0; 95% CI −0.001 to 0; $P = 0.59$), which shows there is no bias due to the horizontal pleiotropy of the IVs. All the data process of IVs is to reduce the bias, and make the result more reliably.

TABLE 2 | Results based on leave-one-out validation.

SNP (leave out)	Wald ratio	95%CI lower	95%CI upper	*p*-value
rs6760105	−0.001	−0.002	0	0.005
rs6748621	−0.001	−0.002	0	0.005
rs7577696	−0.001	−0.002	0	0.006
rs2250417	−0.001	−0.002	0	0.02
rs212713	0	−0.001	0.001	0.485
rs2300702	0	−0.001	0.001	0.489
rs2268797	0	−0.001	0.001	0.49
rs212745	0	−0.001	0.001	0.491

Although there is no direct associations between IVs and Osteoporosis, we also validated their potential linkages. We downloaded associated genes of Osteoporosis from a widely used functional annotation database OAHG (Cheng et al., 2016, 2018a), and compared them with genes that IVs located at (**Table 1**). Up to 440 genes of Osteoporosis were documented in OAHG, and no intersection between them and genes of IVs. This validate further that no potential bias of IVs based on current knowledge.

Although MR analysis show the advantages in distinguishing causal effect from general associations. It also has the limitations. With the incensement of samples, the summarized data of phenotype and disease should be a little alteration. This could lead to the changes in fluctuations about summarized data of GWAS data and even in IVs. This would influence the estimated result. Thus, to avoid this problem, the huge number of samples is very important. In this study, we do a leave-one-out analysis to judge which SNPs could influence our results. As a result, we find out SNPs rs6760105, rs6748621, rs7577696, or rs2250417 shows a little capability in influencing the result. In comparison,

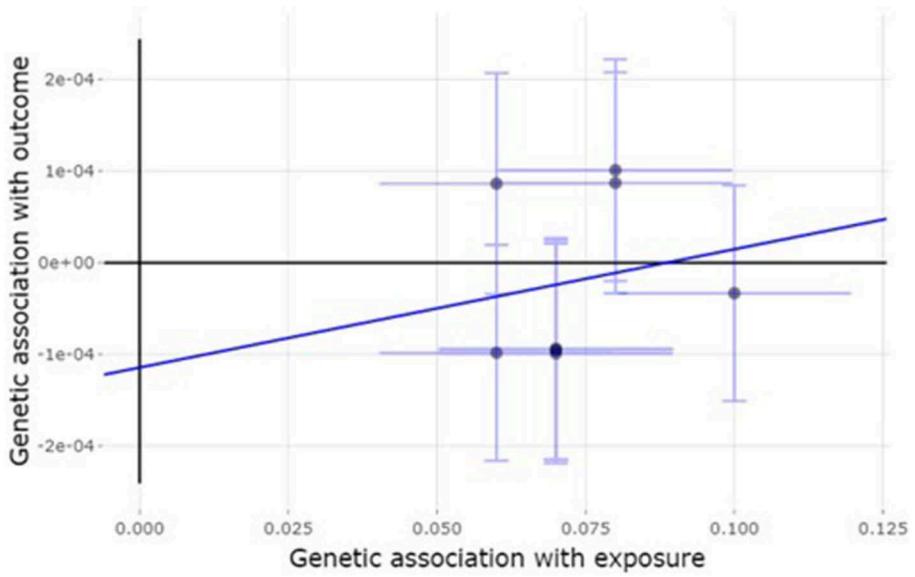


FIGURE 4 | The estimate of horizontal pleiotropy based on MR-Egger analysis.

SNPs rs212713, rs2300702, rs2268797, or rs212745 could largely influence the result, all of which are located at genes SRD5A2, SLC30A6, and NLRC4. With the increase of the GWAS data, the impact of these SNPs could be decreased.

In addition to simple median method, IVW method is another frequently used method for MR analysis. As a result of IVW method, we get SE (0; 95% CI −0.001 to 0.001; $P = 0.625$), which is inconsistent with the result of simple median method. Currently, most of current MR analysis based on different methods often show inconsistent results. Each method has its advantages and limitations. In general, the causal effect could be validated when the relationship is supported by one of these methods. To provide more reliable validation, it needs randomized controlled trial (RCT). Whereas, it is very hard to conduct RCT. Thus, most of current causal effect between clinical phenotypes and diseases are validated based on MR analysis. Although current success in applying MR analysis, it couldn't substitute for RCT.

In summary, 8 SNPs were used as IVs for estimating the causal effect of IL-18 on the development of Osteoporosis. Results show that low IL-18 level could increase the risk of the development of Osteoporosis based on simple median method. In considering the limitation of MR method and current samples of GWAS data, further experiment for the conclusion is expected.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=20002.

AUTHOR CONTRIBUTIONS

HL conceived and designed the experiments. NK and WZ analyzed the data. NK, WZ, and HL wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 61871068).

REFERENCES

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44, 512–525. doi: 10.1093/ije/dyv080

Burgess, S., Bowden, J., Dudbridge, F., and Thompson, S. G. (2016). Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *Methodology. arXiv.org [Preprint]. arXiv*:1606.03729.

Bycroft, C., Freeman, C., Petkova, D., Elliott, L. T., Sharp, K., Motyer, A., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/1566523219041910221133307

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). 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., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019a). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucl. Acids Res.* 48, D554–D560. doi: 10.1093/nar/gkz843

Cheng, L., Sun, J., Xu, W. Y., Dong, L. X., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6:34820. doi: 10.1038/srep34820

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019b). Computational methods for identifying similar diseases. *Mol. Ther. Nucl. Acids* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Cheng, L., Zhuang, H., Ju, H., Yang, S., Han, J., Tan, R., et al. (2019c). Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a Mendelian randomization study. *Front Genet* 10:94. doi: 10.3389/fgene.2019.00094

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018b). Exposing the causal effect of C-reactive protein on the risk of type 2 diabetes mellitus: a Mendelian randomization study. *Front. Genet.* 9:657. doi: 10.3389/fgene.2018.00657

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.* doi: 10.1093/bib/bbaa017. [Epub ahead of print].

Devuyst, O. (2015). The 1000 Genomes project: welcome to a new world. *Perit. Dial. Int.* 35, 676–677. doi: 10.3747/pdi.2015.00261

Du, X. Q., Li, X. R., Li, W., Yan, Y. T., and Zhang, Y. P. (2018). Identification and analysis of cancer diagnosis using probabilistic classification vector machines with feature selection. *Curr. Bioinform.* 13, 625–632. doi: 10.2174/1574893612666170405125637

Edwards, C. J., and Williams, E. (2010). The role of interleukin-6 in rheumatoid arthritis-associated osteoporosis. *Osteoporos. Int.* 21, 1287–1293. doi: 10.1007/s00198-010-1192-7

Han, J., Han, X., Kong, Q., and Cheng, L. (2019). psSubpathway: a software package for flexible identification of phenotype-specific subpathways in cancer progression. *Bioinformatics*. doi: 10.1093/bioinformatics/btz894. [Epub ahead of print].

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife* 7:e34408. doi: 10.7554/eLife.34408

Hennemann, A. (2002). Osteoporosis: prevention, diagnosis and therapy. *Medizinische Monatsschrift Für Pharmazeuten* 25, 164–167.

Kato, H., and Perl, A. (2018). Blockade of treg cell differentiation and function by the interleukin-21–mechanistic target of rapamycin axis via suppression of autophagy in patients with systemic lupus erythematosus. *Arthritis Rheumatol.* 70, 427–438. doi: 10.1002/art.40380

Lewis, D. B., Liggitt, H. D., Effmann, E. L., Motley, S. T., Teitelbaum, S. L., Jepsen, K. J., et al. (1993). Osteoporosis induced in mice by overproduction of interleukin 4. *Proc. Natl. Acad. Sci. U.S.A.* 90, 11618–11622. doi: 10.1073/pnas.90.24.11618

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, 143–155. doi: 10.1093/bfgp/elu036

Liao, Z. J., Li, D. P., Wang, X. R., Li, L. S., and Zou, Q. (2018). Cancer diagnosis through IsomiR expression with machine learning method. *Curr. Bioinform.* 13, 57–63. doi: 10.2174/1574893611666160609081155

Lv, H., Zhang, Z. M., Li, S. H., Tan, J. X., Chen, W., and Lin, H. (2019). Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform*. doi: 10.1093/bib/bbz048. [Epub ahead of print].

Matteini, A. M., Li, J., Lange, E. M., Tanaka, T., Lange, L. A., Tracy, R. P., et al. (2014). Novel gene variants predict serum levels of the cytokines IL-18 and IL-1ra in older adults. *Cytokine* 65, 10–16. doi: 10.1016/j.cyto.2013.10.002

Papaleontiou, M., Banerjee, M., Reyes-Gastelum, D., Hawley, S. T., and Haymart, M. R. (2019). Risk of osteoporosis and fractures in patients with thyroid cancer: a case–control study in U.S. Veterans. *Oncologist* 24, 1166–1173. doi: 10.1634/theoncologist.2019-0234

Prencipe, G., Bracaglia, C., and Benedetti, F. D. (2019). Interleukin-18 in pediatric rheumatic diseases. *Curr. Opin. Rheumatol.* 31, 421–427. doi: 10.1097/BOR.0000000000000634

Rusinska, A., and Chlebnasokół, D. (2005). Evaluation of interleukin-1 and -6 in the etiopathogenesis of idiopathic osteoporosis and osteopenia in children. *Arch Immunol Ther Exp* 53, 257–265.

Sambrook, P. N. (2006). Osteoporosis. *Lancet* 367, 2010–2018. doi: 10.1016/S0140-6736(06)68891-0

Seo, G. H., Kang, H. Y., and Choe, E. K. (2018). Osteoporosis and fracture after gastrectomy for stomach cancer: a nationwide claims study. *Medicine* 97:e0532. doi: 10.1097/MD.0000000000010532

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

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

Tu, K. N., Lie, J. D., C.,Wan, K. V., Cameron, M., and Hyun, D. (2018). Osteoporosis: a review of treatment options. *P T* 43, 92–104.

Valero, N., Mosquera, J., Torres, M., Duran, A., and Veliz, T. (2019). Increased serum ferritin and interleukin-18 levels in children with dengue. *Braz. J. Microbiol.* 50, 649–656. doi: 10.1007/s42770-019-00105-2

Wang, B., Lu, K., Zheng, X., Su, B. Y., Zhou, Y. M., Chen, P., et al. (2018). Early stage identification of Alzheimer's disease using a two-stage ensemble classifier. *Curr. Bioinform.* 13, 529–535. doi: 10.2174/1574893613666180328093114

Wang, J., Chen, L., Zhang, Y., Li, C.-G., and Shu, B. (2019). Association between serum vitamin B 6 concentration and risk of osteoporosis in the middle-aged and older people in China: a cross-sectional study. *BMJ Open* 9:e028129. doi: 10.1136/bmjopen-2018-028129

Wang, L., Ping, P. Y., Kuang, L. N., Ye, S. T., Lqbal, F. M. B., and Pei, T. R. (2018). A novel approach based on Bipartite network to predict human microbe-disease associations. *Curr. Bioinform.* 13, 141–148. doi: 10.2174/1574893612666170911143601

Weiss, E. S., Girard-Guyonvarc'h, C., Holzinger, D., de Jesus, A. A., Tariq, Z., Picarsic, J., et al. (2018). Interleukin-18 diagnostically distinguishes and pathogenically promotes human and murine macrophage activation syndrome. *Blood* 131, 1442–1455. doi: 10.1182/blood-2017-12-820852

Yang, H., Yang, W., Dao, F. Y., Lv, H., Ding, H., Chen, W., et al. (2019). A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief. Bioinform.* bbz123. doi: 10.1093/bib/bbz123. [Epub ahead of print].

Youssef, D. M., Al-Atif, A. M. A., El-Khateeb, S. S. H., and Elshal, A. S. (2018). Evaluation of interleukin-18 in children with steroid-sensitive nephrotic syndrome before and after using levamisole. *Saudi J. Kidney Dis. Transpl.* 29, 591–597. doi: 10.4103/1319-2442.235173

Zeng, X. X., Wang, W., Deng, G. S., Bing, J. X., and Zou, Q. (2019). Prediction of potential disease-associated microRNAs by using neural networks. *Mol. Ther. Nucl. Acids* 16, 566–575. doi: 10.1016/j.omtn.2019.04.010

Zhang, J., Li, Z., Liu, L., Wang, Q., Li, S., Chen, D., et al. (2018). Long noncoding RNA TSLNC8 is a tumor suppressor that inactivates the interleukin-6/STAT3 signaling pathway. *Hepatology* 67, 171–187. doi: 10.1002/hep.29405

Zhang, Z. Y., Yang, Y. H., Ding, H., Wang, D., Chen, W., and Lin, H. (2020). Design powerful predictor for mRNA subcellular location prediction in *Homo sapiens*. *Brief. Bioinform.* doi: 10.1093/bib/bbz177. [Epub ahead of print].

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 Bioinform.* 21:43. doi: 10.1186/s12859-020-3388-y

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2017b). An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. *Mol. Neurobiol.* 55, 3684–3697. doi: 10.1007/s12035-017-0572-9

Zhou, M., Zhao, H. Q., Xu, W. Y., Bao, S. Q., Cheng, L., and Sun, J. (2017a). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 2017:16. doi: 10.1186/s12943-017-0580-4

Zhuang, H., Han, J., Cheng, L., and Liu, S. (2019a). A positive causal influence of IL-18 levels on the risk of T2DM: a mendelian randomization study. *Front. Genet.* 10:295. doi: 10.3389/fgene.2019.00295

Zhuang, H., Zhang, Y., Yang, S., Cheng, L., and Liu, S. L. (2019b). A Mendelian randomization study on infant length and type 2 diabetes mellitus risk. *Curr Gene Ther* 19, 224–231. doi: 10.2174/1566523219666190925115535

Zou, Q., and Ma, Q. (2019). The application of machine learning to disease diagnosis and treatment. *Math. Biosci.* 320:108305. doi: 10.1016/j.mbs.2019.108305

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* 9:515. doi: 10.3389/fgene.2018.00515

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 © 2020 Kou, Zhou, He, Ying, Chai, Fei, Fu, Huang 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: 25 March 2020
doi: 10.3389/fcell.2020.00194

Comprehensive Analysis of the Genetic and Epigenetic Mechanisms of Osteoporosis and Bone Mineral Density

*Hui Dong[1,2†], Wenyang Zhou[3†], Pingping Wang[3†], Enjun Zuo[1], Xiaoxia Ying[1], Songling Chai[1], Tao Fei[1], Laidi Jin[1], Chen Chen[1], Guowu Ma[1] and Huiying Liu[1]**

[1] Department of Oral Prosthodontics, School of Stomatology, Dalian Medical University, Dalian, China, [2] Department of Stomatology, The First Affiliated Hospital of Dalian Medical University, Dalian, China, [3] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China

Osteoporosis is a skeletal disorder characterized by a systemic impairment of bone mineral density (BMD). Genome-wide association studies (GWAS) have identified hundreds of susceptibility loci for osteoporosis and BMD. However, the vast majority of susceptibility loci are located in non-coding regions of the genome and provide limited information about the genetic mechanisms of osteoporosis. Herein we performed a comprehensive functional analysis to investigate the genetic and epigenetic mechanisms of osteoporosis and BMD. BMD and osteoporosis are found to share many common susceptibility loci, and the corresponding susceptibility genes are significantly enriched in bone-related biological pathways. The regulatory element enrichment analysis indicated that BMD and osteoporosis susceptibility loci are significantly enriched in 5′UTR and DNase I hypersensitive sites (DHSs) of peripheral blood immune cells. By integrating GWAS and expression Quantitative Trait Locus (eQTL) data, we found that 15 protein-coding genes are regulated by the osteoporosis and BMD susceptibility loci. Our analysis provides new clues for a better understanding of the pathogenic mechanisms and offers potential therapeutic targets for osteoporosis.

Keywords: genome-wide association study, osteoporosis, bone mineral density, summary data-based Mendelian randomization, functional element enrichment analysis

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Liang Cheng,
Harbin Medical University, China

***Correspondence:**
Huiying Liu
lhy04512000@dmu.edu.com

† These authors have contributed equally to this work

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

Received: *04 February 2020*
Accepted: *09 March 2020*
Published: *25 March 2020*

Citation:
Dong H, Zhou W, Wang P, Zuo E, Ying X, Chai S, Fei T, Jin L, Chen C, Ma G and Liu H (2020) Comprehensive Analysis of the Genetic and Epigenetic Mechanisms of Osteoporosis and Bone Mineral Density. Front. Cell Dev. Biol. 8:194. doi: 10.3389/fcell.2020.00194

INTRODUCTION

Osteoporosis is a systemic skeletal disease characterized by a significant decrease in BMD and microarchitectural deterioration of bone tissue (Rachner et al., 2011). The decline in bone mass and prevalence of osteoporosis increase with age, especially in postmenopausal women (Cummings and Melton, 2002). Researchers estimate there are more than 200 million individuals with osteoporosis worldwide, and the fracture risk of patients with osteoporosis is as high as 40% (Rachner et al., 2011; Al-Barghouthi and Farber, 2019). What's worse, the number of patients with osteoporosis is expected to steadily increase in the near future, as the effects of an aging global population (Aggarwal and Masuda, 2018).

Osteoporosis, a typical of complex polygenic disease, is considered to be the consequence of the genetic interaction of multiple gene mutations (Saad, 2019). Previous studies based on twin and

family data have estimated that both the osteoporosis and BMD show high heritability (h^2 = 0.5–0.8) (Ralston and Uitterlinden, 2010; Al-Barghouthi and Farber, 2019). Clinically, BMD is a strong relevant marker of osteoporosis, as well as a key indicator for its diagnosis and treatment (Kemp et al., 2017). Therefore, a comprehensive understanding of the genetic factors underlying both osteoporosis and BMD is highly necessary to develop effective therapies for osteoporosis.

As early as in 1994, a candidate gene study found that several common allelic variants in vitamin D receptor gene are associated with bone mineral density (BMD) (Morrison et al., 1994). In recent years, with the development of microarray and next-generation sequencing technology, genome-wide association studies (GWAS) have been considered as powerful tools to investigate the genetic architecture of complex diseases (Liu G. et al., 2018). Especially since 2007, the GWAS have identified hundreds of susceptibility loci for osteoporosis and BMD (Estrada et al., 2012; Richards et al., 2012; Kemp et al., 2017; Morris et al., 2019). However, the vast majority (>80%) of reported genome-wide significant susceptibility loci are located in non-coding regions of the genome and provide limited information about the genetic mechanisms of osteoporosis (Liu G. et al., 2019; Manolio et al., 2009).

To clarify the complexities of osteoporosis genetic architecture in both coding and non-coding regions, we provided a comprehensive insight into the genetic and epigenetic mechanisms of osteoporosis and BMD based on the GWAS susceptibility loci. We delineated the genetic architecture of osteoporosis and BMD, and estimated their genetic correlation. Then, we identified the pathways and epigenetic regulatory elements that may be involved in the pathogenesis of osteoporosis and BMD. Finally, we further integrated GWAS and eQTL data to identify the potential functional target genes of the osteoporosis susceptibility loci.

MATERIALS AND METHODS

GWAS Summary Datasets

The GWAS summary data of BMD and osteoporosis were downloaded from the UK Biobank. The BMD dataset includes 206,496 individuals and the osteoporosis dataset includes 35,736 patients and 355,405 controls.

Mapping of SNPs to RefSNP ID and Gene

The human SNP (dpSNP147) and gene (GRCh37) position data were downloaded from NCBI. All the SNPs were mapped to the corresponding RefSNP ID and gene symbol (located within 10 kb upstream or downstream of the SNP) based on the position information (Wang et al., 2010).

Estimation of Genetic Correlation

SNP-based genetic correlation between osteoporosis and BMD was calculated using LD Score regression (LDSC) (Yang et al., 2011). The regression was performed using pre-computed LD scores based on 1000 Genomes European data. To prevent the bias from the variable quality, we removed the variants that are InDels, not in 1000 Genomes European data, strand ambiguous SNPs, SNPs with duplicated rs numbers and SNPs with minor allele frequency (MAF) <0.01.

Pathway Enrichment Analysis

bone mineral density and osteoporosis susceptibility genes that contain at least one significant susceptibility locus were identified (P < 5.00E-08), and the KEGG pathway and GO term enrichment analysis were performed using the CPDB database (Kamburov et al., 2013) based on the BMD and osteoporosis susceptibility genes, respectively. In this study, the adjusted P-value threshold for enrichment analysis is 0.05.

Regulatory Element Enrichment Analysis

Regulatory element enrichment analysis was performed using various regulatory data from the ENCODE and Roadmap Epigenomics projects with GARFIELD software (Iotchkova et al., 2016; Cheng et al., 2018). The fold enrichment (FE) was calculated at different GWAS P-value thresholds (0.1 to 1E-08) after removing the known confounders such as local linkage disequilibrium, local gene density, matched genotyping variants and MAF (Iotchkova et al., 2016). The regulatory element enrichments were tested for various regulatory elements including open chromatin regions, DHSs, transcription factor binding sites and different types of epigenomic markers.

Summary-Data-Based Mendelian Randomization Analysis

We applied a summary-data-based Mendelian randomization (SMR) method integrating osteoporosis and BMD GWAS summary-level data with expression quantitative trait locus (eQTL) data to identify target genes regulated by osteoporosis and BMD susceptibility loci (Zhu et al., 2016). SMR is an instrumental variable analysis approach that uses genetic SNP as an instrumental variable (Z) to test whether the effect of SNP (Z) on the outcome (Y) is mediated by gene expression (X) (Pavlides et al., 2016). The estimation of the effect size of X on Y (β_{XY}) can be expressed as $\beta_{XY} = \beta_{zy}/\beta_{zx}$, where β_{zy} is the effect size of Z on Y and β_{zx} is the effect size of Z on X. GTEx blood eQTL data were used in the SMR analysis (Jiang et al., 2014; GTEx Consortium, 2017), and only SNPs within 1 Mb of the transcription start site are included in this study.

RESULTS

Estimation of Genetic Correlation Between BMD and Osteoporosis

bone mineral density and osteoporosis show a high degree of clinical correlation (Kemp et al., 2017). To investigate whether there is a genetic correlation between the two phenotypes, we analyzed the genetic architecture between BMD and osteoporosis. The BMD and osteoporosis GWAS summary-level data were downloaded from the UK Biobank (Bycroft et al., 2018). We applied a cross-trait LDSC method to estimate genetic correlation by looking for correlations in effect sizes of all

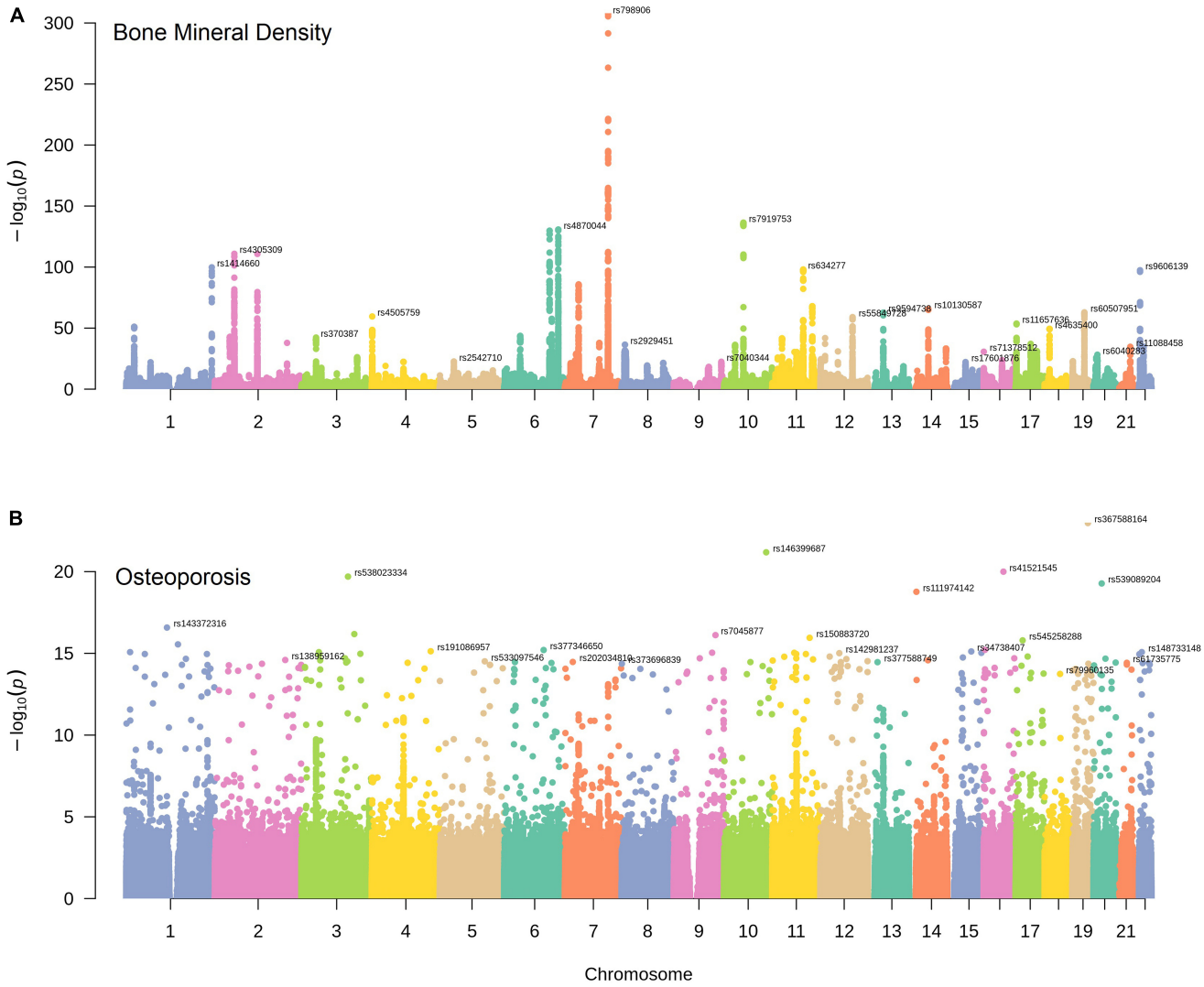


FIGURE 1 | The distribution of BMD and osteoporosis susceptibility loci. Manhattan plots of single nucleotide variant associations for BMD **(A)** and osteoporosis **(B)** identified by UK Biobank. The top-SNP associations on each chromosome are annotated on the plot.

SNPs between BMD and osteoporosis (Yang et al., 2011; Bulik-Sullivan et al., 2015).

The chromosome distribution of BMD and osteoporosis susceptibility loci were displayed graphically in **Figure 1**. We obtained a statistically significant negative genetic correlation between BMD and osteoporosis (ρ_g = −0.57, P = 6.32E-37). In other words, the genetic variants associated with increased risk of osteoporosis tend to induce a decreased BMD level, which conforms with the clinical pathology. It also has been proven that BMD susceptibility loci are significantly enriched for the clinically relevant therapeutic targets of osteoporosis (Richards et al., 2012; Nelson et al., 2015). In our results, BMD and osteoporosis were found to have many common susceptibility loci, 1578 of 5191 osteoporosis susceptibility loci were also associated with the level of BMD (P_{gwas} < 1.0E-04). The common susceptibility loci between BMD and osteoporosis can be found in **Supplementary Table S1** and **Supplementary Figure S1**.

Pathway Enrichment Analysis of BMD and Osteoporosis Susceptibility Loci Located in Protein-Coding Regions

To better understand the potential biological characteristics of the BMD and osteoporosis susceptibility loci located in the protein-coding regions, we mapped all the BMD and osteoporosis susceptibility loci to their corresponding genes based on the position information and conducted a pathway enrichment analysis of all mapped BMD and osteoporosis susceptibility genes with CPDB database. The susceptibility genes of BMD are significantly enriched in 17 KEGG pathways and 58 GO terms, osteoporosis susceptibility genes are enriched in 2 GO terms.

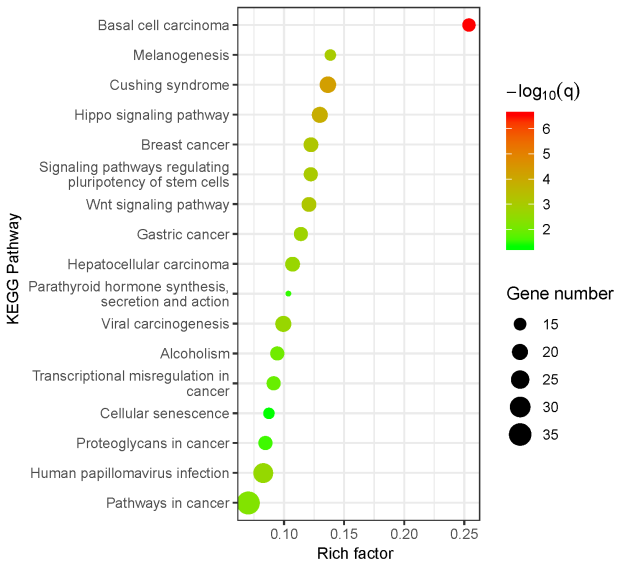


FIGURE 2 | KEGG pathway enrichment analysis for BMD susceptibility loci located in protein-coding regions. The size of the point means the gene number both in BMD susceptibility genes and KEGG pathways. The color of point means enrichment significance ($-\log_{10}$ *Q*-value). The pathways were sorted by the rich factor (the ratio of BMD susceptibility gene number in this pathway to gene number in this pathway term).

More detailed results about all significant pathways can be found in **Supplementary Table S2**.

Our results indicated that the BMD susceptibility genes are significantly enriched in bone-related biological pathways (**Figure 2**). The Wnt/β-catenin and Hippo signaling pathways have been found to play crucial roles in the osteoclasts and osteoblasts balancing (Liu Y. et al., 2019; Saad, 2019), and the GO enrichment results further supported the findings that BMD susceptibility genes are significantly enriched in biological processes of ossification (P = 3.86E-09). What's more, both the glucocorticoid (Cushing syndrome) and parathyroid hormone have profound effects on the homeostasis of bone through multiple cellular and molecular mechanisms (Toth and Grossman, 2013; Wein and Kronenberg, 2018). Besides, we also found that many enriched KEGG pathways for BMD are associated with various types of cancer. It has been reported that cancer is one of the risk factors for osteoporosis due to the effects of cancer cells and cancer-specific therapies on bone cells (Drake, 2013). Therefore, the mutations located in BMD-related susceptibility genes play important roles in the osteoporosis pathogenesis by the dysregulation of the osteoblastic and osteoclastic biological processes.

Regulatory Element Enrichment Analysis of BMD and Osteoporosis Susceptibility Loci Located in Non-coding Regions

The great majority of significant susceptibility loci identified by GWAS are located in non-coding regions such as regulatory elements (Medina-Gomez et al., 2018). In order to investigate whether the susceptibility loci of osteoporosis and BMD are significantly enriched in genomic regulatory elements, we performed a regulatory element enrichment analysis with LD correction (Iotchkova et al., 2016) to calculate FE values for regulatory elements at various genome-wide significant thresholds using regulatory data obtained from the ENCODE and Roadmap Epigenomics project (**Supplementary Table S3**).

We found that both the BMD and osteoporosis susceptibility loci (P < 1.00E-03) are significantly enriched in 5′ UTR region (OR_{BMD} = 3.48, P_{BMD} = 2.6E-02; OR_{OP} = 2.95, P_{OP} = 2.7E-03) according to the physical locations. The 5′ UTR region plays a regulatory role in RNA translation because it contains multiple functional elements (Araujo et al., 2012). The previous pathologic studies found that SNPs in 5′ UTR region can lead to abnormal expression of osteoporosis-related genes, including *TNFRSF11B* and *CYP17* (Tofteng et al., 2004; Krela-Kazmierczak et al., 2016). Our studies provide further evidence that a portion of susceptibility loci may increase osteoporosis risk by 5′ UTR-mediated regulation of BMD-related genes.

We also found that BMD and osteoporosis susceptibility loci (P < 1.00E-04) are significantly enriched in DHSs across different blood cells, especially in various subtypes of leukemia and normal blood lymphocyte (**Figure 3**). DHSs cover many kinds of gene regulatory elements such as enhancers, silencers and locus control regions. It has been proven that there are many osteoclast-specific DHSs located in well-known osteoclast transcription factor binding sites during early osteoclastogenesis (Inoue and Imai, 2014). What's more, both the immune system and leukemia have significant effects on pathogenic mechanisms of osteoporosis (Cagnetta and Patella, 2012; Frisch et al., 2012). The dysregulation of immune cells may directly or indirectly modulate bone metabolism and remodeling through the secretion of various proinflammatory cytokines (Faienza et al., 2013; Srivastava et al., 2018). A significant decrease of BMD was also observed in patients with leukemia (Massenkeil et al., 2001), and the leukemic cells can also influence bone health through altering osteoblastic and osteoclastic functions (Frisch et al., 2012). Therefore, our results of regulatory element enrichment analysis indicated that the BMD and osteoporosis susceptibility loci are significantly enriched in DHSs of peripheral blood immune cells, which may induce bone microarchitectural deterioration by the dysregulation of peripheral blood immune cells.

Identification of Target Genes of BMD and Osteoporosis Susceptibility Loci by Integrating GWAS and eQTL Data

bone mineral density and osteoporosis susceptibility loci were found to be significantly enriched in 5′ UTR and DHSs. To identify the potential functional target genes of the susceptibility loci of osteoporosis, we performed a SMR analysis by integrating the GWAS data with eQTL summary data (Wu et al., 2018). The blood eQTL summary data were obtained from the GTEx project

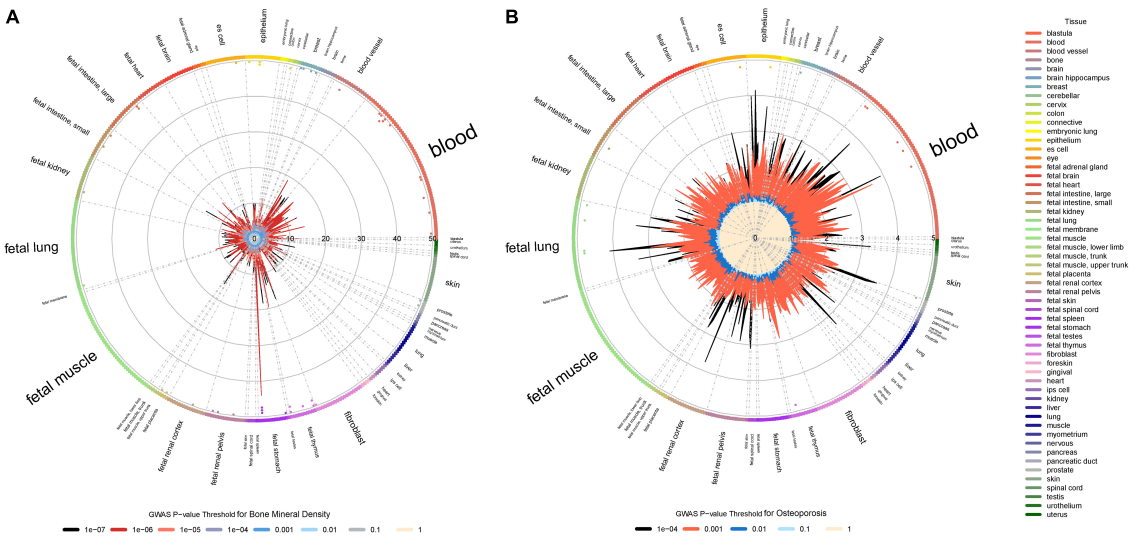


FIGURE 3 | Regulatory element enrichment analysis of BMD and osteoporosis susceptibility loci located in non-coding regions. **(A)** Bone mineral density. **(B)** Osteoporosis. Radial lines with different color show FE values at their corresponding GWAS *P*-value thresholds for all ENCODE and Roadmap Epigenomics DNase I hypersensitive cell lines, sorted by tissue on the outer circle. The font size of tissue is directly proportional to the number of cell lines. The significance of enrichment for a given cell line is marked by the dots in the inner ring of the outer circle.

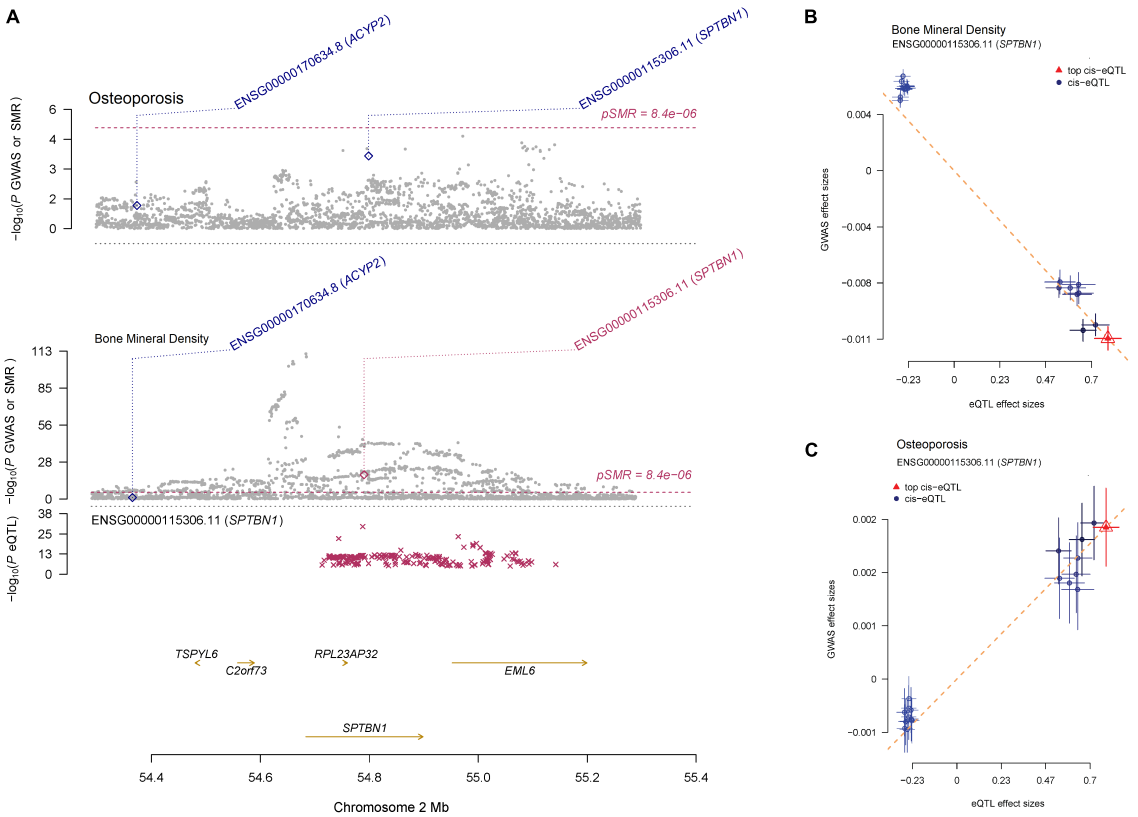


FIGURE 4 | The regulation mechanism of *SPTBN1* locus for BMD and osteoporosis. **(A)** *P*-values of GWAS (gray dots) for BMD (top) and osteoporosis (middle) and *P*-value for the SMR test (diamonds) using the GTEx blood eQTL data. The bottom plot shows the eQTL *P*-values from GTEx blood tissue for the *SPTBN1* gene (red stars). The dots shown in the plot include all the SNPs at these loci in the GWAS and eQTL summary data, respectively. **(B, C)** Effect sizes of SNPs from BMD **(B)** and osteoporosis **(C)** GWAS data against those from the GTEx blood eQTL data. The orange dashed lines represent effect size (b_{xy}) of eQTL on phenotype. Error bars are the standard errors of SNP effects.

(GTEx Consortium, 2017). We mapped all the susceptibility loci to GTEx eQTL target genes in cis-regions and then linked them to BMD and osteoporosis.

Summary-data-based Mendelian randomization identified that 50 target genes were associated with BMD (P_{SMR} < 8.4E-06) (**Supplementary Table S4**), 15 of the 50 target genes were also associated with osteoporosis (P_{SMR} < 0.05), suggesting that the expressions of those genes may have a connection with both BMD and osteoporosis. A notable example is the SNPs located in the *SPTBN1* coding region, where the SNP-association signals are significant and consistent across GWAS and eQTL datasets with P_{SMR} = 3.15E-19 for BMD and P_{SMR} = 2.20E-04 for osteoporosis (**Figure 4A**). The effect sizes of SNPs show the overexpression of SPTBN1 associated with decreased BMD level (**Figure 4B**) and increased osteoporosis risk (**Figure 4C**). The regulation of *SPTBN1* showed a possible pathogenic mechanism for both BMD and osteoporosis.

Among the 15 target genes of BMD and osteoporosis susceptibility loci, *SPTBN1* is a molecular scaffolding protein which has been recognized as an osteoporosis susceptibility gene (Estrada et al., 2012; Chen et al., 2016), and it shows co-expression with alpha-actinin binding and cell adhesion gene in the bone (Calabrese et al., 2017). The overexpression of *ASB16-AS1* also can increase the expression of osteoblastogenesis related genes (Meng et al., 2018). The *SUPT3H-RUNX2* locus associated with bone-related phenotypes including height (Lango Allen et al., 2010) and BMD (Estrada et al., 2012) through regulation of osteoblastic differentiation and skeletogenesis (Rice et al., 2018). Above all, our integration research of GWAS and eQTL data identified 15 target genes of both BMD and osteoporosis susceptibility loci, and the expression of those genes may play important roles in osteoblastic biological processes.

DISCUSSION

In this study, we provided comprehensive insights into the genetics and epigenetic mechanisms of osteoporosis and BMD based on the GWAS summary data. The genetic architecture demonstrated that BMD and osteoporosis share many common susceptibility loci, which allowed us to elucidate pathogenic mechanisms of osteoporosis by integrating the BMD and osteoporosis GWAS summary data. Further, the pathway and regulatory element enrichment analysis found that the susceptibility loci located in both coding and non-coding regions play crucial roles in determining BMD and the pathophysiology of osteoporosis.

The pathway analysis demonstrated that both cancer, Cushing syndrome, Parathyroid hormone, Wnt/β-catenin and Hippo signaling pathways may play crucial roles in osteoporosis pathogenesis by the dysregulation of the osteoblastic and osteoclastic biological processes. What's more, the regulatory element enrichment analysis found BMD and osteoporosis susceptibility loci are more often located in 5′ UTR and DHSs of peripheral blood immune cells. Finally, we found the expression of 15 target genes (*SPTBN1, ASB16-AS1, SUPT3H-RUNX2*, etc.) regulated by susceptibility loci, are associated with BMD and osteoporosis by integrating GWAS summary data with eQTL data.

In summary, we integrated GWAS susceptibility loci with multi-omics datasets to uncover the molecular mechanisms and identify target genes of osteoporosis-associated variants. Our findings shed light on the genetics and epigenetic mechanisms of osteoporosis and provided new clues for effective target therapeutics of osteoporosis. In the future, further researches about the function of those target genes are promising to gain a better understanding of osteoporosis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The GWAS summary data can be downloaded from UK Biobank (http://www.nealelab.is/uk-biobank) under the phenotype code 3148 (bone mineral density) and 20002_1309 (osteoporosis).

AUTHOR CONTRIBUTIONS

HL designed the experiments. WZ obtained the data from UK Biobank. WZ, PW, HD, EZ, XY, SC, TF, LJ, GM, and CC analyzed the data. WZ and HL wrote the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos: 61871068 and 81671827).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.00194/full#supplementary-material

FIGURE S1 | The common susceptibility loci between BMD and osteoporosis.

TABLE S1 | The shared genetic susceptibility loci between osteoporosis (OP) and bone mineral density (BMD).

TABLE S2 | (A) The results of KEGG pathways enrichment analysis for BMD susceptibility genes. **(B)** The results of GO pathways enrichment analysis for BMD susceptibility genes. **(C)** The results of GO pathways enrichment analysis for Osteoporosis susceptibility genes.

TABLE S3 | (A) The results of regulatory element enrichment analysis for BMD susceptibility loci. **(B)** The results of regulatory element enrichment analysis for Osteoporosis susceptibility loci.

TABLE S4 | The results of summary-data-based Mendelian randomization analysis.

REFERENCES

Aggarwal, L., and Masuda, C. (2018). Osteoporosis: a quick update. *J. Fam. Pract.* 67, 59–65.

Al-Barghouthi, B. M., and Farber, C. R. (2019). Dissecting the genetics of osteoporosis using systems approaches. *Trends Genet.* 35, 55–67. doi: 10.1016/j.tig.2018.10.004

Araujo, P. R., Yoon, K., Ko, D., Smith, A. D., Qiao, M., Suresh, U., et al. (2012). Before it gets started: regulating translation at the 5′. *UTR. Comp. Funct. Genomics* 2012:475731. doi: 10.1155/2012/475731

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47, 1236–1241. doi: 10.1038/ng.3406

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z

Cagnetta, V., and Patella, V. (2012). The role of the immune system in the physiopathology of osteoporosis. *Clin. Cases Miner Bone Metab.* 9, 85–88.

Calabrese, G. M., Mesner, L. D., Stains, J. P., Tommasini, S. M., Horowitz, M. C., Rosen, C. J., et al. (2017). Integrating GWAS and Co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell Syst.* 4, 46–59.e4. doi: 10.1016/j.cels.2016.10.014

Chen, Y. C., Guo, Y. F., He, H., Lin, X., Wang, X. F., Zhou, R., et al. (2016). Integrative analysis of genomics and transcriptome data to identify potential functional genes of BMDs in females. *J. Bone. Miner. Res.* 31, 1041–1049. doi: 10.1002/jbmr.2781

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

Cummings, S. R., and Melton, L. J. (2002). Epidemiology and outcomes of osteoporotic fractures. *Lancet* 359, 1761–1767. doi: 10.1016/s0140-6736(02)08657-9

Drake, M. T. (2013). Osteoporosis and cancer. *Curr. Osteoporos. Rep.* 11, 163–170. doi: 10.1007/s11914-013-0154-3

Estrada, K., Styrkarsdottir, U., Evangelou, E., Hsu, Y. H., Duncan, E. L., Ntzani, E. E., et al. (2012). Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* 44, 491–501. doi: 10.1038/ng.2249

Faienza, M. F., Ventura, A., Marzano, F., and Cavallo, L. (2013). Postmenopausal osteoporosis: the role of immune system cells. *Clin. Dev. Immunol.* 2013:575936. doi: 10.1155/2013/575936

Frisch, B. J., Ashton, J. M., Xing, L., Becker, M. W., Jordan, C. T., and Calvi, L. M. (2012). Functional inhibition of osteoblastic cells in an in vivo mouse model of myeloid leukemia. *Blood* 119, 540–550. doi: 10.1182/blood-2011-04-348151

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Inoue, K., and Imai, Y. (2014). Identification of novel transcription factors in osteoclast differentiation using genome-wide analysis of open chromatin determined by DNase-seq. *J. Bone. Miner. Res.* 29, 1823–1832. doi: 10.1002/jbmr.2229

Iotchkova, V., Ritchie, G. R. S., Geihs, M., Morganella, S., Min, J. L., Walter, K., et al. (2016). GARFIELD - GWAS analysis of regulatory or functional information enrichment with LD correction. *bioRxiv* [preprint]. doi: 10.1101/085738

Jiang, Q., Wang, J., Wang, Y., Ma, R., Wu, X., and Li, Y. (2014). TF2LncRNA: identifying common transcription factors for a list of lncRNA genes from ChIP-Seq data. *Biomed. Res. Int.* 2014:317642. doi: 10.1155/2014/317642

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* 41, D793–D800. doi: 10.1093/nar/gks1055

Kemp, J. P., Morris, J. A., Medina-Gomez, C., Forgetta, V., Warrington, N. M., Youlten, S. E., et al. (2017). Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* 49, 1468–1475. doi: 10.1038/ng.3949

Krela-Kazmierczak, I., Kaczmarek-Rys, M., Szymczak, A., Michalak, M., Skrzypczak-Zielinska, M., Drweska-Matelska, N., et al. (2016). Bone metabolism and the c.-223C (T polymorphism in the 5′UTR region of the osteoprotegerin gene in patients with inflammatory bowel disease. *Calcif. Tissue Int.* 99, 616–624. doi: 10.1007/s00223-016-0192-9

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838. doi: 10.1038/nature09410

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. U.S.A.* 116, 3347–3348. doi: 10.1073/pnas.1821550116

Liu, Y., Wang, J., Liu, S., Kuang, M., Jing, Y., Zhao, Y., et al. (2019). A novel transgenic murine model with persistently brittle bones simulating osteogenesis imperfecta type I. *Bone* 127, 646–655. doi: 10.1016/j.bone.2019.07.021

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 gene TRIB3. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10519–E10520. doi: 10.1073/pnas.1812975115

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. doi: 10.1038/nature08494

Massenkeil, G., Fiene, C., Rosen, O., Michael, R., Reisinger, W., and Arnold, R. (2001). Loss of bone mass and vitamin D deficiency after hematopoietic stem cell transplantation: standard prophylactic measures fail to prevent osteoporosis. *Leukemia* 15, 1701–1705. doi: 10.1038/sj.leu.2402264

Medina-Gomez, C., Kemp, J. P., Trajanoska, K., Luan, J., Chesi, A., Ahluwalia, T. S., et al. (2018). Life-Course genome-wide association study meta-analysis of total body BMD and assessment of age-specific effects. *Am. J. Hum. Genet.* 102, 88–102. doi: 10.1016/j.ajhg.2017.12.005

Meng, X. H., Chen, X. D., Greenbaum, J., Zeng, Q., You, S. L., Xiao, H. M., et al. (2018). Integration of summary data from GWAS and eQTL studies identified novel causal BMD genes with functional predictions. *Bone* 113, 41–48. doi: 10.1016/j.bone.2018.05.012

Morris, J. A., Kemp, J. P., Youlten, S. E., Laurent, L., Logan, J. G., Chai, R. C., et al. (2019). An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* 51, 258–266. doi: 10.1038/s41588-018-0302-x

Morrison, N. A., Qi, J. C., Tokita, A., Kelly, P. J., Crofts, L., Nguyen, T. V., et al. (1994). Prediction of bone density from vitamin D receptor alleles. *Nature* 367, 284–287. doi: 10.1038/367284a0

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., et al. (2015). The support of human genetic evidence for approved drug indications. *Nat. Genet.* 47, 856–860. doi: 10.1038/ng.3314

Pavlides, J. M., Zhu, Z., Gratten, J., McRae, A. F., Wray, N. R., and Yang, J. (2016). Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med.* 8:84. doi: 10.1186/s13073-016-0338-4

Rachner, T. D., Khosla, S., and Hofbauer, L. C. (2011). Osteoporosis: now and the future. *Lancet* 377, 1276–1287. doi: 10.1016/S0140-6736(10)62349-5

Ralston, S. H., and Uitterlinden, A. G. (2010). Genetics of osteoporosis. *Endocr. Rev.* 31, 629–662. doi: 10.1210/er.2009-0044

Rice, S. J., Aubourg, G., Sorial, A. K., Almarza, D., Tselepi, M., Deehan, D. J., et al. (2018). Identification of a novel, methylation-dependent, RUNX2 regulatory region associated with osteoarthritis risk. *Hum. Mol. Genet.* 27, 3464–3474. doi: 10.1093/hmg/ddy257

Richards, J. B., Zheng, H. F., and Spector, T. D. (2012). Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nat. Rev. Genet.* 13, 576–588. doi: 10.1038/nrg3228

Saad, F. A. (2019). Novel insights into the complex architecture of osteoporosis molecular genetics. *Ann. N. Y. Acad. Sci.* 1462, 37–52. doi: 10.1111/nyas.14231

Srivastava, R. K., Dar, H. Y., and Mishra, P. K. (2018). Immunoporosis: immunology of osteoporosis-role of T cells. *Front. Immunol.* 9:657. doi: 10.3389/fimmu.2018.00657

Tofteng, C. L., Abrahamsen, B., Jensen, J. E., Petersen, S., Teilmann, J., Kindmark, A., et al. (2004). Two single nucleotide polymorphisms in the CYP17 and COMT Genes–relation to bone mass and longitudinal bone changes in postmenopausal women with or without hormone replacement therapy. The danish osteoporosis prevention study. *Calcif. Tissue Int.* 75, 123–132. doi: 10.1007/s00223-004-0176-z

Toth, M., and Grossman, A. (2013). Glucocorticoid-induced osteoporosis: lessons from Cushing's syndrome. *Clin. Endocrinol.* 79, 1–11. doi: 10.1111/cen.12189

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Wein, M. N., and Kronenberg, H. M. (2018). Regulation of bone remodeling by parathyroid hormone. *Cold Spring Harb. Perspect. Med.* 8:a031237. doi: 10.1101/cshperspect.a031237

Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., et al. (2018). Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.* 9:918. doi: 10.1038/s41467-018-03371-0

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

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 © 2020 Dong, Zhou, Wang, Zuo, Ying, Chai, Fei, Jin, Chen, Ma 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 Genetics

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fgene.2020.00294

Methylation-Driven Genes Identified as Novel Prognostic Indicators for Thyroid Carcinoma

Liting Lv, Liyu Cao, Guinv Hu, Qinyan Shen and Jinzhong Wu*

Department of Thyroid and Breast Surgery, Affiliated Dongyang Hospital of Wenzhou Medical University, Dongyang, China

Background: Aberrant DNA methylation plays an crucial role in tumorigenesis through regulating gene expression. Nevertheless, the exact role of methylation in the carcinogenesis of thyroid cancer and its association with prognosis remains unclear. The purpose of this study is to explore the DNA methylation-driven genes in thyroid cancer by integrative bioinformatics analysis.

Methods: The transcriptome profiling data and DNA methylation data of thyroid cancer were downloaded from The Cancer Genome Atlas (TCGA) database. The methylmix R package was used to screen DNA methylation-driven genes in thyroid cancer. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis were conducted to annotate the function of methylation-driven genes. Univariate Cox regression analyses was performed to distinguish prognosis-related methylation-driven genes. Multivariate Cox regression analyses was utilized to build a prognostic multi-gene signature. A survival analysis was carried out to determine the individual prognostic significance of this multi-gene signature.

Results: A total of 51 methylation-driven genes were identified. The functional analysis indicated that these genes were significantly enriched in diverse biological processes (BP) and pathways related to the malignancy processes. Four of these genes (RDH5, TREM1, BIRC7, and SLC26A7) were selected to construct the risk evaluation model. Patients in the low-risk group had an conspicuously better overall survival (OS) than those in high-risk group ($p < 0.001$). The area under the receiver operating characteristic (ROC) curve for this model was 0.836, suggesting a good specificity and sensitivity. Subsequent survival analysis revealed that this four-gene signature served as an independent indicator for the prognosis of thyroid cancer. Moreover, the prognostic signature was well validated in a external thyroid cancer cohort.

Conclusion: We identified methylation-driven genes in thyroid cancer with independent prognostic value, which may offer new insight into molecular mechanisms of thyroid cancer and provide new possibility for individualized treatment of thyroid cancer patients.

Keywords: thyroid carcinoma, methylation-driven genes, biomarker, TCGA, prognostic indicators

OPEN ACCESS

Edited by:
Xiaotian Zhang,
Van Andel Research Institute (VARI), United States

Reviewed by:
Cheng Zhang,
Fourth Affiliated Hospital of China Medical University, China
Yangyang Hao,
Veracyte, Inc., United States

****Correspondence:***
Liting Lv
lv_liting@163.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *26 December 2019*
Accepted: *12 March 2020*
Published: *31 March 2020*

Citation:
Lv L, Cao L, Hu G, Shen Q and Wu J (2020) Methylation-Driven Genes Identified as Novel Prognostic Indicators for Thyroid Carcinoma. Front. Genet. 11:294. doi: 10.3389/fgene.2020.00294

INTRODUCTION

The incidence of thyroid cancer has increased rapidly in the United States over the last four decades (Lim et al., 2017; Bray et al., 2018). As the fifth most commonly diagnosed cancer in women, thyroid cancer accounts for 40900 new cases estimated by the latest cancer statistic report in the United States (Siegel et al., 2018). This is driven largely by increasing prevalence of papillary thyroid cancer (PTC) which is identified as the most common and least aggressive histologic type in thyroid cancer (Lim et al., 2017). Although thyroid cancer is considered as an indolent malignancy with favorable prognosis, a few patients may suffer local and/or distant recurrence and metastasis, even after surgery or adjuvant radioactive iodine therapy, leading to a poor prognosis (Leboulleux et al., 2016; Yang et al., 2019). Hence, exploring effective and promising biomarkers capable of distinguishing thyroid cancer patients with worse prognosis is of important clinical significance, which is in huge demand.

DNA methylation, a critical element in epigenetic modifications, plays a vital role in the transcriptional regulation and maintains the genome stability (Pu et al., 2016). Accumulating evidence have indicated that aberrant DNA methylation occurred on CpG islands of promoters is capable of regulating expression of many tumor-associated genes and is critical for cancer development (Ferry et al., 2017; Zheng et al., 2017). It was well demonstrated that hypomethylation of oncogene or hypermethylation of tumor suppressor acts as crucial initial events in carcinogenesis (Huang et al., 2011; Yang et al., 2017). For instance, Gao et al. (2017) demonstrated that methylation of TMEM176A was associated with human colorectal cancer development. Aberrant promoter methylation of SOCS-1 was proved to contribute to the pathogenesis of hepatocellular carcinoma (Zhao et al., 2016).

DNA methylation can also be utilized to diagnose cancer as well as predict clinical outcomes. For example, a panel of DNA methylation biomarkers were validated by Lasseigne et al. (2014) to have a good performance in early clinical detection of renal cell carcinoma. Casadio et al. reported that the detection of the methylation frequencies of GSTP1, HIC1, and RASSF1A could be used to predict recurrence of bladder cancer (Casadio et al., 2013). Hence, studying DNA methylation may help elucidate the mechanism of cancer development as well as explore promising diagnostic and prognostic biomarker. Recently, a computational protocol called MethylMix, an algorithm implemented in the R program, can be utilized to identify disease specific hyper/hypomethylation genes significantly associated with their expression (methylation-driven genes) (Cedoz et al., 2018). Several studies focusing on identifying DNA methylation-driven genes using the MethylMix algorithm have been reported in various cancer, such as lung adenocarcinoma, hepatocellular carcinoma, prostate adenocarcinoma. Nevertheless, there is still lack of studies on assessing methylation-driven genes in thyroid carcinoma.

In this study, we used an integrative approach to identify thyroid carcinoma related methylation-driven genes by combining the transcriptomic and DNA methylation profiles from the TCGA database. A four methylation-driven signatures was successfully identified by constructing a Cox survival predictive model, which could effectively distinguish thyroid carcinoma patients with bad prognosis. Our findings will provide new insights into the molecular mechanisms of thyroid carcinoma and prompt a more individualized therapies for this prevalent disease.

MATERIALS AND METHODS

Data Acquisition and Preprocessing

The available RNA-seq transcriptome data, DNA methylation data, and clinicopathological information of thyroid carcinoma were downloaded from the TCGA database[1]. There were 567 samples with gene transcriptome data (58 normal and 509 tumor), 570 samples with DNA methylation data (56 normal and 514 tumor), and 506 patients with available survival data.

The differentially expressed genes (DEGs) between tumor and normal samples were screened firstly by utilizing the "limma" R package. The cutoff criteria was set as $|\log2 \text{ fold change (FC)}| > 1$ and $p < 0.05$.

Screening for DNA Methylation-Driven Genes

The "MethylMix" R package was conducted for screening the DNA methylation-driven genes in thyroid carcinoma by integrating DNA methylation data and paired gene expression data. The cutoff criteria was set as $|\log2 \text{ FC}| > 0$, $p < 0.05$ and correlation coefficient < -0.3. In brief, genes of which the expression was significantly affected by the corresponding DNA methylation events were chosen for further analysis. Then, a beta mixture model was constructed for defining the degree of methylation across the large samples. Finally, Wilcoxon rank test was utilized to calculate differential methylation in tumor and normal samples, and genes met the cutoff criteria were considered as DNA methylation-driven gene (Cedoz et al., 2018). The expression and methylation pattern of those DNA methylation-driven genes in thyroid carcinoma were visualized by "pheatmap" R package.

Functional Enrichment Analysis

Gene ontology (GO) enrichment analysis were conducted to annotate those identified DNA methylation-driven genes by utilizing the Database of Annotation, Visualization and Integrated Discovery (DAVID) v6.8[2]. The top significantly enriched ($p < 0.05$) GO terms of biological process (BP) were visualized by "GOplot" R package. Subsequently, the Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis was also used to perform the pathway enrichment analysis for those DNA methylation-driven genes through KOBAS 3.0[3]. The "GOplot" R package was also used to visualize the significantly enriched pathways ($p < 0.05$).

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

[2]https://david.ncifcrf.gov/

[3]http://kobas.cbi.pku.edu.cn/

TABLE 1 | Methylation-driven genes.

genes	Normal mean	Tumor mean	log FC	*p*-value	Cor	Cor *p*-value
RAET1E	0.914157	0.697235	−0.390797	4.45E−29	−0.534397	6.26E−39
SERPINA1	0.632612	0.514767	−0.297402	1.51E−26	−0.535371	4.32E−39
LPAR5	0.457776	0.273762	−0.741721	3.01E−26	−0.646354	1.54E−61
RDH5	0.898498	0.430300	−1.062172	1.02E−24	−0.644569	4.21E−61
TREM1	0.795173	0.638780	−0.315949	2.06E−24	−0.342375	1.91E−15
LIPH	0.768724	0.665275	−0.208516	2.95E−24	−0.603783	7.03E−52
LRP4	0.325453	0.191412	−0.765766	7.06E−24	−0.494515	9.51E−33
COL8A2	0.496770	0.384904	−0.368082	8.23E−24	−0.442363	8.42E−26
DCSTAMP	0.682648	0.386206	−0.821769	8.25E−22	−0.653196	3.09E−63
LGALS1	0.376251	0.190589	−0.981231	9.00E−22	−0.477586	2.32E−30
TMEM173	0.336766	0.205043	−0.715818	6.24E−21	−0.764352	1.11E−98
MUC21	0.824024	0.723442	−0.187809	1.78E−20	−0.333290	1.14E−14
CTXN1	0.831012	0.659250	−0.334044	3.22E−20	−0.599326	5.96E−51
LONRF2	0.790843	0.531867	−0.572325	5.67E−20	−0.327829	3.23E−14
AHNAK2	0.765057	0.615665	−0.313421	6.59E−20	−0.714908	8.18E−81
AK1	0.775374	0.590682	−0.392510	9.14E−20	−0.493644	1.27E−32
ABTB2	0.810591	0.447253	−0.857882	1.79E−18	−0.678615	5.92E−70
CDSN	0.798050	0.618537	−0.367620	2.49E−18	−0.498284	2.68E−33
SCEL	0.699450	0.568227	−0.299754	9.04E−18	−0.480317	9.76E−31
S100A16	0.794216	0.607415	−0.386850	2.26E−17	−0.547830	3.33E−41
PLAU	0.904954	0.853874	−0.083821	7.95E−17	−0.478322	1.84E−30
CDH16	0.262682	0.383591	0.546251	1.06E−16	−0.393862	2.46E−20
TMEM40	0.797271	0.733412	−0.120446	1.47E−16	−0.465595	9.50E−29
MYBPHL	0.542402	0.425328	−0.350787	1.80E−15	−0.482369	5.07E−31
PDZK1IP1	0.577887	0.459305	−0.331335	2.19E−15	−0.504215	3.54E−34
BPIFB1	0.832361	0.785808	−0.083033	7.10E−15	−0.396836	1.20E−20
RIN1	0.510654	0.425439	−0.263393	7.68E−15	−0.640856	3.33E−60
MYO1G	0.588103	0.410445	−0.518881	2.73E−14	−0.664549	3.73E−66
TGM1	0.758261	0.556121	−0.447296	4.46E−14	−0.612998	7.60E−54
ALOX15B	0.341945	0.265716	−0.363879	5.72E−13	−0.438093	2.77E−25
LAMB3	0.728829	0.638443	−0.191021	1.51E−12	−0.674408	8.53E−69
TMEM100	0.638705	0.506043	−0.335889	1.58E−12	−0.523246	4.06E−37
CD3G	0.635594	0.735447	0.210517	2.10E−12	−0.453344	3.63E−27
LINC00607	0.847477	0.699606	−0.276632	3.24E−12	−0.474113	6.91E−30
SLC26A4	0.187955	0.269248	0.518549	1.25E−11	−0.577639	1.24E−46
SLC1A5	0.506373	0.378163	−0.421193	1.52E−11	−0.548142	2.94E−41
DUSP5	0.493559	0.374514	−0.398205	4.39E−11	−0.513583	1.33E−35
BIRC7	0.655324	0.584382	−0.165297	6.04E−10	−0.480075	1.05E−30
S100A10	0.657108	0.510199	−0.365071	9.06E−10	−0.705123	1.01E−77
CD151	0.714099	0.649581	−0.136613	1.44E−09	−0.484509	2.54E−31
VTCN1	0.785455	0.709718	−0.146284	2.22E−09	−0.492672	1.76E−32
NAPSA	0.847424	0.788968	−0.103117	6.69E−08	−0.608248	7.99E−53
KLK11	0.684627	0.610142	−0.166173	1.74E−07	−0.740069	2.14E−89
EPPK1	0.765503	0.683879	−0.162667	2.20E−07	−0.507768	1.03E−34
SGCD	0.911930	0.923126	0.017604	2.58E−05	−0.465736	9.10E−29
BASP1	0.656438	0.574277	−0.192911	4.44E−05	−0.323740	6.96E−14
S100A4	0.613862	0.580597	−0.080377	9.76E−05	−0.604334	5.39E−52
SLC26A7	0.322965	0.356954	0.144364	0.000287214	−0.408232	7.30E−22
TMEM105	0.596893	0.545987	−0.128605	0.000421052	−0.482211	5.33E−31
C1orf116	0.903649	0.830643	−0.121534	0.014746146	−0.332451	1.34E−14
EPHA4	0.305452	0.293696	−0.056623	0.026836525	−0.507321	1.21E−34

log FC, log fold change; Cor, Correlation.

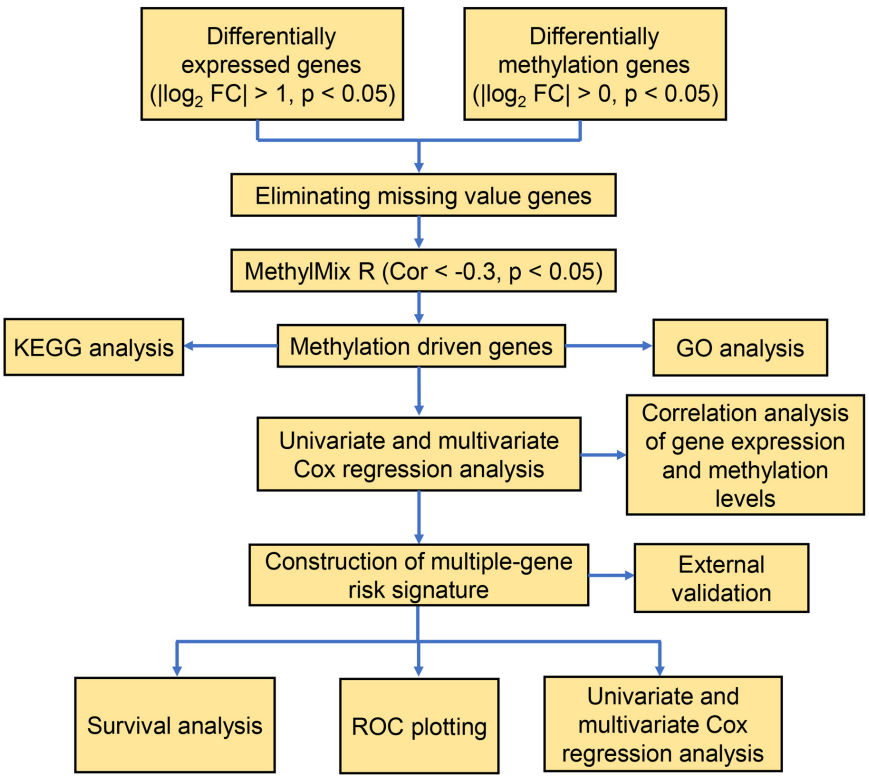


FIGURE 1 | A flow chart of the exploration of methylation-driven genes in thyroid cancer.

Gene Set Enrichment Analysis (GSEA) was performed using the software GSEA v4.0.3[4]. After normalization of gene expressions data of 567 samples (509 tumor and 58 normal), GO analysis was conducted by employment of GSEA software mentioned above. GO gene sets from MSigDB (Molecular Signature Database) were used as reference. The analysis process was repeated 1,000 times using the default weighted enrichment statistics methods. All other parameters were set based on their default values.

Construction of Prognostic Signatures and Survival Analysis

Univariate Cox regression analysis was conducted to determine the relationship between the expression of DNA methylation-driven genes and prognosis of thyroid carcinoma patients. Genes with a p-value < 0.05 were regarded as prognostic methylation-driven genes and were subsequently fitted into the multivariate Cox regression analysis. A DNA methylation-driven gene-based prediction model was constructed by the linear combination of the expression levels of methylation-driven genes using coefficients (β) calculated from multivariate Cox regression as the weights. The risk score for each patient was calculated based on the risk score formula: risk score = expression of gene1 × β1 + expression of gene2 × β2 + . . .expression of genen × βn. After that, patients were divided into high-risk and low-risk groups by setting the median value of risk scores as cut-off value. The overall survival (OS) of these two groups was calculated by the Kaplan-Meier method with log-rank test. Receiver operating characteristic (ROC) curve were performed to assess the predictive performance of the prognostic model. The expression patterns of DNA methylation-driven genes in this prognostic model were visualized by "pheatmap" R package.

Univariate and multivariate Cox regression analyses were conducted to determine whether the risk score calculated from the prognostic model was independent prognostic factors for thyroid carcinoma patients after considering other clinical features, including age, gender and AJCC stage.

Cell Culture

The papillary thyroid carcinoma cell lines (TPC-1 and K1) were cultured in RPMI 1640 medium (Gibco, Life Technologies, CA, United States) with 10% fetal bovine serum (Biological Industries, CT, United States) at 37°C and 5% CO_2.

The Validation Patient Cohort

The validation study was approved by the First Affiliated Hospital of Zhejiang University. A total of 200 specimens of thyroid cancer were included in the validation cohort. The detailed clinicopathological information of the validated cohort was summarized in **Supplementary Table S1**. Written informed consent was obtained from all the patients.

[4]www.gsea-msigdb.org

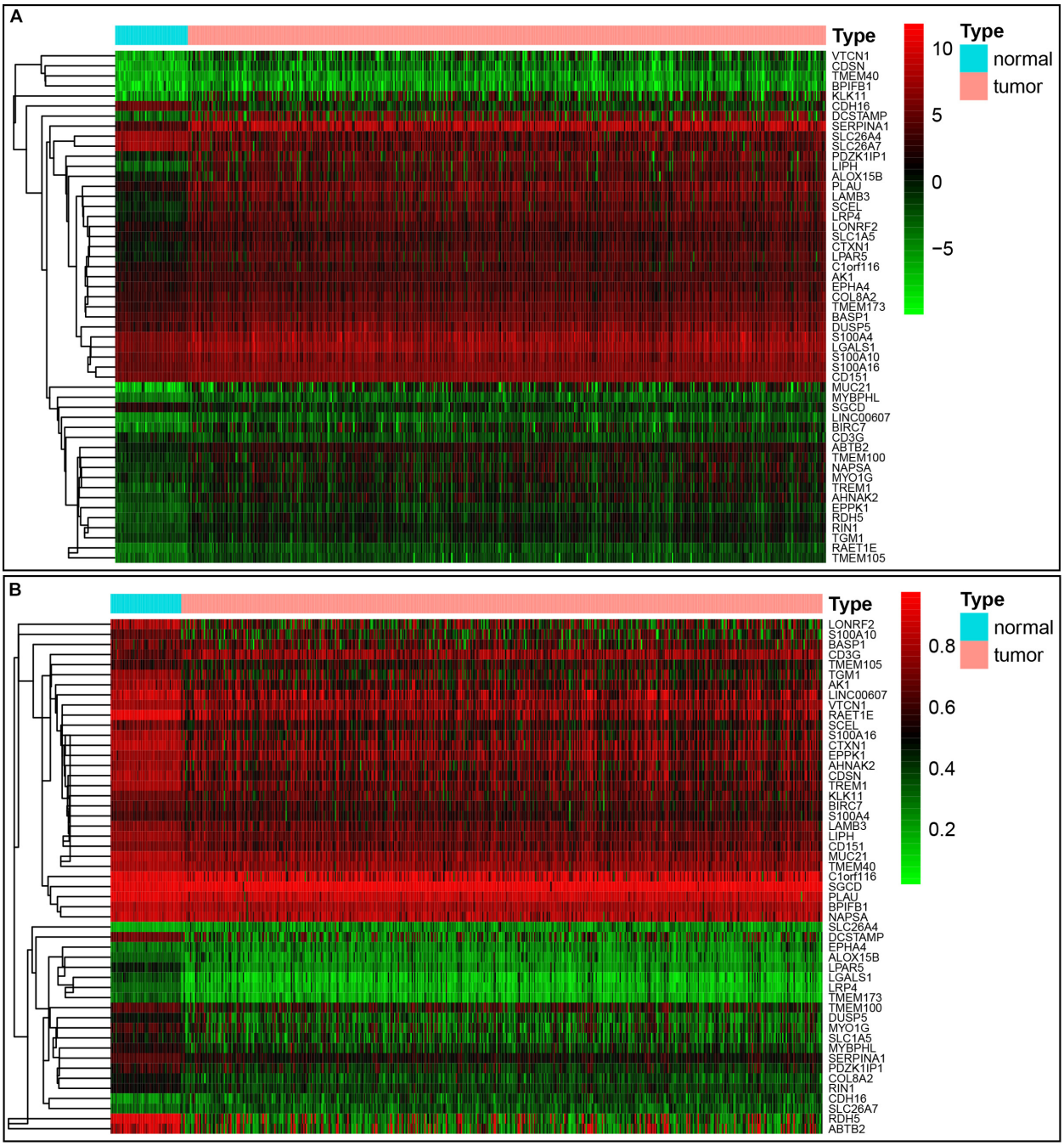
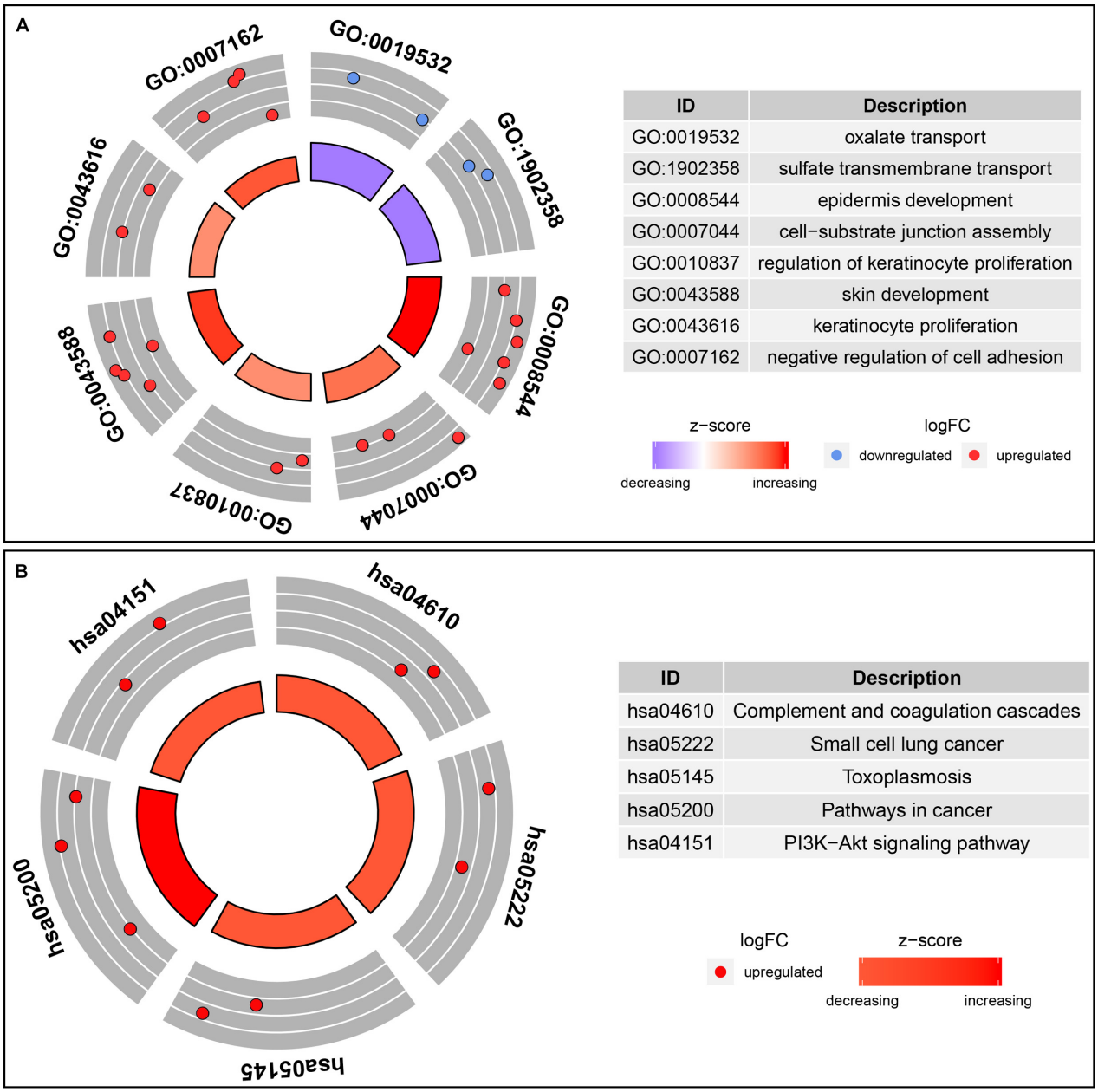


FIGURE 2 | Heatmap of 51 methylation-driven genes in thyroid cancer. **(A)** The expression pattern of 51 methylation-driven genes. Red represents upregulated genes and green represents downregulated genes between tumor and normal tissues. **(B)** The methylation pattern of 51 methylation-driven genes. Red represents highly methylated genes and green represents low methylated genes between tumor and normal tissues.

Quantitative Reverse Transcription PCR

The total RNA was extracted using Trizol reagent (Invitrogen, United States). The cDNA for each cell line and tissue specimens was reverse transcribed using the PrimeScript 1st Strand cDNA Synthesis Kit (TaKaRa, Dalian, China). qRT-PCR analysis was conducted using the SYBR-Green method according to standard protocols. The sequences of the primers used were as follows: RDH5, 5′-TGGGTGGAGATGCACGTTAAG-3′ (forward), 5′-GTGTGGGTCCGATGATACCAG-3′ (reverse); TREM1, 5′-GAACTCCGAGCTGCAACTAAA-3′ (forward), 5′-TCTAGCGT GTAGTCACATTTCAC-3′ (reverse); BIRC7, 5′-GCTCTGAGG AGTTGCGTCTG-3′ (forward), 5′-CACACTGTGGACAAAGT



ID	Description
GO:0019532	oxalate transport
GO:1902358	sulfate transmembrane transport
GO:0008544	epidermis development
GO:0007044	cell−substrate junction assembly
GO:0010837	regulation of keratinocyte proliferation
GO:0043588	skin development
GO:0043616	keratinocyte proliferation
GO:0007162	negative regulation of cell adhesion

ID	Description
hsa04610	Complement and coagulation cascades
hsa05222	Small cell lung cancer
hsa05145	Toxoplasmosis
hsa05200	Pathways in cancer
hsa04151	PI3K−Akt signaling pathway

FIGURE 3 | Functional enrichment analysis of methylation-driven genes in thyroid cancer. **(A)** GO enrichment analysis. **(B)** KEGG enrichment analysis. The color of inner circle represents *z* score while the band thickness of inner circle represents the significance of GO terms (log10-adjusted *p* values). The outer circle represents the expression (log_2 FC) of methylation-driven mRNAs in each enriched GO (gene ontology) term: red dots indicate upregulated methylation-driven mRNAs while blue dots indicate downregulated methylation-driven mRNAs.

CTCTT-3′ (reverse); SLC26A7, 5′-AGAAGGCGACTGCCCA TTTT-3′ (forward), 5′-ACTGCCAACATTATCCCAGACA-3′ (reverse).

Validation of the Prognostic Signature

The validated cohort was divided into high-risk group and low-risk group based on their risk scores. The cut-off value was set as the median of the risk scores in TCGA thyroid cancer cohort. Then, the difference of OS between high-risk and low-risk group was calculated. Univariate and multivariate Cox regression analyses were utilized to determine whether risk score was an independent prognostic factor.

Statistical Analysis

All data in the present study were analyzed by utilizing the R statistical package (R version 3.6.1) unless otherwise stated. A two-tailed $p < 0.05$ was considered statistically significant. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

TABLE 2 | Coefficients based on a multivariate Cox regression analysis of four genes.

genes	Coef.	HR	95% CI of HR	*p*-value
RDH5	−0.331	0.719	0.457–1.129	0.151
TREM1	0.165	1.180	1.059–1.314	0.003
BIRC7	0.017	1.018	1.007–1.028	0.001
SLC26A7	0.016	1.017	1.007–1.026	0.001

HR, hazard ratio.

RESULTS

Identification of Methylation-Driven Genes in Thyroid Carcinoma

Our study included RNA-sequencing data from 567 samples from thyroid carcinoma patients, including 58 normal samples and 509 tumor samples. DNA methylation data were extracted from 570 thyroid carcinoma specimens, including 56 normal samples and 514 tumor samples. Using the cutoff criteria of FDR < 0.05 and $|\log2FC| > 1$, a total of 3430 DEGs (1751 upregulated and 1679 downregulated) were screened for further analysis. The gene expression data and DNA methylation data for 3430 DEGs were included in the MethylMix analysis with a screening criteria set as $|\log2FC| > 0$, $p < 0.05$ and Cor < -0.3. As a result, we totally identified 51 DNA methylation-driven genes of which 46 were hypomethylated while 5 were hypermethylated (**Table 1**). A flow chart of the exploration of methylation-driven genes was shown in **Figure 1**. The expression pattern and methylation value of methylation-driven genes were shown as heat map (**Figures 2A,B**).

Functional Analysis of Methylation-Driven Genes in Thyroid Carcinoma

In order to understand the possible function of those DNA methylation-driven genes, the GO functional enrichment analysis and KEGG pathway enrichment analysis were conducted. A BP analysis was mainly performed in GO analysis. The result indicated that DNA methylation-driven genes were significantly enriched ($p < 0.05$) in terms associated with cell proliferation, including epidermis development, regulation of keratinocyte proliferation, skin development, keratinocyte proliferation (**Figure 3A**). In addition, a GO terms with regard to negative regulation of cell adhesion was also statistically significant ($p < 0.05$) (**Figure 3A**). Moreover, pathway analysis also showed that the methylation-driven genes were significantly enriched in malignancy-related pathways, such as small cell lung cancer, pathways in cancer, and PI3K-Akt signaling pathway. The pathway analysis was shown in **Figure 3B**. In addition, the GSEA analysis of 51 DNA methylation-driven genes was also performed. The results showed that regulation of wnt signaling pathway, cell signaling, endoplasmic reticulum part, and molecular function regulator were significantly enriched for those DNA methylation-driven genes (**Supplementary Figures S1A–D**).

Construction of a Methylation-Driven Gene-Based Risk Signature

For the purpose of determining the prognostic role of DNA methylation-driven genes in thyroid carcinoma, univariate Cox regression analysis was performed firstly to identify prognosis

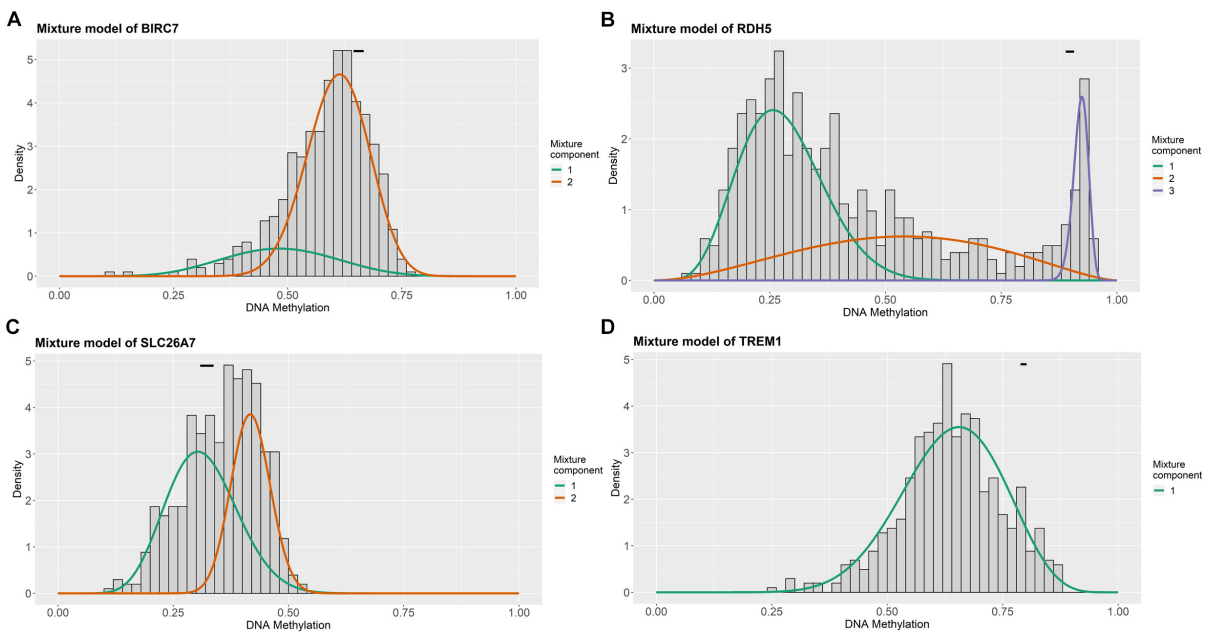


FIGURE 4 | Distribution map of the methylation degree of BIRC7 **(A)**, RDH5 **(B)**, SLC26A7 **(C)**, and TREM1 **(D)** in the risk model. *X*-axis represents the degree of methylation and *Y*-axis represents the number of methylated samples; The black horizontal line represents the methylation degree distribution in the normal samples.

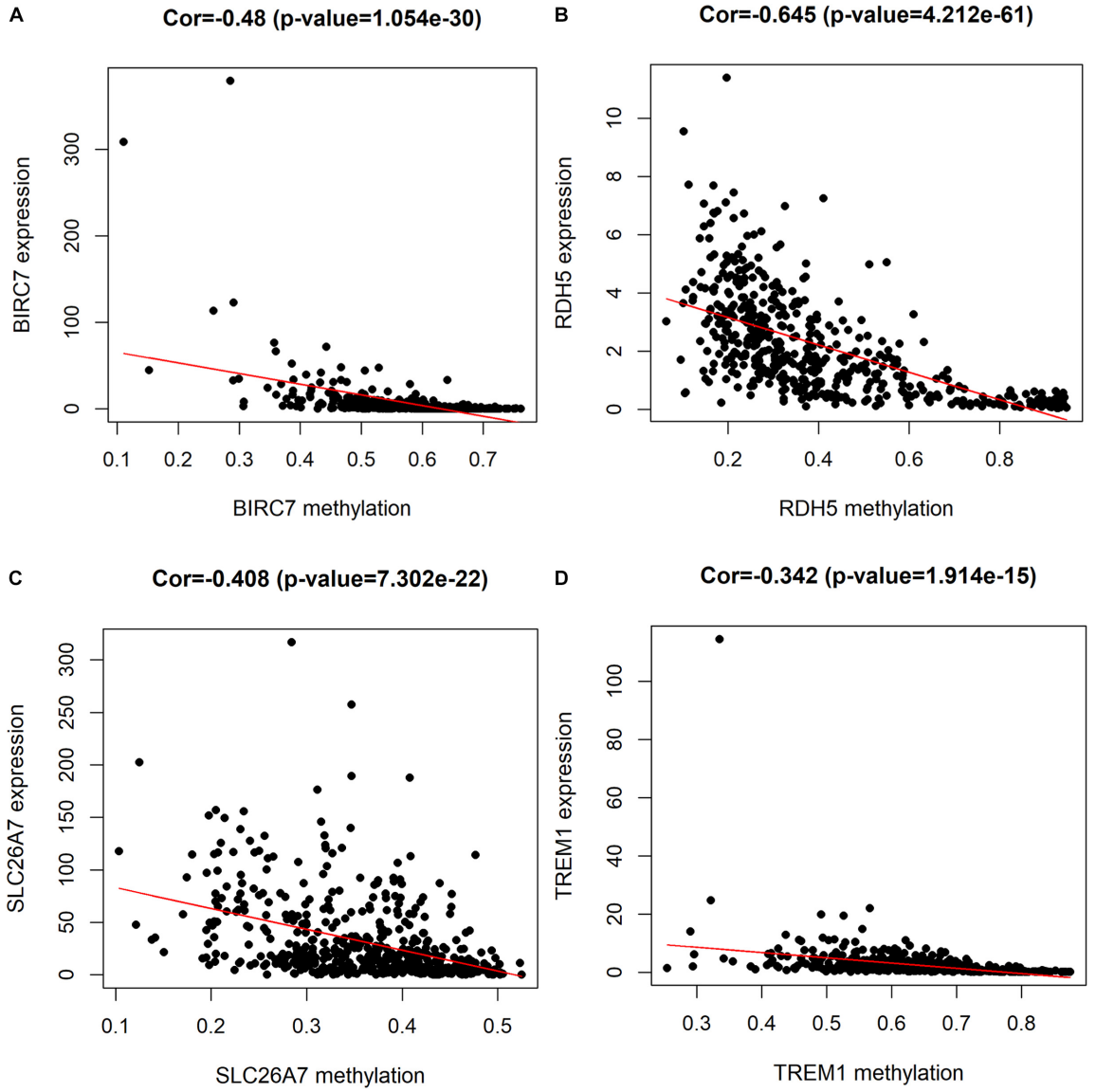


FIGURE 5 | Correlation between the expression and methylation degree of BIRC7 **(A)**, RDH5 **(B)**, SLC26A7 **(C)**, and TREM1 **(D)** in the risk model. *X*-axis represents the methylation degree and *Y*-axis represents gene expression level.

associated methylation-driven genes in TCGA cohort. The results indicated that 4 methylation-driven genes (TREM1, CDH16, BIRC7, and SLC26A7) were risky genes with HR > 1, while 3 genes (LPAR5, RDH5, and LIPH) served as protective genes with HR < 1 (**Supplementary Table S2**). Subsequently, multivariate Cox regression analysis was utilized and showed that four methylation-driven genes (RDH5, TREM1, BIRC7, and SLC26A7) were eventually chosen to build a predictive model. The result of multivariate Cox regression analysis was shown in **Table 2**. Through linear combination of the expression of the 4 methylation-driven genes, the coefficient of each gene were calculated from the multivariate Cox regression analysis. As a result, a risk score for each patient could be calculated using the following formula: risk score = (−0.331) × expression value of RDH5 + (0.165) × expression value of TREM1 + (0.017) × expression value of BIRC7 + (0.016) × expression value of SLC26A7. The mixed models for the four genes in the prognostic model with regard to the methylation degree in normal and tumor tissues were visualized in **Figure 4**. As shown in **Figure 5**, all the methylation degrees of RDH5, TREM1, BIRC7, and SLC26A7 were negatively correlated with corresponding gene expression. To validate the role of methylation in the regulation of the expression of these 4 genes, the thyroid cancer cell lines, TPC-1 and K1, were treated with methylation inhibitor 5-aza. As shown in **Supplementary Figure S2**, the expression of all these 4 genes was significantly up-regulated in a dose-dependent manner, which further demonstrated it as methylation-driven genes.

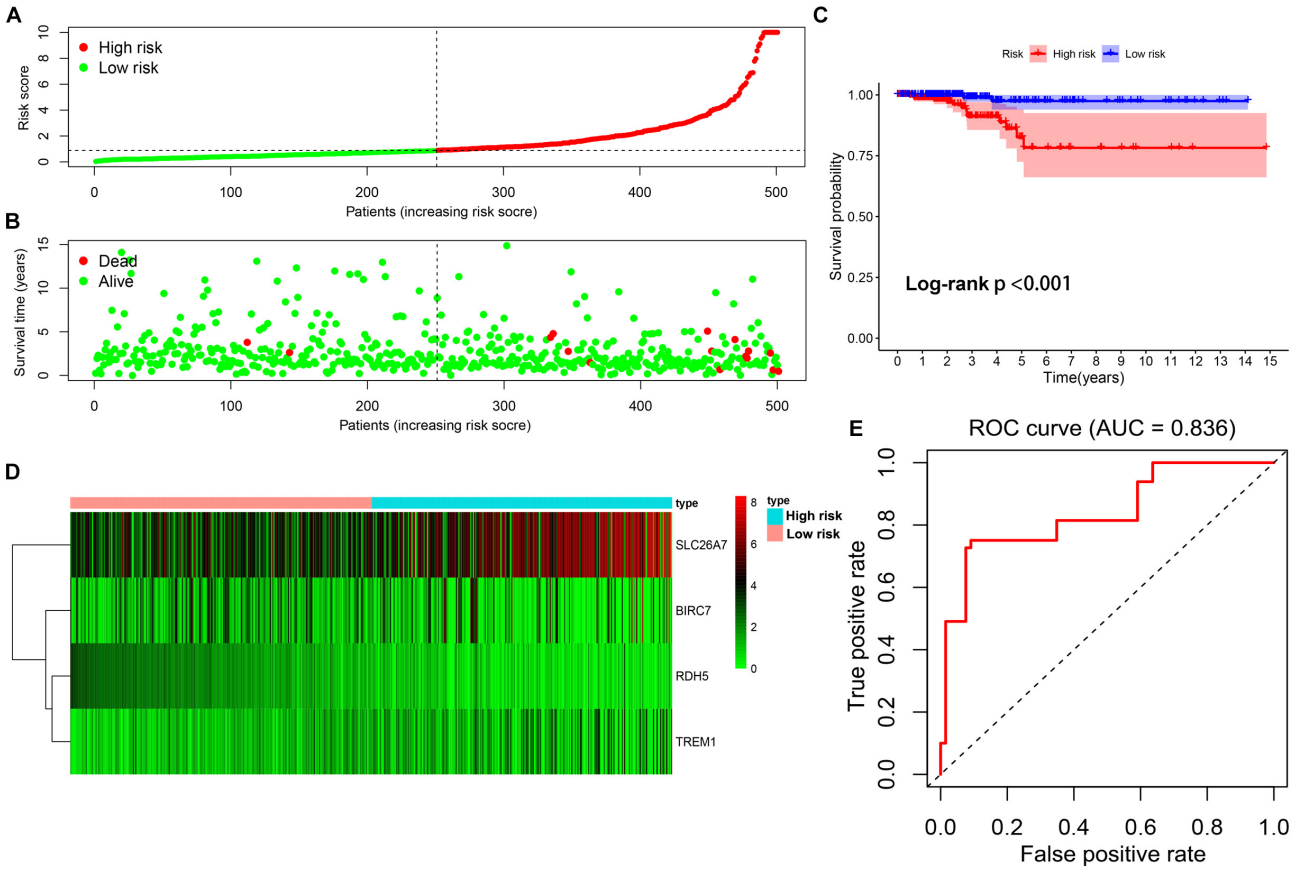


FIGURE 6 | Construction of the multi-gene prognostic signature. **(A)** The distributions of patients' risk scores. **(B)** The distributions of risk scores and OS status. **(C)** The survival analysis of the two subgroups stratified based on the median of risk scores. **(D)** The expression pattern of the four methylation-driven genes in low- and high-risk groups. **(E)** The ROC curve for evaluating the prediction efficiency of the prognostic signature.

Survival and ROC Curve Analysis

By using the median of risk scores as cut-off value (0.887), a total of 501 thyroid carcinoma patients with complete survival information were divided into the low-risk group (n = 251) and high-risk group (n = 250). The distributions of the four gene signature-based risk scores were showed in **Figure 6A**. Moreover, the distributions of risk scores and OS status of each patient were displayed in **Figure 6B**, suggesting a good discrimination between low-risk and high-risk group. By means of plotting Kaplan-Meier curve, survival analysis demonstrated that patients in the low-risk group had a conspicuously better OS than those in the high-risk group ($p < 0.001$) (**Figure 6C**). In addition, **Figure 6D** exhibited the expression pattern of these 4 methylation-driven genes in thyroid carcinoma. Finally, the ROC curve analysis further showed an excellent prediction efficiency with a AUC value equal to 0.836 (**Figure 6E**).

The Signature-Based Risk Score Was an Independent Prognostic Factor

Univariate and multivariate Cox analyses were conducted to determine if the four-gene signature-based risk score was an independent prognostic factor. The univariate Cox analysis indicated that age ($p < 0.001$, HR = 1.152, 95% CI = 1.095–1.213), AJCC stage ($p < 0.001$, HR = 2.478, 95% CI = 1.557–3.943), T stage ($p = 0.006$, HR = 2.382, 95% CI = 1.289–4.403), and risk score ($p < 0.001$, HR = 1.344, 95% CI = 1.142–1.582) were dramatically associated with the OS (**Figure 7A**). When all these factors were enlisted into the multivariate Cox regression analysis, only age ($p < 0.001$, HR = 1.163, 95% CI = 1.100–1.230) and risk score ($p < 0.001$, HR = 1.602, 95% CI = 1.287–1.995) were identified as the independent prognostic factors (**Figure 7B**).

Validation of the Prognostic Signature

The prognostic value of the four-gene risk signature was validated in validation cohort (n = 200). Based on aforementioned cut-off value, a total of 68 patients were grouped into high-risk subgroup while the remaining 132 patients were categorized into low-risk group. The distributions of the risk scores and OS status were shown in **Figures 8A,B**. The Kaplan-Meier curve demonstrated that patients in high risk group had an obviously poorer OS compared to patients with low risk ($p < 0.05$) (**Figure 8C**). The ROC curves also demonstrated that risk score (AUC = 0.714) had a good predictive ability (**Figure 8D**). Then,

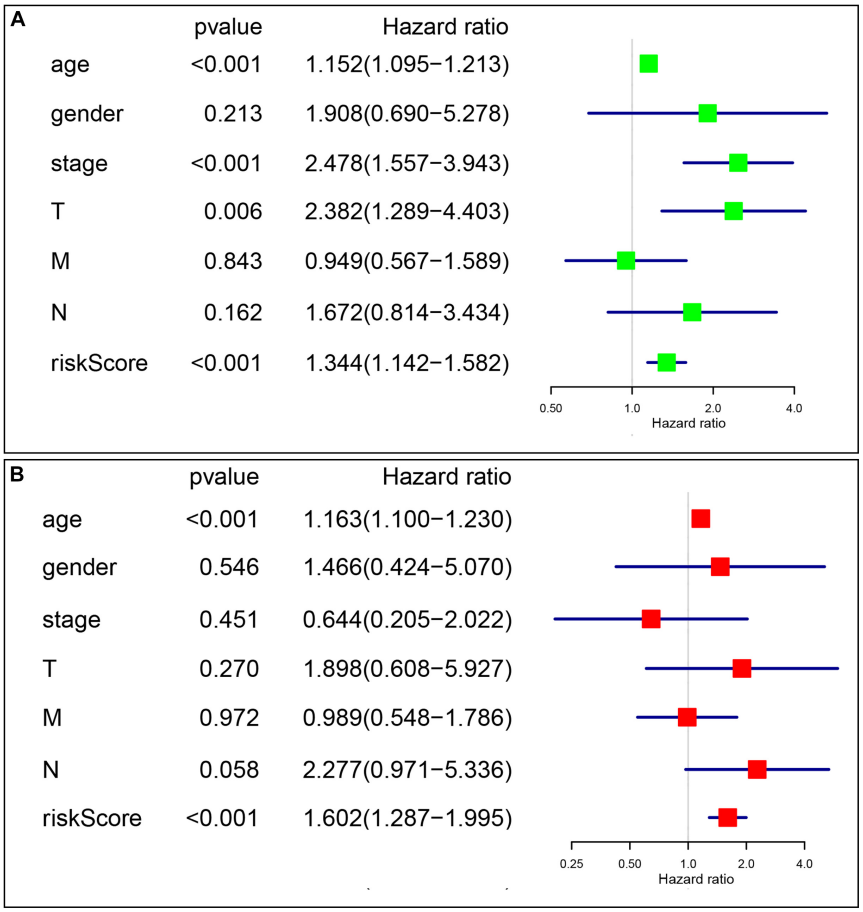


FIGURE 7 | The signature-based risk score was an independent prognostic factor. **(A)** Univariate analysis of the risk score and clinicopathological features in TCGA thyroid cancer cohort. **(B)** Multivariate analysis of the risk score and clinicopathological features in TCGA thyroid cancer cohort.

the univariate analysis demonstrated that age (HR = 1.175, 95% CI [1.103–1.252], $p < 0.001$), T stage (HR = 3.472, 95% CI [1.298–9.286], $p = 0.013$), M stage (HR = 2.712, 95% CI [1.080–6.810], $p = 0.034$), and signature-based risk score (HR = 1.422, 95% CI [1.081–1.869], $p = 0.012$) were significantly associated with the OS in validation cohort (**Figure 9A**). The multivariate analysis further showed that signature-based risk score served as independent prognostic indicators (HR = 1.405, 95% CI [1.055–1.870], $p = 0.020$) (**Figure 9B**). These results, taken together, convincingly verified the prognostic value of this four-gene risk signature in patients with thyroid cancer.

DISCUSSION

According to statistics, the overall incidence of thyroid cancer in United States increased significantly from 1994 to 2013 (approximately 3% per year) (Lim et al., 2017). In addition, a dramatically increase in thyroid cancer mortality rate (approximately 1.1% annually), especially in advanced stage PTC (2.9% per year), were also reported in this period (Lim et al., 2017). Indeed, although most of the patients with thyroid cancer have excellent prognosis, part of them have worse prognosis due to tumor recurrence or distant metastasis (Kazaure et al., 2012). Therefore, it is of great importance to excavate novel biomarkers that can indicate these patients with bad prognosis.

A great deal of research have demonstrated a strong relationship between epigenetic aberrations and genetic aberrations in tumorigenesis (You and Jones, 2012; Chang et al., 2013). It is commonly believed that epigenetic changes, such as DNA methylation, can drive abnormal gene expression of crucial genes involved in the development and progression of cancer, including prostate cancer (Chen et al., 2014), liver cancer (Feitelson, 2006), head and neck cancer (Worsham et al., 2014), etc. In addition, previous studies also reported that the methylation status of specific genes significantly associated with worse prognosis (Lee et al., 2006; Zhu et al., 2017). Therefore, the methylation-driven genes could serve as attractive prognostic indicator in tumor patients. For example, Long et al. used two DNA methylation-driven genes, SPP1 and LCAT, to construct two-gene signature which acted as an independent predictor for prognosis of liver cancer (Long et al., 2019). Methylated hub genes, including HOXD3, LAT, and NFE2L3, were proved to be a novel prognostic indicators in

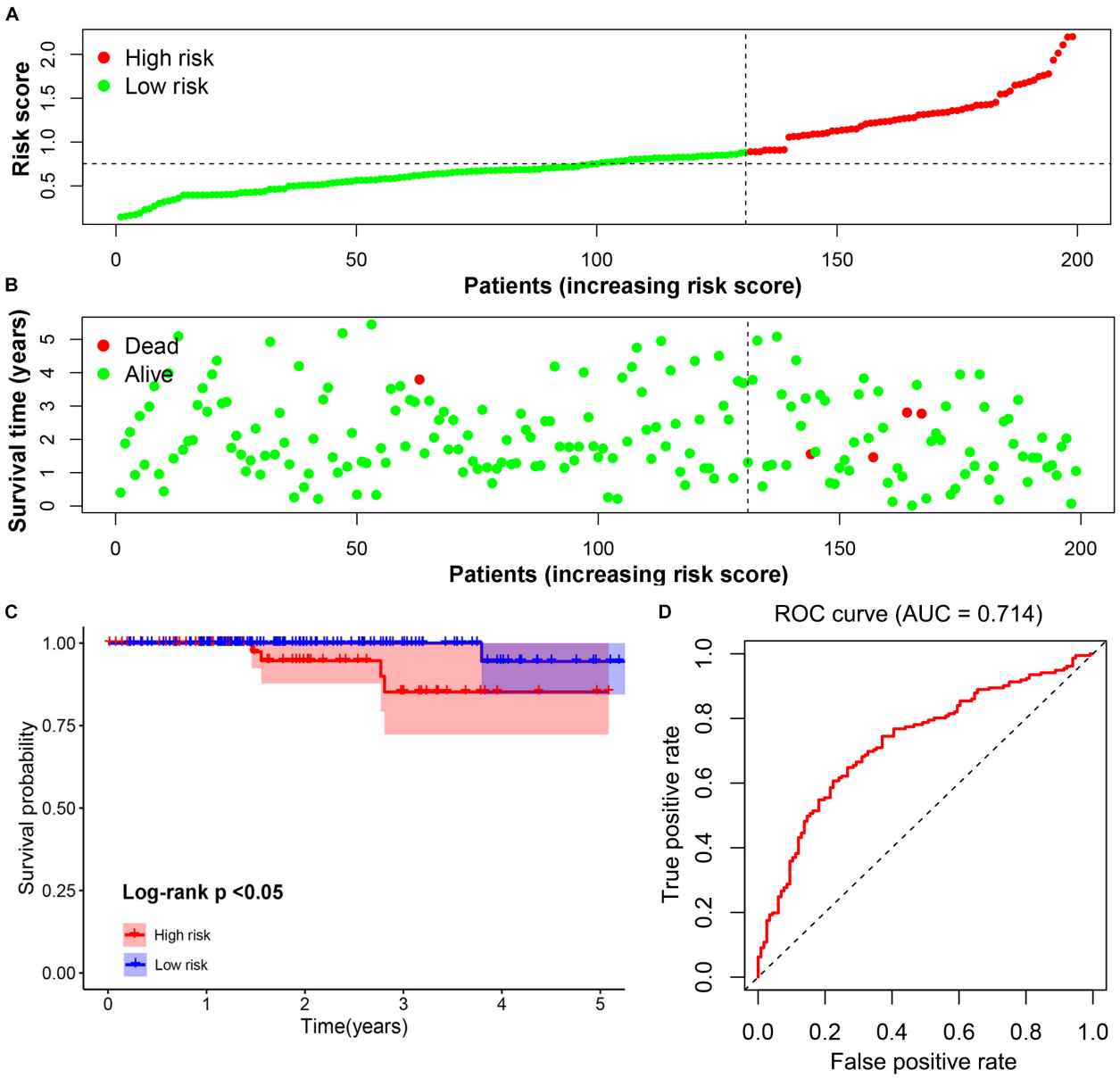


FIGURE 8 | Validation of the prognostic risk signature. **(A,B)** The distributions of prognostic signature-based risk scores. The red dots represent high-risk patients while green dots represent low-risk patients. **(C)** The survival analysis of the two subgroups. **(D)** The ROC curve for evaluating the prediction efficiency of the prognostic signature.

clear cell renal cell carcinoma (Wang et al., 2019). However, to our best knowledge, there is still a lack of research on screening DNA methylation-driven genes as prognostic biomarker in thyroid cancer.

In our study, we conducted a comprehensive view of DNA methylation-driven genes in thyroid cancer and developed a prognostic signature based on the expression values of four methylation-driven genes. A cohort of 51 DNA methylation-driven genes was identified firstly in thyroid cancer. The functional analysis indicated that these genes were significantly enriched in diverse BP and pathways ranging from cell proliferation, cell adhesion and pathways in cancer. These results implied that DNA methylation might functionally relate to the malignancy processes of thyroid cancer. Subsequently, a risk multi-genes signature including four methylation-driven genes (RDH5, TREM1, BIRC7, and SLC26A7) was constructed to serve as a reliable predictor by means of univariate and multivariate Cox analysis. Based on the risk score calculated from the four-genes signature, the thyroid cancer patients in TCGA cohort could be divided into two groups with high- or low-risk. Survival analysis indicated that thyroid cancer patients with high risk had significantly inferior OS than those with low

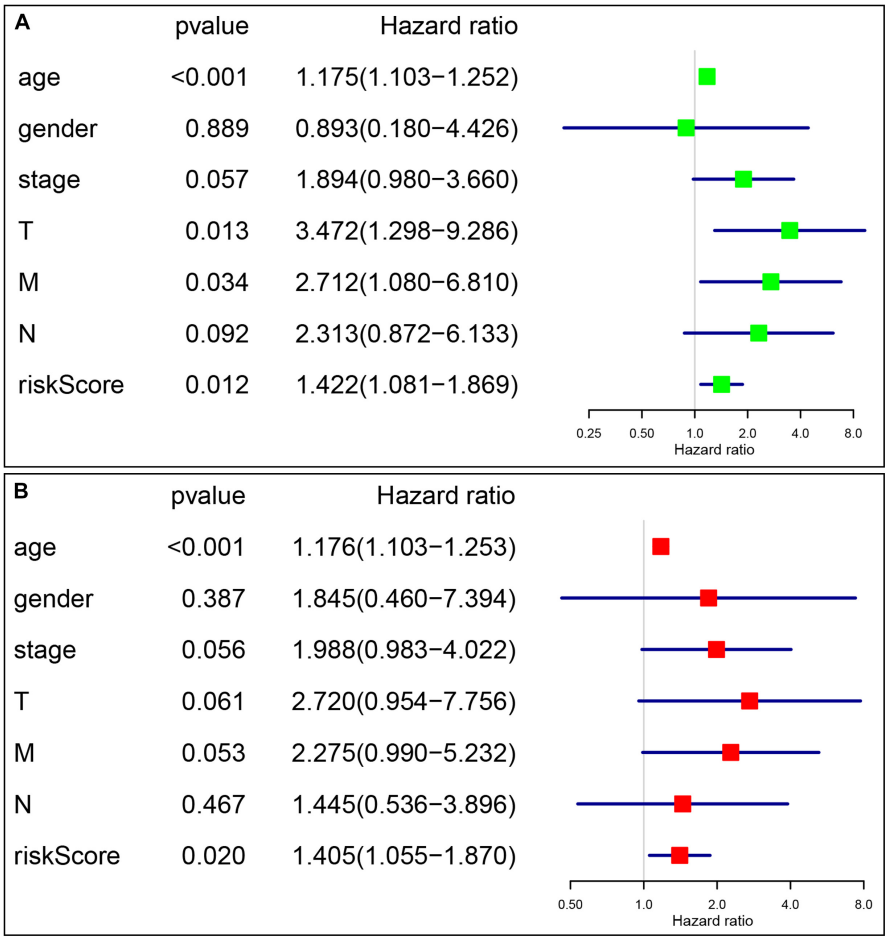


FIGURE 9 | Identification of the independent prognostic factors in the validation group. **(A)** Univariate Cox analyses of the signature based risk score and clinicopathological parameters in validation cohort; **(B)** Multivariate Cox analyses of the signature based risk score and clinicopathological parameters in validation cohort.

risk. The AUC of the ROC curve based on this signature was as high as 0.836 at 5 years of OS. The results of univariate and multivariate Cox analyses further demonstrated that signature-based risk score was an independent prognostic factor. Finally, we validated this four-gene risk signature in validation cohort, which further suggested its convincing prognostic value in thyroid cancer patients. Interesting, we found that traditional pathological indicators, such as AJCC stage, was no longer independent prognostic factors, implying that our risk signature could emerge as a stable and reliable indicator capable of predicting the prognosis of thyroid cancer. Although we are unable to confirm whether the DNA-methylation driven genes based risk factor can perform better than other biomarkers, such as DNA-methylation driven genes and variants/fusions, the prognostic signature built in our study could contribute to distinguish thyroid cancer patients with poor outcome. Importantly, we hold the opinion that for those thyroid cancer patients classified as high-risk, an aggressive transdisciplinary management, such as surgery and adjuvant radioactive iodine, should be considered.

Among the four methylation-driven genes, the high expression level of TREM1, BIRC7, and SLC26A7 prognosticated low survival rate, whereas RDH5 acted as protective genes to suggest good prognosis of thyroid cancer. TREM1 is a activating member of the Ig superfamily that functions as a potent amplifier of pro-inflammatory innate immune responses (Bouchon et al., 2000). Mounting evidence indicated that the overexpression of TREM1 is associated with the development of several types of cancer, such as colorectal carcinoma (Saurer et al., 2017) and hepatocellular tumor (Duan et al., 2015). In terms of BIRC7, Ge et al. (2019) revealed that BIRC7, an important member of the human inhibitor of apoptosis proteins (IAPs) family, promoted colon cancer progression. Wei et al. (2018) also reported that the hypomethylation of BIRC7 was closely related to the pathogenesis of bone tumor. Meanwhile, overexpression of BIRC7 was also used to predict the worse prognosis of various cancer patients (Ibrahim et al., 2014; Sun et al., 2018). Consistently, our study demonstrated that the hypomethylation of TREM1 and BIRC7 contributed to its overexpression, suggesting its potential protumorigenic role in thyroid cancer.

In addition, SLC26A7 and RDH5 also have been previously identified to be associated with anaplastic thyroid carcinoma (ATC) and papillary thyroid carcinoma, respectively (Beltrami et al., 2017; Weinberger et al., 2017). Therefore, our integrative analysis provided a convincing clue that genes potentially regulated by DNA methylation may serve as potential drivers and biomarkers related to thyroid cancer development. Our findings also support the notion that DNA methylation-driven genes are likely to be associated with clinical outcomes and can be utilized as potential biomarkers for predicting the prognosis of thyroid cancer.

CONCLUSION

In conclusion, we screened DNA methylation-driven genes in thyroid cancer for the first time by using bioinformatics analysis from the TCGA database. A four-gene signature was constructed firstly by employing DNA methylation-driven genes, which served as an independent prognostic indicator for thyroid cancer. The results of our study may provide new method for the identification of thyroid cancer patients with clinical high-risk, and may open the way for the possible clinical application of methylation-driven genes.

DATA AVAILABILITY STATEMENT

All RNA-seq transcriptome data, DNA methylation data, and clinicopathological information are available in the TCGA database.

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

LL, LC, GH, QS, and JW conceived and designed the present study. LL, LC, and GH analyzed the data. LL and LC interpreted the data and wrote 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/fgene.2020.00294/full#supplementary-material

FIGURE S1 | The partial enriched GO terms **(A–D)** in GSEA analysis.

FIGURE S2 | Validation of DNA methylation-driven genes in thyroid cancer cells. **(A)** Relative expression of RDH5 in cells treated with 5-aza-2-deoxycytidine (5-aza); **(B)** Relative expression of TREM1 in cells treated with 5-aza; **(C)** Relative expression of BIRC7 in cells treated with 5-aza; **(D)** Relative expression of SLC26A7 in cells treated with 5-aza. $^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$.

TABLE S1 | The results of univariate Cox analysis. HR, hazard ratio.

TABLE S2 | The clinical information of the validated thyroid cancer cohort.

REFERENCES

Beltrami, C. M., Dos Reis, M. B., Barros-Filho, M. C., Marchi, F. A., Kuasne, H., Pinto, C. A. L., et al. (2017). Integrated data analysis reveals potential drivers and pathways disrupted by DNA methylation in papillary thyroid carcinomas. *Clin. Epigenet.* 9:45. doi: 10.1186/s13148-017-0346-342

Bouchon, A., Dietrich, J., and Colonna, M. (2000). Cutting edge: inflammatory responses can be triggered by TREM-1, a novel receptor expressed on neutrophils and monocytes. *J. Immunol.* 164, 4991–4995. doi: 10.4049/jimmunol.164.10.4991

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

Casadio, V., Molinari, C., Calistri, D., Tebaldi, M., Gunelli, R., Serra, L., et al., (2013). DNA Methylation profiles as predictors of recurrence in non muscle invasive bladder cancer: an MS-MLPA approach. *J. Exp. Clin. Cancer Res.* 32:94. doi: 10.1186/1756-9966-32-94

Cedoz, P. L., Prunello, M., Brennan, K., and Gevaert, O. (2018). MethylMix 2.0: an R package for identifying(DNA)methylation genes. *Bioinformatics* 34, 3044–3046. doi: 10.1093/bioinformatics/bty156

Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Chin. J. Lung Cancer* 45, 1113–1120. doi: 10.1038/ng.2764

Chen, L. N., Rubin, R. S., Othepa, E., Cer, C., and Suy, S. (2014). Correlation of HOXD3 promoter hypermethylation with clinical and pathologic features in screening prostate biopsies. *Prostate* 74, 714–721. doi: 10.1002/pros.22790

Duan, M., Wang, Z. C., Wang, X. Y., Shi, J. Y., Yang, L. X., Ding, Z. B., et al. (2015). TREM-1, an inflammatory modulator, is expressed in hepatocellular carcinoma cells and significantly promotes tumor progression. *Ann. Surg. Oncol.* 22, 3121–3129. doi: 10.1245/s10434-014-4191-4197

Feitelson, M. A. (2006). Parallel epigenetic and genetic changes in the pathogenesis of hepatitis virus-associated hepatocellular carcinoma. *Cancer Lett.* 239, 0–20.

Ferry, L., Fournier, A., Tsusaka, T., Adelmant, G., Shimazu, T., Matano, S., et al. (2017). Methylation of DNA Ligase 1 by G9a/GLP recruits UHRF1 to replicating DNA and regulates DNA Methylation. *Mol. Cell.* 67:550-565.e555. doi: 10.1016/j.molcel.2017.07.012

Gao, D., Han, Y., Yang, Y., Herman, J. G., Linghu, E., Zhan, Q., et al. (2017). Methylation of TMEM176A is an independent prognostic marker and is involved in human colorectal cancer development. *Epigenetics* 12, 575–583. doi: 10.1080/15592294.2017.1341027

Ge, Y., Liu, B. L., Cui, J. P., and Li, S. Q. (2019). Livin promotes colon cancer progression by regulation of H2A.X(Y39ph) via JMJD6. *Life Sci.* 234:116788. doi: 10.1016/j.lfs.2019.116788

Huang, Z. H., Hu, Y., Hua, D., Wu, Y. Y., Song, M. X., and Cheng, Z. H. (2011). Quantitative analysis of multiple methylated genes in plasma for the diagnosis and prognosis of hepatocellular carcinoma. *Exp. Mol. Pathol.* 91, 702–707. doi: 10.1016/j.yexmp.2011.08.004

Ibrahim, L., Aladle, D., Mansour, A., Hammad, A., Al Wakeel, A. A., and Abd El-Hameed, S. A. (2014). Expression and prognostic significance of livin/BIRC7 in childhood acute lymphoblastic leukemia. *Med. Oncol.* 31:941. doi: 10.1007/s12032-014-0941-944

Kazaure, H. S., Roman, S. A., and Sosa, J. A. (2012). Aggressive variants of papillary thyroid cancer: incidence, characteristics and predictors of survival among 43,738 patients. *Ann. Surg. Oncol. Oncol. J. Surg.* 19, 1874–1880. doi: 10.1245/s10434-011-2129-x

Lasseigne, B. N., Burwell, T. C., Patil, M. A., Absher, D. M., Brooks, J. D., and Myers, R. M. (2014). DNA methylation profiling reveals novel diagnostic biomarkers

in renal cell carcinoma. *BMC Med.* 12:235. doi: 10.1186/s12916-014-0235-x

Leboulleux, S., Tuttle, R. M., Pacini, F., and Schlumberger, M. (2016). Papillary thyroid microcarcinoma: time to shift from surgery to active surveillance? *Lancet Diabetes Endocrinol.* 4, 933–942. doi: 10.1016/s2213-8587(16)30180-30182

Lee, E. J., Lee, B. B., Kim, J. W., Shim, Y. M., Hoseok, I., Han, J., et al. (2006). Aberrant methylation of Fragile Histidine Triad gene is associated with poor prognosis in early stage esophageal squamous cell carcinoma. *Eur. J. Cancer* 42, 0–980.

Lim, H., Devesa, S. S., Sosa, J. A., Check, D., and Kitahara, C. M. (2017). Trends in thyroid cancer incidence and mortality in the United States1974-2013. *JAMA* 317, 1338–1348. doi: 10.1001/jama.2017.2719

Long, J., Chen, P., Lin, J., Bai, Y., Yang, X., Bian, J., et al. (2019). DNA methylation-driven genes for constructing diagnostic, prognostic, and recurrence models for hepatocellular carcinoma. *Theranostics* 9, 7251–7267. doi: 10.7150/thno.31155

Pu, W., Geng, X., Chen, S., Tan, L., Tan, Y., Wang, A., et al. (2016). Aberrant methylation of CDH13 can be a diagnostic biomarker for lung adenocarcinoma. *J Cancer* 7, 2280–2289. doi: 10.7150/jca.15758

Saurer, L., Zysset, D., Rihs, S., Mager, L., Gusberti, M., Simillion, C., et al. (2017). TREM-1 promotes intestinal tumorigenesis. *Sci. Rep.* 7:14870. doi: 10.1038/s41598-017-14516-4

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA A Cancer J. Clin.* 68:7. doi: 10.3322/caac.21442

Sun, K., Liao, Q., Chen, Z., Chen, T., and Zhang, J. (2018). Expression of Livin and PlGF in human osteosarcoma is associated with tumor progression and clinical outcome. *Oncol. Lett.* 16, 4953–4960. doi: 10.3892/ol.2018.9239

Wang, J., Zhao, H., Dong, H., Zhu, L., Wang, S., Wang, P., et al. (2019). LAT, HOXD3 and NFE2L3 identified as novel DNA methylation-driven genes and prognostic markers in human clear cell renal cell carcinoma by integrative bioinformatics approaches. *J. Cancer* 10, 6726–6737. doi: 10.7150/jca.35641

Wei, H., Pan, L., Tao, D. Y., and Li, R. G. (2018). Correlation between the methylation of LIVIN gene and the pathogenesis of bone tumor. *Eur. Rev. Med. Pharmacol. Sci.* 22(1 Suppl.), 23–28. doi: 10.26355/eurrev_201807_15355

Weinberger, P., Ponny, S. R., Xu, H., Bai, S., Smallridge, R., Copland, J., et al. (2017). Cell Cycle M-phase genes are highly upregulated in anaplastic thyroid carcinoma. *Thyroid* 27, 236–252. doi: 10.1089/thy.2016.0285

Worsham, M. J., Stephen, J. K., Chen, K. M., Havard, S., Shah, V., Gardner, G., et al. (2014). Delineating an epigenetic continuum in head and neck cancer. *Cancer Lett.* 342, 178–184. doi: 10.1016/j.canlet.2012.02.018

Yang, F., Zhong, Q., Huang, Z., Lian, M., and Fang, J. (2019). Survival in papillary thyroid microcarcinoma: a comparative analysis between the 7th and 8th versions of the AJCC/UICC staging system based on the SEER database. *Front. Endocrinol.* 10:10. doi: 10.3389/fendo.2019.00010

Yang, X., Gao, L., and Zhang, S. (2017). Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Brief Bioinform.* 18, 761–773. doi: 10.1093/bib/bbw063

You, J. S., and Jones, P. A. (2012). Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* 22, 9–20. doi: 10.1016/j.ccr.2012.06.008

Zhao, R. C., Zhou, J., He, J. Y., Wei, Y. G., Qin, Y., and Li, B. (2016). Aberrant promoter methylation of SOCS-1 gene may contribute to the pathogenesis of hepatocellular carcinoma: a meta-analysis. *J. Buon* 21, 142–151.

Zheng, X., Zhang, N., Wu, H. J., and Wu, H. (2017). Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.* 18:17. doi: 10.1186/s13059-016-1143-1145

Zhu, Y., Zhang, W., and Wang, P. (2017). Smoking and gender modify the effect of TWIST on patient survival in head and neck squamous carcinoma. *Oncotarget* 8, 85816–85827. doi: 10.18632/oncotarget.20682

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 © 2020 Lv, Cao, Hu, Shen and Wu. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fbioe.2020.00245

PSBP-SVM: A Machine Learning-Based Computational Identifier for Predicting Polystyrene Binding Peptides

Chaolu Meng[1,2†], Yang Hu[3†], Ying Zhang[4] and Fei Guo[1*]*

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China, [2] College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot, China, [3] School of Life Sciences and Technology, Harbin Institute of Technology, Harbin, China, [4] Department of Pharmacy, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China

Polystyrene binding peptides (PSBPs) play a key role in the immobilization process. The correct identification of PSBPs is the first step of all related works. In this paper, we proposed a novel support vector machine-based bioinformatic identification model. This model contains four machine learning steps, including feature extraction, feature selection, model training and optimization. In a five-fold cross validation test, this model achieves 90.38, 84.62, 87.50, and 0.90% SN, SP, ACC, and AUC, respectively. The performance of this model outperforms the state-of-the-art identifier in terms of the SN and ACC with a smaller feature set. Furthermore, we constructed a web server that includes the proposed model, which is freely accessible at http://server.malab.cn/PSBP-SVM/index.jsp.

Keywords: polystyrene binding peptides, support vector machine, bioinformatic, machine learning, identifier

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Hao Lin,
University of Electronic Science and Technology of China, China
Xiangxiang Zeng,
Xiamen University, China

****Correspondence:***
Ying Zhang
zhangying_hmu@163.com
Fei Guo
fguo@tju.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *31 January 2020*
Accepted: *09 March 2020*
Published: *31 March 2020*

Citation:
Meng C, Hu Y, Zhang Y and Guo F (2020) PSBP-SVM: A Machine Learning-Based Computational Identifier for Predicting Polystyrene Binding Peptides. Front. Bioeng. Biotechnol. 8:245. doi: 10.3389/fbioe.2020.00245

INTRODUCTION

The immobilization of a biological functional molecule on a solid surface is one of the most important topics in the field of biology. Immobilized enzymes are a typical application of this technology and are commonly used in industrial reactors (Eş et al., 2015). The nature of the biocompatibility on the implant surface is considered to be a protein absorption process (Yin et al., 2020). The enzyme-linked immunosorbent assay (ELISA) (Engvall and Perlmann, 1971) is a well-known method for identifying counterparts in biological interactions. This assay is derived from the immobilization target antigen molecules (Li et al., 2018). There are two principles in immobilization: one principle is orienting the target part in the preferred direction, and the other principle is avoiding any unnecessary interaction between the target and the solid surface.

Polystyrene (PS) is used as a protein solid surface in ELISAs and animal cell cultures because of its biological inertia (Kumada et al., 2010). Polystyrene with binding peptides can be used to immobilize bioactive peptides, enzymes and antigens in water at room temperature. These functional monolayer protein layers can be widely applied in the medicine, textile and automobile industries (Yaman et al., 2009; Moritomi et al., 2010; Mrozek and Malysiak-Mrozek, 2011; Modjarrad, 2013). Polystyrene binding peptides (PSBPs) can combine with target proteins or peptides to determine their improper orientation in the immobilization process (Bakhshinejad and Sadeghizadeh, 2016; Wang et al., 2020). The correct recognition of PSBPs is the first and most important step of its related application. It is time-consuming and expensive to use a wet experiment to verify these peptides. To identify PSBPs, we turn to machine learning-based

computing strategies. To date, machine learning algorithms have been widely used in biological sequence recognition (Wang et al., 2008, 2010, 2018; Zhou et al., 2017, 2018, 2019; He et al., 2018; Liao et al., 2018; Xu et al., 2018a,b; Bao et al., 2019; Cheng et al., 2019; Ding et al., 2019; Fang et al., 2019; Jin et al., 2019; Liu et al., 2019a; Meng et al., 2019; Shen et al., 2019; Zhu et al., 2019). This process generally includes data collection, feature extraction, feature selection and model training. The positive and negative samples are collected to form a training dataset, and the sequence recognition problem is transformed into a binary classification problem. Discrete features are extracted from training datasets via the feature extraction process. The pseudo amino acid composition (PseAAC) is one of the most commonly used feature extraction algorithms, and many improved algorithms have been produced (Shen and Chou, 2008). The MRMD (Max-Relevance-Max-Distance) (Zou et al., 2018), ANOVA (analysis of variance) (Anderson, 2001) and mRMR (Minimal Redundancy Maximal Relevance) (Ding and Peng, 2005) are commonly used feature selection algorithms. The aim of these feature selection algorithms is overcoming the data redundancy problem. Choosing a good classification algorithm is another particularly important step, and the SVM (support vector machine), random forest, and Bayes classifiers have been widely used to address sequence recognition problems (Li et al., 2019). Ning et al. combined a SVM and the dipeptide composition (DPC) feature, named PSBinder, to construct an identifier to recognize PSBPs (Li N. et al., 2017)[1]. In this study, we used the same training dataset as PSBinder.

In the "Materials and methods" section, we describe the data collection process, the feature extraction method, the ANOVA feature selection, the SVM and the evaluation metrics. We depict the workflow of the proposed identifier and comprehensively analyze the performance of the identifier in the "Results and discussion" section. In the "Conclusion" section, we analyze the shortcomings of the model and look forward to its future improvement.

MATERIALS AND METHODS

Data Collection

We use the same training dataset as PSBinder. This benchmark dataset includes 104 positive samples (PSBPs) and 104 negative samples (non-PSBPs). This dataset is collected from the BDB database (released in January 2017) according to the following criteria. The raw positive samples are selected from nine different phage display libraries. Furthermore, in order to ensure the difference between the positive and negative samples, we attempt to select the same numbers of negative and positive samples from each of the above-mentioned libraries. For those libraries that do not have enough negative samples, we select the same length sequences from the other libraries instead. Then, cysteine amino acids are deleted because they found are at both ends of the circular peptides (Fu et al., 2018). Peptides that contain two specific kinds of characters are removed. One kind is ambiguous characters including "B," "J," "O," "U," "X," and "Z." The other kind is non-alphabetic characters. Two measures are used to screen the above data. Then, we compare each sequence in the positive and negative sample sets, delete the same negative sample sequences and positive sample sequences and replace them with other new negative samples (Yang et al., 2019b). Moreover, the Generalized Jaccard similarity is applied to keep the similarity between the positive and negative samples below 90% (Pan et al., 2009).

Feature Extraction

The amino acid residue frequency is one of the most important features of protein sequences (Małysiak-Mrozek et al., 2018a; Liu, 2019). The frequency feature can be calculated via the single amino acid composition (AAC), the DPC, three or more peptides' composition or peptides with a certain gap. There are several proteins or peptide identifiers that have been proposed based on these features. In this paper, we use the weighted frequency of the single AAC and the DPC as the discrete extraction feature.

A peptide consists of 20 kinds of amino acid residues. Thus, a peptide can be presented as follows:

$$p = A_1A_2A_3 \ldots A_i \ldots A_{L-1}A_L \quad (1)$$

where A_i is the ith amino acid residue of peptide p with a length of L.

(i) 20-dimensional amino acid composition (AAC)

The weighted frequency of the single AAC is defined as follows:

$$Feature_{AAC} = \{ (f_1, f_2, \ldots, f_i, \ldots, f_{20}) | \mathrm{f}_n = 20/420 \times (count(A_i) \sum_{i=1}^{L} count(A_i) \} \quad (2)$$

where $count(A_i)$ is the number of A_i in peptide p. $Feature_{AAC}$ consists of 20 vectors, and these vectors represent the weighted frequency of "G," "A," "V," "L," "I," "P," "F," "Y," "W," "S," "T," "C," "M," "N," "Q," "D," "E," "K," "R" and "H."

(ii) 400-dimensional dipeptide composition (DPC)

The weighted frequency of the DPC is defined as follows:

$$Feature_{DPC} = \{ (f_1, f_2, \ldots, f_i, \ldots, f_{400}) | \mathrm{f}_n = 400/420 \times (count(A_iA_j) \sum_{i=1}^{L} count(A_iA_j) \} \quad (3)$$

where $count(A_iA_j)$ represents the number of amino acid residue pairs that consist of A_i and A_j. $Feature_{DPC}$ includes 400 vectors. These vectors represent the weighted frequencies of {"GG," "GA,"..., "GH," "AG," "AA,"..., "HR" and "HH"}.

Feature Selection

Generally, the extracted discrete features cannot be directly used in the training of the recognition model because there is noise in them (Yan et al., 2019). Therefore, after feature extraction,

[1] http://i.uestc.edu.cn/sarotup3/cgi-bin/PSBinder.pl

we need to use feature selection algorithms to filter the optimal features (Malysiak-Mrozek et al., 2018b). This process is also often considered to be a feature dimensionality reduction process in which noisy features are removed. In this paper, we use ANOVA and the IFS (incremental feature selection) strategy to rank and select the optimal feature set. First, all the extracted features are ranked by their ANOVA scores, and then optimal feature set is selected via incremental feature selection according to a certain criterion (Tang et al., 2019a,b).

(i) ANOVA

The training dataset is composed of positive and negative samples. Thus, each feature can naturally be divided into two groups, that is, the positive group and the negative group. If the difference between the positive and negative groups of a feature is large, then the discriminative ability is good. In ANOVA, the mean square between (MSB) groups and the mean square within (MSW) groups are used to measure the discriminative ability of a feature (Li B. et al., 2017). The MSB groups and the MSW groups of the ξth feature are calculated as follows:

$$MSB^2(\xi) = \sum_{i=1}^{2} m_i \left(\frac{\sum_{j=1}^{m_i} fea_\xi(i,j)}{m_i} - \frac{\sum_{i=1}^{2}\sum_{j=1}^{m_i} fea_\xi(i,j)}{\sum_{i=1}^{2} m_i}\right)^2 \quad (4)$$

$$MSW^2(\xi) = \sum_{i=1}^{2}\sum_{j=1}^{m_i} \left(fea_\xi(i,j) - \frac{\sum_{j=1}^{m_i} fea_\xi(i,j)}{m_i}\right)^2 \quad (5)$$

where m_i is total number of samples in the ith group. $fea_\xi(i,j)$ represents the value of the jth sample in the ith group of the ξth feature. $MSB^2(\xi)$ and $MSW^2(\xi)$ follow a chi-square distribution with 1 and $\sum_{i=1}^{k} m_i - 2$ degrees of freedom, respectively.

$$MSB^2(\xi) \sim \chi^2(1) \quad (6)$$

$$MSW^2(\xi) \sim \chi^2\left(\sum_{i=1}^{k} m_i - 2\right) \quad (7)$$

From eqs 6 and 7, can deduce the following equation:

$$F(\xi) = \frac{MSB^2(\xi)/1}{MSW^2(\xi)/\sum_{i=1}^{2} m_i - 2} \sim F\left(1, \sum_{i=1}^{2} m_i - 2\right) \quad (8)$$

$F(\xi)$ follows an F-distribution with $(1, \sum_{i=1}^{2} m_i - 2)$ degrees of freedom. The larger $F(\xi)$ is, the greater the contribution of the ξth feature to the classification is.

(ii) Incremental feature selection

All the features are sorted in descending order after calculating eq. 8. The feature sets are generated by adding one new feature at a time as follows:$[fea_1']$, $[fea_1', fea_2'] \ldots \left[fea_1', fea_2' \ldots, fea_{n-1}'\right]$and $\left[fea_1', fea_2' \ldots, fea_{n-1}', fea_n'\right]$. The classification models are generated using the above new feature sets, and the best model is selected according to some criteria, such as the accuracy, F1 score or another.

Support Vector Machine

A support vector machine (SVM) is a kind of generalized linear classifier that classifies data via supervised learning. The SVM maps labeled data to a high-dimensional space and then uses the maximum-margin hyperplane to classify those data. In addition, the SVM is also one of the common kernel learning methods for non-linear classification (Yang et al., 2019a). In recent years, SVMs have been successfully applied in bioinformatics fields (Xiong et al., 2012, 2019; Zhang et al., 2015; Zhang J. et al., 2019; Ding et al., 2016a,b; Wei et al., 2016; Zeng et al., 2017; Zhao et al., 2017; Bu et al., 2018; Xu et al., 2018c; Hu et al., 2019; Liu and Li, 2019; Liu et al., 2019b; Wang et al., 2019; Dou et al., 2020). The LIBSVM is a widely used SVM tool. In addition to the standard SVM algorithm, LIBSVM also includes a support vector regression, multiple classifiers and probability output functions. The source code of LIBSVM is written using C, and it provides a call interface for the mainstream development languages including Java, Python, R and MATLAB. In this paper, the radial basis function (RBF) is used as the kernel function of the SVM. In addition, the grid.py program is used to find the kernel width parameter γ and the penalty constant C that optimize the model. In this paper, the search range of log_2^γ is set to [6, 20] and the step size is −0.5. Similarly, the search range of $log_2^C$is [−10, 20], and the step size is 0.5. We use LIBSVM version 3.24, and it can be downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Evaluation Measurement

K-fold cross validation, leave-one-out cross-validation (LOOCV) and independent tests are three major validation methods. In this paper, we use five-fold cross validation to evaluate and compare the different identifiers (Jiang et al., 2013; Ding et al., 2017; Wei et al., 2017a,b,c, 2019; Chu et al., 2019; Liu et al., 2019c,d; Shan et al., 2019; Xu et al., 2019c; Zeng et al., 2019a,c; Zhang X. et al., 2019). five-fold cross validation first divides the whole training dataset into five parts. Then, this validation selects four parts to train the model, and the remaining part is used for testing. The above process iterates until all five subsets are used as test datasets. Finally, the five groups of evaluation metric scores are averaged to evaluate the trained model's performance. To evaluate the model's performance, we employ the sensitivity (SN), specificity (SP) and accuracy (ACC) to compare the different models. It is worth mentioning that ACC is also used as the objective of model optimization. These evaluation metrics are defined as follows:

$$SN = \frac{TP}{TP + FN} \quad (9)$$

$$SP = \frac{TN}{TN + FP} \quad (10)$$

$$ACC = \frac{TN + TP}{TN + FP + FN + TP} \quad (11)$$

where TN represents true negatives, and TP represents true positives. FN and FP represent false negatives and false positives, respectively.

In addition, the area under the curve (AUC) is also used to evaluate the overall performance of the model. The AUC is the value of the area enclosed by the X, Y coordinates and the receiver operating characteristic curve (ROC curve). The AUC reflects the performance stability of the model. The greater the AUC is, the better the stability of the model.

RESULTS AND DISCUSSION

The Framework of the Proposed PSBP-SVM Identifier

There are four steps in the process of constructing our proposed identifier. As illustrated in **Figure 1**, these steps are data collection, feature extraction, feature selection and model generation and optimization. In the data collection step, the positive and negative samples are collected as described in the "data collection" section. The 420-dimensional AAC and DPC feature is generated from the above benchmark dataset in the feature extraction step. Then, the resulting feature vectors are ranked via their ANOVA scores and a 123-dimensional optimal feature set (123D optimal set) is selected via the IFS process using the ACC as the criterion. This optimal feature set is input into the SVM classifier to train and optimize the model. Finally, the proposed identifier is obtained and called the PSBP-SVM. "PSBP" refers to PSBPs, and the SVM is applied as the classification algorithm.

The identification of a peptide is as follows. ①The 420-dimensional (AAC + DPC) feature is extracted from this peptide. ② Then, we select the feature vectors from the above feature according to the optimal feature set. ③ ④Finally, the selected feature vectors are put into the proposed model (PSBP-SVM) to identify whether a peptide is a PSBP or not.

Comparison With Other Identifiers

To comprehensively investigate the performance of the PSBP-SVM, we compare it with other identifiers including the state-of-the-art identifier. All models presented in this section have been optimized. The optimization conditions of SVM related models are the same as PSBP-SVM.

The 188-bit (Wei et al., 2018) and Izlti (Diener et al., 2016) feature extraction algorithms are combined with the SVM classifier to generate the 188D_SVM and Iztli_SVM, respectively. The comparison of the PSBP-SVM with the 188D_SVM and Iztli_SVM is illustrated in **Figure 2A**. In the five-fold cross validation test, the PSBP-SVM achieves 90.38, 84.62, 87.50, and 0.90% SN, SP, ACC, and AUC, respectively. It is observed that the PSBP-SVM is better than the other two identifiers by approximately 20% in terms of the SN, SP, ACC and AUC. This finding demonstrates that the 188-bit and Iztli extraction features might not include important discriminative features of PSBPs and non-PSBPs compared with the 123-dimensional optimal feature set.

To investigate the effectiveness of the SVM classifier, the naive Bayes, random forest and J48 are used to train the identifiers on the 123D optimal set. The generated identifiers of these algorithms are named 123D_NB, 123D_RF and 123D_J48, respectively. From **Figure 2B**, it is observed that the PSBP-SVM still outperforms the other three identifiers. The performance of 123D_NB follows. The SN, SP, ACC and AUC of 123D_NB are 84.6, 83.7, 84.13, and 0.897%, respectively, which are 5.78, 0.92, 3.37 and 0.003% lower than those of the PSBP-SVM, respectively. 123D_J48 is the worst of all. 123D_J48 only exhibits 47.1, 70.2, 58.65, and 0.641% SN, SP, ACC and AUC, respectively. In particular, the SN of 123D_J48 is below that of random classification. The performance of 123D_RF is worse than that of 123D_NB and better than that of 123D_J48. Thus, it can be concluded that the SVM classifier performs better than other classifier on the 123D optimal feature set.

Different feature selection algorithms lead to different classification effects. **Figure 2C** represents the influence of two different feature selection algorithms on the model. The MRMD-SVM is generated by replacing part C of **Figure 1** with MRMD, that is, MRMD is used as the feature selection algorithm. Finally, a 178-dimensional new optimal feature set is selected by MRMD. From the comparison result, we observe that the MRMD-SVM only achieves 66.3, 68.3, 67.31, and 0.673% in terms of the SN, SP, ACC, and AUC, respectively. The performance of MRMD-SVM is much worse than that of the PSBP-SVM. This result indicates that MRMD may not select important features from the 420-dimensional feature set (Hong et al., 2019).

As shown in **Figure 2D**, the SN, SP, ACC and AUC values of PSBinder are 88.46, 85.58, 87.02 and 0.91%, respectively, according to the five-fold cross validation test. The SN and ACC of the PSBP-SVM are higher than those of PSBinder by 1.92 and 0.48%, respectively, although the other two metrics are slightly lower. It is worth mentioning that the number of features for the PSBP-SVM is 123, which is smaller than the 146 of PSBinder. Therefore, the PSBP-SVM can effectively avoid overfitting problems compared with PSBinder. For the computing model, the SN value is more significant because it can improve the positive sample identification accuracy by reducing its scope.

Feature Contribution and Importance Analysis

Figure 3A shows that the ACC values vary with the incremental feature selection process. When the top 123 features are selected, the ACC reaches the highest value of 87.5%. This is also the reason why 123 features are chosen as training features. The analysis of the composition of these optimal features is represented in **Figure 3B**. It is found that there are 8 AAC

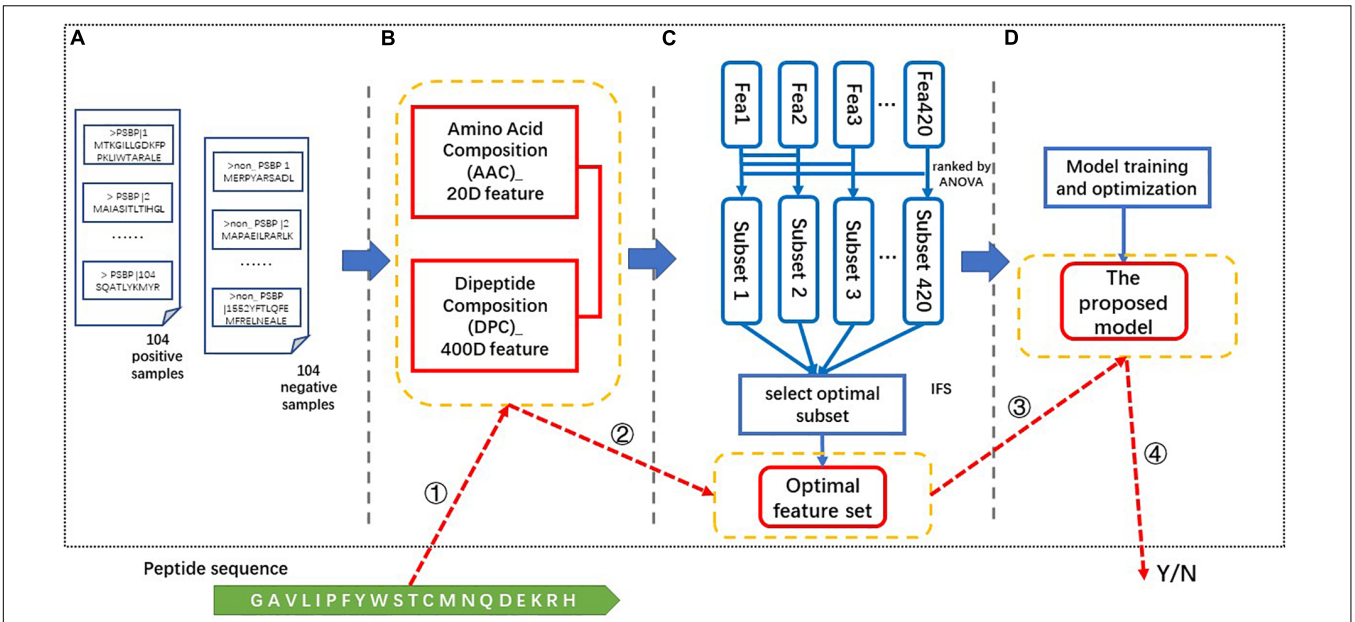


FIGURE 1 | The framework and identification process of the PSBP-SVM. **(A)** Data collection. The benchmark dataset consists of 104 positive samples and 104 negative samples. **(B)** Feature extraction. A 420-dimensional feature is extracted from the benchmark dataset. **(C)** Feature selection. The optimal feature set is generated by the ANOVA ranking algorithm and the IFS process. **(D)** Model training and optimization. The optimal feature set is used to train and optimize the model. The PSBP identification process is based on the four parts in yellow boxes. These parts are ① feature extraction, ② feature selection, ③ PSBP identification, and ④ the result.

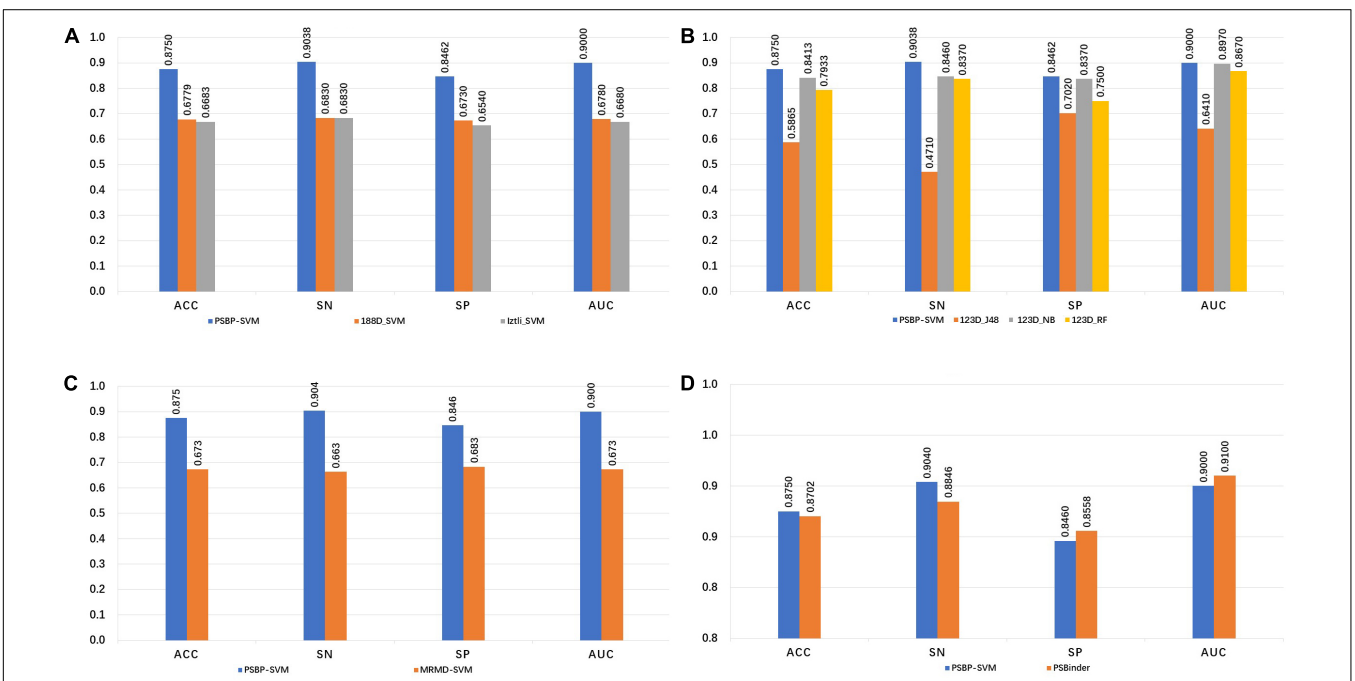


FIGURE 2 | Comparison of the PSBP-SVM and other identifiers. **(A)** Comparison with other feature extraction identifiers. **(B)** Comparison with other classification algorithm identifiers. **(C)** Comparison with other feature selection identifiers. **(D)** Comparison with the state-of-the-art identifier.

features and 115 DPC features, respectively accounting for 40 and 28.75% of the original features. This finding indicates that the AAC features have higher participation rates. Furthermore, the appearance frequencies of 20 amino acids are calculated using the AAC and DPC separately. From the result shown in **Figure 3C**, we can observe that the top six amino acids

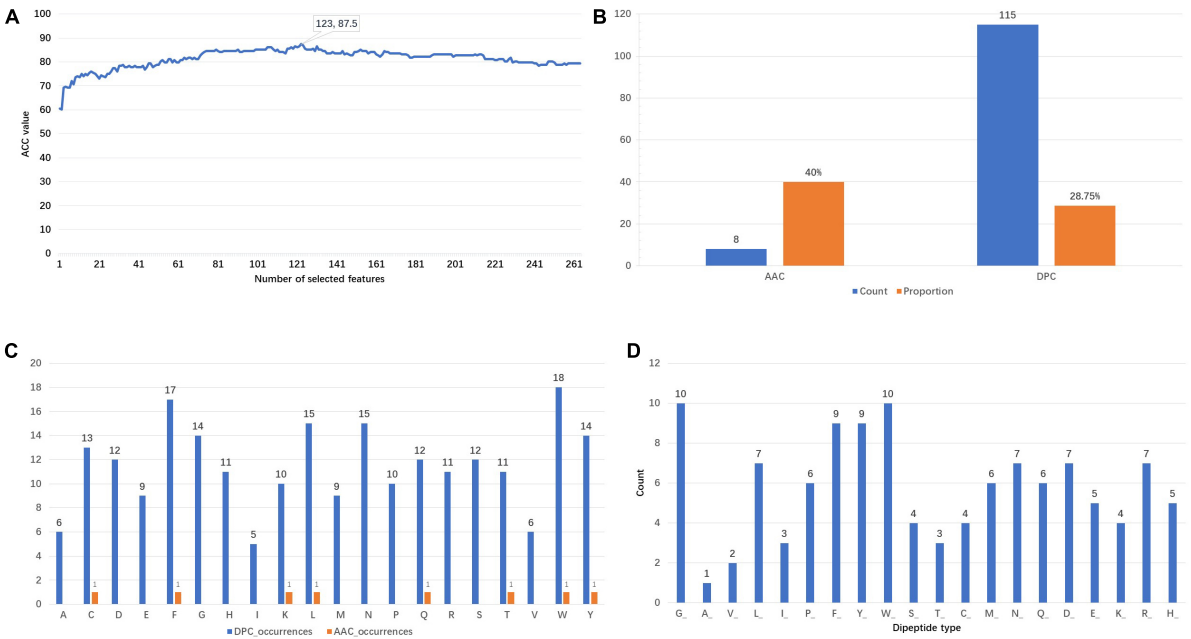


FIGURE 3 | Analysis of the 123D optimal feature set. **(A)** Plot of the accuracy of incremental feature selection. **(B)** Composition of the optimal feature set. **(C)** DPC and AAC occurrences. **(D)** Number of dipeptide types in the DPC.

both in the AAC and DPC are tryptophan (W), phenylalanine (F), leucine (L), tyrosine (Y), cysteine (C) and glutamine (Q). The counts of the dipeptide types are presented in **Figure 3D**. The dipeptides that begin with glycine (G), tryptophan (W), phenylalanine (F) and tyrosine (Y) are the top four dipeptide types in the 123D optimal feature set. From the above analysis, we can conclude that tryptophan (W), phenylalanine (F) and tyrosine (Y) play important roles in identifying PSBPs from non-PSBPs.

Web Server Guidelines

For the convenience of other researchers, we have constructed a web server including the PSBP-SVM, and free access is provided at http://server.malab.cn/PSBP-SVM/index.jsp. This web server includes "Home," "Dataset," "About" and "Contact us" pages. One can enter a sequence into the input box of the "Home" page and click the "submit" button to identify whether it is a PSBP or not. Note that only the FASTA format is supported. The "Dataset" page provides a link to download positive and negative samples. The "About" and "Contact us" pages give related information about our proposed model and the authors, respectively.

CONCLUSION

In this study, we proposed a novel SVM-based polystyrene binding peptide identification model and incorporated it in an identifier called the PSBP-SVM. The construction process of this model includes feature extraction, feature selection, model training and optimization. The performance comparison shows that the PSBP-SVM outperforms other identifiers, including the state-of-the-art identifier. Furthermore, in order to investigate the contribution of features, we comprehensively analyzed the composition and importance of the optimal feature set used in model training. However, there is still room for improvement in the future. With the help of multiview learning, ensemble learning strategies (Liu and Zhu, 2019; Ru et al., 2019; Zeng et al., 2019b) and evolutionary optimization (Xu et al., 2019a,b), the accuracy can be improved, and the range of the effective features can be further reduced.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://server.malab.cn/PSBP-SVM/data.jsp.

AUTHOR CONTRIBUTIONS

CM and YH wrote the manuscript, participated in the research design and developed the web server. YZ and FG participated in preparation of the manuscript. CM, YH, FG, and YZ read and approved the final manuscript.

FUNDING

The work was supported by the National Key R&D Program of China (2018YFC0910405) and the National Natural Science Foundation of China (Nos. 61772362 and 61801147).

REFERENCES

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* 26, 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Bakhshinejad, B., and Sadeghizadeh, M. (2016). A polystyrene binding target-unrelated peptide isolated in the screening of phage display library. *Anal. Biochem.* 512, 120–128. doi: 10.1016/j.ab.2016.08.013

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief. Bioinform.* doi: 10.1093/bib/bbz118 [Epub ahead of print].

Bu, H. D., Hao, J. Q., Guan, J. H., and Zhou, S. G. (2018). Predicting enhancers from multiple cell lines and tissues across different developmental stages based on SVM method. *Curr. Bioinform.* 13, 655–660. doi: 10.2174/1574893613666180726163429

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

Chu, Y., Kaushik, A. C., Wang, X., Wang, W., Zhang, Y., Shan, X., et al. (2019). DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Brief. Bioinform.* bbz152. doi: 10.1093/bib/bbz152

Diener, C., Martínez, G. G. R., Blas, D. M., González, D. A. C., Corzo, G., Castro-Obregon, S., et al. (2016). Effective design of multifunctional peptides by combining compatible functions. *PLoS Comput. Biol.* 12:e1004786. doi: 10.1371/journal.pcbi.1004786

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3, 185–205. doi: 10.1142/s0219720005001004

Ding, Y., Tang, J., and Guo, F. (2016a). Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information. *Int. J. Mol. Sci.* 17:1623. doi: 10.3390/ijms17101623

Ding, Y., Tang, J., and Guo, F. (2016b). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinformatics* 17:398. doi: 10.1186/s12859-016-1253-9

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

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

Dou, L., Li, X., Ding, H., Xu, L., and Xiang, H. (2020). Is there any sequence feature in the RNA pseudouridine modification prediction problem? *Mol. Ther. Nucleic Acids* 19, 293–303. doi: 10.1016/j.omtn.2019.11.014

Engvall, E., and Perlmann, P. (1971). "Enzyme-linked immunosorbent assay (ELISA)," in *Proceedings of the Twenty-Second Colloquium Protides of the Biological Fluids*, (London: Pergamon Press), 553–556.

Eş, I., Vieira, J. D. G., and Amaral, A. C. (2015). Principles, techniques, and applications of biocatalyst immobilization for industrial application. *Appl. Microbiol. Biotechnol.* 99, 2065–2082. doi: 10.1007/s00253-015-6390-y

Fang, T., Zhang, Z., Sun, R., Zhu, L., He, J., Huang, B., et al. (2019). RNAm5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition. *Mol. Ther. Nucleic Acids* 18, 739–747. doi: 10.1016/j.omtn.2019.10.008

Fu, J., Tang, J., Wang, Y., Cui, X., Yang, Q., Hong, J., et al. (2018). Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front. Pharmacol.* 9:681. doi: 10.3389/fphar.2018.00681

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics* 19:306. doi: 10.1186/s12859-018-2321-0

Hong, J., Luo, Y., Mou, M., Fu, J., Zhang, Y., Xue, W., et al. (2019). Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief. Bioinform.* bbz120. doi: 10.1093/bib/bbz120

Hu, Y., Zhao, T., Zhang, N., Zang, T., Zhang, J., and Cheng, L. (2019). Identifying diseases-related metabolites using random walk. *BMC Bioinformatics* 19(Suppl. 5):116. doi: 10.1186/s12859-018-2098-1

Jiang, Q. H., Wang, G. H., Jin, S. L., Li, Y., and Wang, Y. D. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/ijdmb.2013.056078

Jin, S., Wang, L., Guo, F., and Zou, Q. (2019). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7:224. doi: 10.3389/fbioe.2019.00224

Kumada, Y., Kuroki, D., Yasui, H., Ohse, T., and Kishimoto, M. (2010). Characterization of polystyrene-binding peptides (PS-tags) for site-specific immobilization of proteins. *J. Biosci. Bioeng.* 109, 583–587. doi: 10.1016/j.jbiosc.2009.11.005

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, W162–W170. doi: 10.1093/nar/gkx449

Li, N., Kang, J., Jiang, L., He, B., Lin, H., and Huang, J. (2017). PSBinder: a web service for predicting polystyrene surface-binding peptides. *Biomed Res. Int.* 2017:5761517. doi: 10.1155/2017/5761517

Li, Y. H., Li, X. X., Hong, J. J., Wang, Y. X., Fu, J. B., Yang, H., et al. (2019). Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief. Bioinform.* doi: 10.1093/bib/bby130 [Epub ahead of print].

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, D1121–D1127. doi: 10.1093/nar/gkx1076

Liao, Z. J., Li, D. P., Wang, X. R., Li, L. S., and Zou, Q. (2018). Cancer diagnosis through IsomiR expression with machine learning method. *Curr. Bioinform.* 13, 57–63. doi: 10.2174/1574893611666160609081155

Liu, B. (2019). BioSeq-analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019a). 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:e127. doi: 10.1093/nar/gkz740

Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., Zou, Q., et al. (2019b). Computational methods for identifying the critical nodes in biological networks. *Brief. Bioinform.* bbz011. doi: 10.1093/bib/bbz011

Liu, B., Li, C., and Yan, K. (2019c). DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinform.* bbz098. doi: 10.1093/bib/bbz098

Liu, B., Zhu, Y., and Yan, K. (2019d). Fold-LTR-TCP: protein fold recognition based on triadic closure principle. *Brief. Bioinform.* bbz139. doi: 10.1093/bib/bbz139

Liu, B., and Li, K. (2019). iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids* 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Liu, B., and Zhu, Y. (2019). ProtDec-LTR3.0: protein remote homology detection by incorporating profile-based features into Learning to Rank. *IEEE Access* 7, 102499–102507. doi: 10.1109/access.2019.2929363

Małysiak-Mrozek, B., Baron, T., and Mrozek, D. (2018a). Spark-IDPP: high-throughput and scalable prediction of intrinsically disordered protein regions with Spark clusters on the Cloud. *Cluster Comput.* 22, 487–508. doi: 10.1007/s10586-018-2857-9

Malysiak-Mrozek, B., Stabla, M., and Mrozek, D. (2018b). Soft and declarative fishing of information in big data lake. *IEEE Trans. Fuzzy Syst.* 26, 2732–2747. doi: 10.1109/tfuzz.2018.2812157

Meng, C., Wei, L., and Zou, Q. (2019). SecProMTB: support vector machine-based classifier for secretory proteins using imbalanced data sets applied to *Mycobacterium tuberculosis*. *Proteomics* 19:1900007. doi: 10.1002/pmic.201900007

Modjarrad, K. (2013). *Handbook of Polymer Applications in Medicine and Medical Devices*, 1st Edn. Amsterdam: Elsevier.

Moritomi, S., Watanabe, T., and Kanzaki, S. (2010). Polypropylene compounds for automotive applications. *Sumitomo Kagaku* 1, 1–16.

Mrozek, D., and Malysiak-Mrozek, B. (2011). An improved method for protein similarity searching by alignment of fuzzy energy signatures. *Int. J. Comput. Intell. Syst.* 4, 75–88. doi: 10.2991/ijcis.2011.4.1.7

Pan, L., Lei, Y., Wang, C., and Xie, J. (2009). Method on entity identification using similarity measure based on weight of Jaccard. *J. Beijing Jiaotong Univ.* 34, 141–145.

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

Shan, X., Wang, X., Li, C. D., Chu, Y., Zhang, Y., Xiong, Y., et al. (2019). Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method. *J. Chem. Inf. Model.* 59, 4577–4586. doi: 10.1021/acs.jcim.9b00749

Shen, H.-B., and Chou, K.-C. (2008). PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* 373, 386–388. doi: 10.1016/j.ab.2007.10.012

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

Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., et al. (2019a). ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief. Bioinform.* doi: 10.1093/bib/bby127 [Epub ahead of print].

Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., et al. (2019b). Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol. Cell. Proteomics* 18, 1683–1699. doi: 10.1074/mcp.RA118.001169

Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151. doi: 10.1093/nar/gkx1096

Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J. Y., Zhao, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics* 9(Suppl. 2):S22. doi: 10.1186/1471-2164-9-S2-S22

Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., and Liu, Y. (2010). Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells. *PLoS One* 5:e11794. doi: 10.1371/journal.pone.0011794

Wang, Y., Shi, F. Q., Cao, L. Y., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinform.* 14, 282–294. doi: 10.2174/1574893614666190304125221

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, D1031–D1041. doi: 10.1093/nar/gkz981

Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2018). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Brief. Bioinform.* 10, 1–14. doi: 10.1093/bib/bby107

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z., and Zou, Q. (2017b). CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017c). Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast prediction of methylation sites using sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1264–1273. doi: 10.1109/TCBB.2017.2670558

Wei, L., Zou, Q., Liao, M., Lu, H., and Zhao, Y. (2016). A novel machine learning method for cytokine-receptor interaction prediction. *Comb. Chem. High Throughput Screen.* 19, 144–152. doi: 10.2174/1386207319666151110122621

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Sci.* 10(Suppl. 1):S20. doi: 10.1186/1477-5956-10-S1-S20

Xiong, Y., Qiao, Y., Kihara, D., Zhang, H. Y., Zhu, X., and Wei, D. Q. (2019). Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates. *Curr. Drug Metab.* 20, 229–235. doi: 10.2174/1389200219666181019094526

Xu, H., Zeng, W., Zeng, X., and Yen, G. G. (2019a). An evolutionary algorithm based on minkowski distance for many-objective optimization. *IEEE Trans. Cybern.* 49, 3968–3979. doi: 10.1109/tcyb.2018.2856208

Xu, H., Zeng, W., Zhang, D., and Zeng, X. (2019b). MOEA/HD: a multiobjective evolutionary algorithm based on hierarchical decomposition. *IEEE Trans. Cybern.* 49, 517–526. doi: 10.1109/TCYB.2017.2779450

Xu, L., Liang, G., Liao, C., Chen, G.-D., and Chang, C.-C. (2019c). k-Skip-n-Gram-RF: a random forest based method for Alzheimer's disease protein identification. *Front. Genet.* 10:33. doi: 10.3389/fgene.2019.00033

Xu, L., Liang, G., Liao, C., Chen, G.-D., and Chang, C.-C. (2018a). An efficient classifier for Alzheimer's disease genes identification. *Molecules* 23:3140. doi: 10.3390/molecules23123140

Xu, L., Liang, G., Shi, S., and Liao, C. (2018b). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Int. J. Mol. Sci.* 19:1773. doi: 10.3390/ijms19061773

Xu, L., Liang, G., Wang, L., and Liao, C. (2018c). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* 9:E158. doi: 10.3390/genes9030158

Yaman, N., Özdoğan, E., Seventekin, N., and Ayhan, H. (2009). Plasma treatment of polypropylene fabric for improved dyeability with soluble textile dyestuff. *Appl. Surf. Sci.* 255, 6764–6770. doi: 10.1016/j.apsusc.2008.10.121

Yan, K., Fang, X., Xu, Y., and Liu, B. (2019). Protein fold recognition based on multi-view modeling. *Bioinformatics* 35, 2982–2990. doi: 10.1093/bioinformatics/btz040

Yang, Q., Hong, J., Li, Y., Xue, W., Li, S., Yang, H., et al. (2019a). A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies. *Brief. Bioinform.* bbz137. doi: 10.1093/bib/bbz137

Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., Li, X., et al. (2019b). Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief. Bioinform.* bbz049. doi: 10.1093/bib/bbz049

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:D1171. doi: 10.1093/nar/gkz878

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and validation of disease genes using HeteSim scores. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 687–695. doi: 10.1109/tcbb.2016.2520947

Zeng, X., Wang, W., Chen, C., and Yen, G. (2019a). A consensus community-based particle swarm optimization for dynamic community detection. *IEEE Trans. Cybern.* doi: 10.1109/TCYB.2019.2938895 [Epub ahead of print].

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2019b). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief. Bioinform.* bbz080. doi: 10.1093/bib/bbz080

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019c). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zhang, C., Song, X., Zhao, Y., Zhang, H., Zhao, S., Mao, F., et al. (2015). *Mycobacterium tuberculosis* secreted proteins as potential biomarkers for the diagnosis of active tuberculosis and latent tuberculosis infection. *J. Clin. Lab. Anal.* 29, 375–382. doi: 10.1002/jcla.21782

Zhang, J., Chen, Q., and Liu, B. (2019). DeepDRBP-2L: a new genome annotation predictor for identifying DNA-binding proteins and RNA-binding proteins using Convolutional neural network and long short-term memory. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2952338 [Epub ahead of print].

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. (2019). Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 283–291. doi: 10.1109/TCBB.2017.2776280

Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network. *Biomed Res. Int.* 2017:7049406. doi: 10.1155/2017/7049406

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2019). Analysis of long non-coding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Brief. Bioinform.* 20, 598–608. doi: 10.1093/bib/bby021

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

Zhu, X., He, J., Zhao, S., Tao, W., Xiong, Y., and Bi, S. (2019). A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of *Saccharomyces cerevisiae*. *Brief. Funct. Genomics* 18, 367–376. doi: 10.1093/bfgp/elz018

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* 21, 1–10.

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.

Copyright © 2020 Meng, Hu, Zhang and Guo. 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 Genetics

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fgene.2020.00278

A Systematic Way to Infer the Regulation Relations of miRNAs on Target Genes and Critical miRNAs in Cancers

Peng Xu[1,2], Qian Wu[1], Jian Yu[1], Yongsheng Rao[1], Zheng Kou[1], Gang Fang[1], Xiaolong Shi[1], Wenbin Liu[1] and Henry Han[3]**

[1] *Institute of Computational Science and Technology, Guangzhou University, Guangzhou, China,* [2] *School of Computer Science of Information Technology, Qiannan Normal University for Nationalities, Duyun, China,* [3] *Department of Computer and Information Science, Fordham University, New York, NY, United States*

MicroRNAs (miRNAs) are a class of important non-coding RNAs, which play important roles in tumorigenesis and development by targeting oncogenes or tumor suppressor genes. One miRNA can regulate multiple genes, and one gene can be regulated by multiple miRNAs. To promote the clinical application of miRNAs, two fundamental questions should be answered: what's the regulatory mechanism of a miRNA to a gene, and which miRNAs are important for a specific type of cancer. In this study, we propose a miRNA influence capturing (miRNAInf) to decipher regulation relations of miRNAs on target genes and identify critical miRNAs in cancers in a systematic approach. With the pair-wise miRNA/gene expression profiles data, we consider the assigning problem of a miRNA on target genes and determine the regulatory mechanisms by computing the Pearson correlation coefficient between the expression changes of a miRNA and that of its target gene. Furthermore, we compute the relative local influence strength of a miRNA on its target gene. Finally, integrate the local influence strength and target gene's importance to determine the critical miRNAs involved in specific cancer. Results on breast, liver and prostate cancers show that positive regulations are as common as negative regulations. The top-ranked miRNAs show great potential as therapeutic targets driving cancer to a normal state, and they are demonstrated to be closely related to cancers based on biological functional analysis, drug sensitivity/resistance analysis and survival analysis. This study will be helpful for the discovery of critical miRNAs and development of miRNAs-based clinical therapeutics.

Keywords: miRNA, influence, cancer, target gene, regulation relation

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Guimin Qin,
Xidian University, China
Xuequn Shang,
Northwestern Polytechnical University, China

***Correspondence:**
Wenbin Liu
wbliu6910@126.com
Henry Han
xhan9@fordham.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *14 January 2020*
Accepted: *09 March 2020*
Published: *31 March 2020*

Citation:
Xu P, Wu Q, Yu J, Rao Y, Kou Z, Fang G, Shi X, Liu W and Han H (2020) A Systematic Way to Infer the Regulation Relations of miRNAs on Target Genes and Critical miRNAs in Cancers. Front. Genet. 11:278. doi: 10.3389/fgene.2020.00278

INTRODUCTION

MicroRNAs (miRNAs) are a class of small non-coding RNAs and have been proved to play important roles in regulating more than two thirds of human genes (Bandyopadhyay et al., 2010; Song et al., 2017). They usually regulate their target genes by binding to the complementary seed sequence at the 3′ untranslated region. The binding of miRNAs usually leads to the translation

repression or degradation of the target mRNAs and ultimately affects the production of the corresponding proteins (Bartel, 2009; Fabian et al., 2010; Hata and Lieberman, 2015). For example, miR-21 was demonstrated to negatively regulate the expression of SAV1 (Sun et al., 2019) and Smad6 (Xu et al., 2016) in colorectal cancer. One single miRNA usually targets many genes and one gene might be regulated by multiple miRNAs. To decipher the relationships between miRNAs and their target genes and unveil miRNAs' biological functions, many miRNA targets databases, such as Targetscan (Agarwal et al., 2015), miRDB (Wong and Wang, 2015), miRanda (Betel et al., 2010), and mirTarbase (Chou et al., 2018), have been built based on various biological experiments and/or different computation methods.

The dysfunctions of miRNAs have been reported to be involved in the tumorigenesis of various cancers (Bartel, 2004; Gotte, 2010; Dela Cruz and Matushansky, 2011; Lovat et al., 2011; Liu W. et al., 2018; Xu et al., 2018). For this reason, miRNAs have become potential biomarkers in cancer diagnosis and treatment (Slack and Chinnaiyan, 2019). Furthermore, some miRNA-based therapeutics have entered into the clinical research, i.e., miR-16-based mimics in phase I clinical trial for treating advanced non-small cell lung cancer, and antimiRs targeted at miR-122 in phase II trial for treating hepatitis (Rupaimoole and Slack, 2017).

However, accumulating evidence indicates that miRNAs can also promote the expression of their target genes. For example, Vasudevan et al. found that miR396-3 could direct the AGO complex binding with the AU-rich elements to promote the translation of its target gene in Vasudevan et al. (2007). They further demonstrated that let-7 and synthetic miRcxcr4 could induce target mRNAs up-regulation on cell cycle arrest while repressing translation in proliferation cells (Vasudevan and Steitz, 2007). In addition to functioning in the cytoplasm, mature miRNAs are also found in the nucleus. Xiao et al. demonstrated that miR-24-1 in the nucleus can activate gene transcription by targeting their enhancers (Xiao et al., 2017). Up to now, more than 200 positive regulations of miRNAs on genes have been experimentally identified in the literature.

It becomes a fundamental problem to elucidate the regulatory relations between miRNAs and their target genes in systems biology. Specifically, we need to know which genes are positively regulated by one miRNA and which genes are negatively regulated by it. The answer to this problem will provide a foundation to study the critical roles of miRNAs in tumorigenesis. Recently, Tan et al. *first* investigated this problem based on the Pearson correlation coefficients between the expression of miRNAs and their target genes in pan-cancer datasets (Tan et al., 2019). Surprisingly, they found many positive correlated miRNA-gene pairs. This demonstrates that miRNAs could exert their important roles in various cancers by positively regulating many genes.

Another important issue is to determine the critical miRNAs potential to affect the overexpression or under-expression of cancer-related genes. The answer to this question will help to determine a few "level point" miRNAs for designing miRNA-based therapeutic strategies. Cui et al. combined the miRNA sequence features and miRNA disease spectrum width (DSW) to define the importance of miRNAs (Cui et al., 2019). However, this static definition could not reflect the different regulatory mechanisms between miRNAs and genes involved in specific cancer well.

In this paper, we propose a novel miRNA influence capturing (MiRNAInf) to decipher regulation relations of miRNAs on target genes and identify critical miRNAs in cancers in a systematic approach. We study miRNA-gene regulations by assuming that the expression of one gene is determined by its upstream miRNAs. We model the expression of a gene as a function of the expression of the miRNAs targeting it. Through the Taylor expansion, we employ the first partial derivative to a miRNA to denote its regulatory effect on the target gene. The first partial derivative is then approximated by the Pearson correlation coefficient of the expression change of a miRNA and that of its target gene between disease and normal control.

We finally define the global influence of a miRNA by combining its local influence strength in an individual cancer and the degree of its target gene in a PPI network. Our results on breast cancer, prostate adenocarcinoma and liver cancer datasets further demonstrate that positive regulations are as common as negative ones in miRNA-gene interactions. We also find that only a few miRNAs have significant influences on the cancer-related differentially expressed genes. The identified top miRNAs in the three datasets are not only highly correlated in a functional network, but also significantly enriched in some important functions such as inflammation, cell proliferation, apoptosis and cell cycle. It demonstrates that they are very likely to play essential roles coherently in tumorigenesis. More importantly, we find that the intervention of a few critical miRNAs may alleviate the abnormal expressions of most genes according to the regulatory effect and differential expression situation between miRNAs and their target genes. Besides, the identified important miRNAs influence patients' survival time of prognosis besides the sensitivity/resistance of some anti-cancer drugs. In sum, our study provides a systematic way to understand the key roles of miRNAs in cancers and to screen potential intervention miRNA biomarkers for future miRNA-based therapy and diagnosis in precision medicine.

MATERIALS AND METHODS

Data Acquisition and Preprocessing

We collected three types of data: miRNA and gene expression data for three cancers, miRNA-gene interaction data, and protein-protein interaction (PPI) data from three different databases: TCGA, miRTarbase and STRING (Chou et al., 2016; Szklarczyk et al., 2017).

The Cancer Genome Atlas (TCGA) Data

We collected datasets with both miRNA and gene expression profiles for cancer samples and corresponding normal samples in this study. Three cancers with abundant gene expression and miRNA expression "pair datasets" were acquired from TCGA (http://tcga-data.nci.nih.gov/tcga/). They included 102 samples for breast cancer, 52 samples for prostate adenocarcinoma, and 49 samples for liver cancer.

MiRNA-Gene Interaction Data

Besides gene expression and miRNA expression data, we further downloaded miRNA-gene targets data from miRTarbase (Chou et al., 2016), a widely-used state-of-the-art database for miRNA-gene targets. miRTarbase includes 502,652 high quality experimentally validated miRNA-gene interactions between 2,599 miRNAs and 15,064 genes for the human species.

Protein-Protein Interaction (PPI) Data

The STRING database, which includes 10,048,286 interactions between the 19,576 proteins for human beings, integrates the experimentally validated and computationally predicted protein-protein interactions (Szklarczyk et al., 2017). In order to use the highly confident interactions, we selected the interactions with a combined score >150. The distribution of proteins' degrees indicates that a small number of proteins have interactions with hundreds of other nodes, while most proteins only have interactions with a few of other proteins, which satisfies power-law distribution (Please go to **Figure S1** for details).

Methods

In this study, we propose a miRNAInf method to identify critical miRNAs involved in cancer as well as conducting comprehensive functional analysis for miRNAs. The flowchart of the proposed miRNAInf methods is illustrated in **Figure 1**. The proposed method consists of the three steps. First, we determine the significant differentially expressed miRNAs and genes based on their expression data from TCGA. Second, we compute the local influence strength of a miRNA to its target gene. Finally, we evaluate the global influence of a miRNA in a specific cancer by integrating the local influence strength and gene's importance.

Identify Differentially Expressed miRNAs and Genes

We conduct normalization for miRNA and gene expression data before identifying differentially expressed miRNAs and genes.

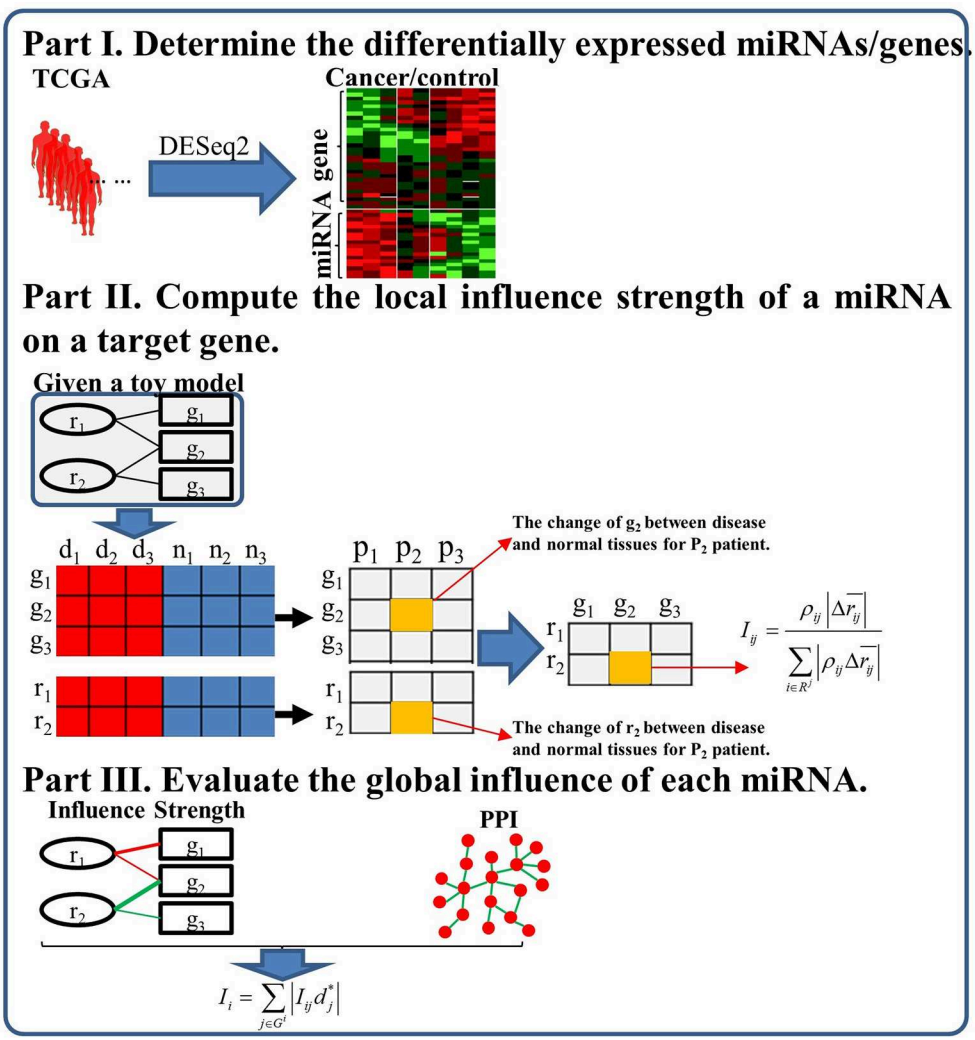


FIGURE 1 | The flowchart of the proposed study framework.

For each miRNA epression sample, we apply the RPM (Reads Per Million) method for normalization as described in Equation (1) for its simplicity and efficiency (Faraldi et al., 2019).

$$r_i^{RPM} = \frac{E_{r_i}}{Total_r} \times 10^6 \quad (1)$$

where E_{r_i} means the read counts of *miRNA i* and $Total_r$ indicates the total counts of all miRNAs in a specific sample. RPM normalizes all read counts with respect to the ratio between library size and a million number. Similarly, we normalize gene expression data as follows

$$g_j^{RPM} = \frac{E_{g_j}}{Total_g} \times 10^6 \quad (2)$$

where E_{g_j} means the read counts of *gene j* and $Total_g$ indicates the total counts of all genes in a specific sample.

We then apply DESeq2 (Love et al., 2014) to identify differentially expressed genes (DEGs) and differentially expressed miRNAs (DEmiRs) by collecting all genes and miRNAs with $p < 0.05$ adjusted by FDR method in differential expression analysis of the normalized data. Finally, we get 546,457 and 313 DemiRs and 5,057, 3,665, 3,064 DEGs for breast, liver, and prostate cancer, respectively.

Compute miRNA Local Influence Strength on a Target Gene

The expression of a gene can be assumed as a function of the expression of the miRNAs targeting it. Given a gene j regulated by m miRNAs $r_1, r_2, \ldots r_i, \ldots r_m$, then its expression $g_j = f^j(r_1, r_2, \ldots r_i, \ldots r_m)$ in disease state can be approximated by first-level Taylor expansion as

$$f^j(r_1, r_2, \ldots r_i, \ldots r_m) = f^j\left(r_0^1, r_0^2, \ldots r_0^i, \ldots r_0^m\right) + \sum_{i \in R^j} \frac{\partial f^j(r_1, r_2, \ldots r_i, \ldots r_m)}{\partial r_i} \Delta r_i \quad (3)$$

where $f^j\left(r_0^1, r_0^2, \ldots r_0^i, \ldots r_0^m\right)$ is the expression value of gene j in normal state, R^j is the index set of the miRNAs regulating gene j, $\Delta r_i = r_i - r_0^i$ ($i \in \{1, 2, \ldots, m\}$) is the change of miRNA i between the disease state and normal state.

The gene expression change Δg_j of gene j can be calculated by moving the gene expression in the normal state $f^j\left(r_0^1, r_0^2, \ldots r_0^i, \ldots r_0^m\right)$ to the left side of Equation 3:

$$\Delta g_j = f^j(r_1, r_2, \ldots r_i, \ldots r_m) - f^j\left(r_0^1, r_0^2, \ldots r_0^i, \ldots r_0^m\right) = \sum_{i \in R^j} \frac{\partial f^j(r_1, r_2, \ldots r_i, \ldots r_m)}{\partial r_i} \Delta r_i \quad (4)$$

where the right side represents the sum of expression change of gene j induced by the perturbation of each miRNA i. The partial derivative $\frac{\partial f^j(r_1, r_2, \ldots r_i, \ldots r_m)}{\partial r_i}$ actually reflects the influence strength of miRNA i on gene j. A given miRNA i may target many other genes, we assume that Δr_{ij}, the portion of the ith miRNA expression difference Δr_i between the disease and normal states, affecting gene j, is positively proportional to the expression change Δg_j. Given the target gene index set G^i of miRNA i, we define Δr_{ij} as the product of the absolute change ratio of gene j $\left|\Delta g_i\right| / \sum_{j \in G^i} \left|\Delta g_j\right|$ and the change of miRNA i between the disease state and normal state Δr_i:

$$\Delta r_{ij} = \frac{\left|\Delta g_j\right|}{\sum_{j \in G^i} \left|\Delta g_j\right|} \times \Delta r_i \quad (5)$$

Based on this, we can calculate the Pearson correlation coefficient ρ_{ij} between Δr_{ij} and Δg_j. $\rho_{ij} > 0$ indicates that miRNA i upregulate gene j, otherwise, it suppresses gene j. Then the partial derivative $\frac{\partial f^j(r_1, r_2, \ldots r_i, \ldots r_m)}{\partial r_i}$ can be approximated by

$$\frac{\partial f^j(r_1, r_2, \ldots r_i, \ldots r_m)}{\partial r_i} = k_{ij} \rho_{ij} \quad (6)$$

where the coefficient k_{ij} can be assumed and approximated by a constant k for all miRNA-genes.

Given one miRNA i, it may have different influence strengths on each target gene. Thus, we compute its local influence strength I_{ij} as follows

$$I_{ij} = \frac{\rho_{ij} \left|\overline{\Delta r_{ij}}\right|}{\sum_{i \in R^j} \left|\rho_{ij} \overline{\Delta r_{ij}}\right|} \quad (7)$$

where the $\overline{\Delta r_{ij}}$ means the average difference between disease and normal tissues for all patients. The local influence strength I_{ij} considers both the correlation coefficient ρ_{ij} and the average difference $\overline{\Delta r_{ij}}$. The higher correlation ρ_{ij} between them, the larger its influence on gene j. Similarly, the larger average difference $\overline{\Delta r_{ij}}$, the larger also its influence on gene j.

Evaluate the Global Influence of Each miRNA

In order to describe the importance of a miRNA in a specific disease, we consider both the number of its target genes and the importance of each gene. Here, we define the global influence of a miRNA i for the disease by weighting its local influence by the importance of its target genes:

$$I_i = \sum_{j \in G^i} \left| I_{ij} d_j^* \right| = \sum_{j \in G^i} \left| I_{ij} \frac{d_j}{d_{\max}} \right| \quad (8)$$

where d_j and $d_{\max}$ represent the degree of the gene j and the maximum degree in the PPI network, respectively. The importance of the gene j is modeled as the ratio of its degree between the maximum degree in the PPI: $\frac{d_i}{d_{\max}}$. The global influence of a miRNA i involves both the local influence strength and the importance of its target genes. The more targets regulated by miRNA i, the larger its global influence. Simultaneously, the larger the degrees of its target genes in PPI, the larger the global influence.

RESULTS

Positive Regulations Are as Common as Negative Ones in Cancers

As mentioned before, it suggests that positive regulations also play important roles in cancers. We first study the distribution of the Pearson correlation coefficient ρ_{ij} between the change of a miRNA and that of its target genes. **Figure 2** shows the distribution of the number of $|\rho_{ij}| > 0.3$ in breast cancer, liver cancer, and prostate cancer, respectively. First, the distribution of the number of $\rho_{ij} > 0.3$ and that of $\rho_{ij} < -0.3$ are very similar. Our observation further supports the results in Tan et al. (2019) that miRNAs exert both positive and negative regulations on their target genes. Second, most of the absolute ρ_{ij} are smaller than or equal to 0.5, which indicates that most of the regulatory strengths are relatively weak because one miRNA may target even hundreds of genes; On the other hand, only a few absolute $|\rho_{ij}|$ are larger than 0.8, which indicates that several individual genes may be strongly regulated by very few miRNAs.

As an example, **Figure 3** shows the scatter plots between the change of five miRNAs and that of their target gene KLF4, which regulates many critical physiologic and cellular processes (Wang et al., 2018). The X-axis of **Figures 3A–E** represents the expression change of a specific miRNA binding on target gene KLF4, X-axis of **Figure 3F** represents the total change of five miRNAs regulating KLF4, and the Y-axis of **Figure 3** represents the expression change of KLF4. We can see that three miRNAs (hsa-miR-10b-5p, hsa-145-5p, and hsa-miR-335-5p) positively correlate with KLF4, while the other two miRNAs (hsa-32-5p and hsa-miR-7-5p) negatively correlate with it. This demonstrates that miRNAs targeting one gene may affect it differently. Their total expression changes positively correlate with that of KLF4 as shown in the last subplot of **Figure 3**. It indicates that some impacts of the negatively correlated miRNAs can be offset by those of dominant ones.

miRNAs Regulate Their Target Genes in a Complex Way

In this section, we select a portion of miRNAs and their target genes and display their local influence relation in a bipartite graph in **Figure 4**. We can see that one miRNA may promote the expression of some genes while repressing that of the others. On the other hand, one gene may be upregulated by some miRNAs while being downregulated by other miRNAs. Furthermore, one miRNA may have a larger influence (wide lines) on some genes while having relatively smaller influence (thin lines) on the others. The observations demonstrate that miRNAs interact with their target genes in a complex way. The inference of these complex interactions forms the basis for us to understand the detailed roles of each miRNA on a specific gene. For example, PIK3CA is regulated by hsa-miR-10b-5p (Influence strength, 0.9509), hsa-miR-335-5p (Influence strength, 2.83E-4), hsa-miR-17-5p (Influence strength,-7.75E-4), hsa-miR-19a-3p (Influence strength,-7.47E-5), and hsa-miR-155-5p (Influence strength,-7.29E-4). Then its expression is thus mainly upregulated by hsa-miR-10b-5p.

Only a Few miRNAs Have Significant Global Influences on Cancers

From the perspective of systems biology, we are more interested in the most critical miRNAs, i.e., dominant miRNAs

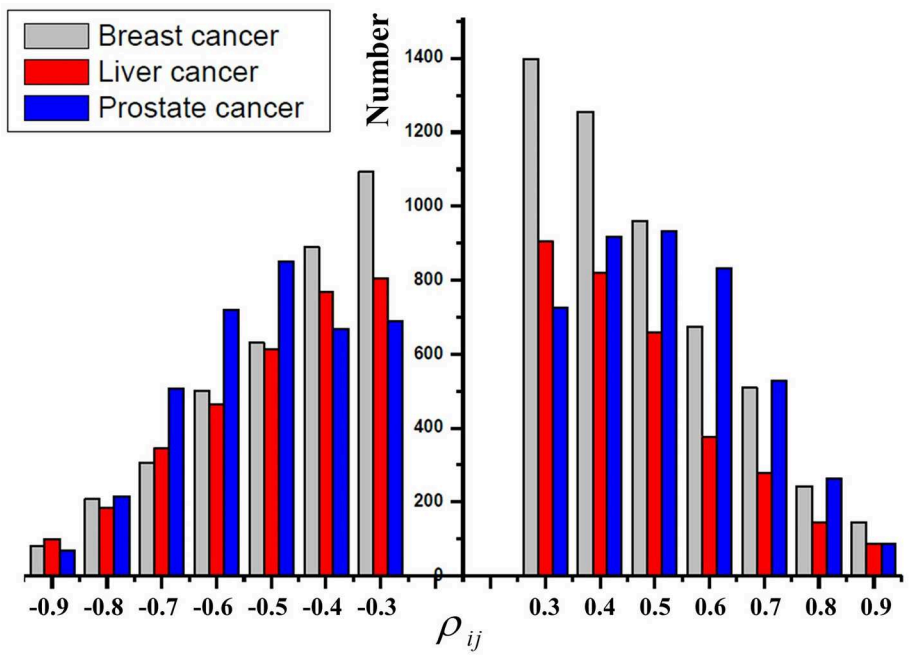


FIGURE 2 | The distribution of the Pearson correlation coefficients ρ_{ij} in breast cancer, liver cancer, and prostate cancer, respectively.

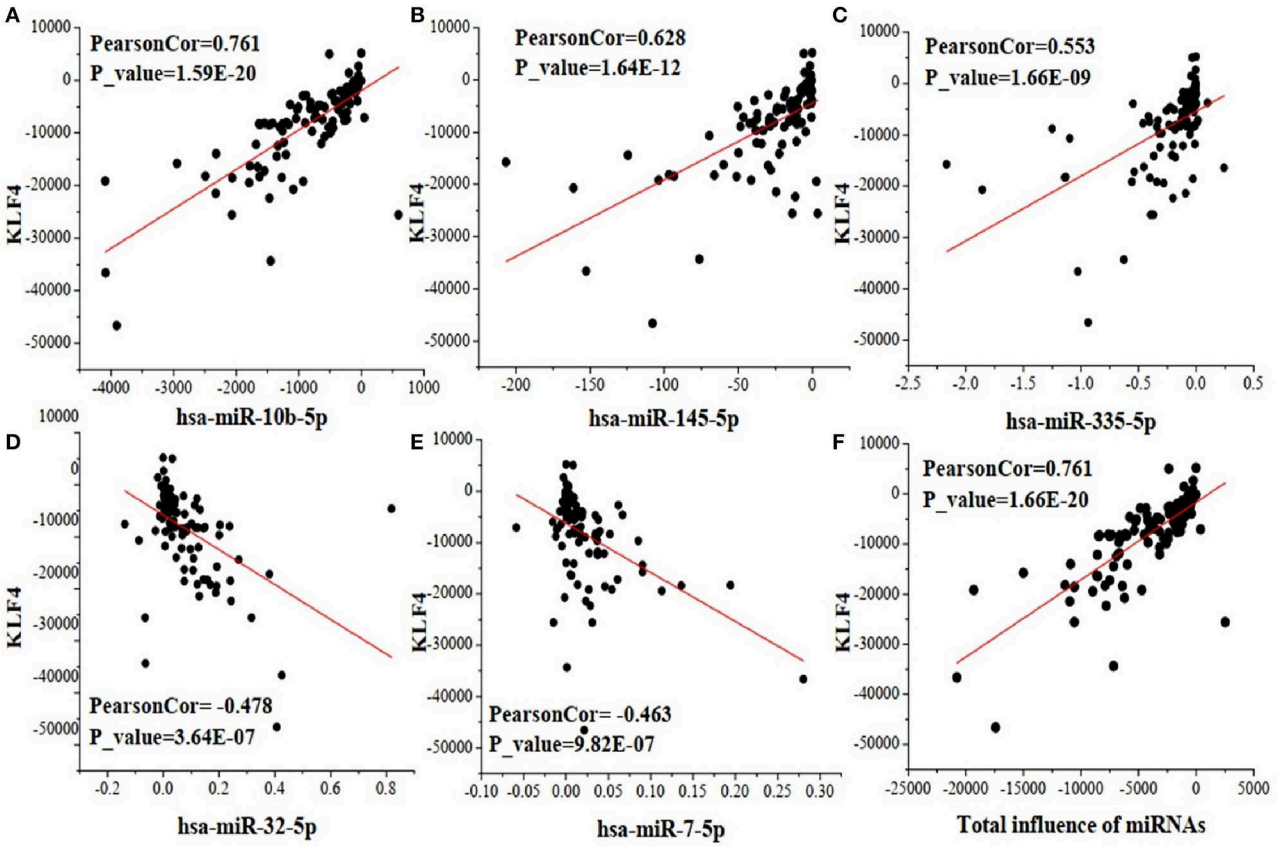


FIGURE 3 | The scatter plots between the change of five miRNAs and that of their target gene KLF4. **(A)** the correlation between hsa-miR-10b-5p and KLF4; **(B)** the correlation between hsa-miR-145-5p and KLF4; **(C)** the correlation between hsa-miR-335-5p and KLF4; **(D)** the correlation between hsa-miR-32-5p and KLF4; **(E)** the correlation between hsa-miR-7-5p and KLF4; **(F)** the correlation of total influence of the five miRNAs and their target gene KLF4.

that have the greatest influence on the whole regulatory network. Identifying the dominant miRNAs will answer the key question: which miRNAs are regulators of the most cancer-related genes?

Figure 5 illustrates the global influence of the top-ranked 20 miRNAs in breast, liver and prostate cancers, respectively. We can see that some miRNAs appear in all the three cancers whereas others may only show in one or two cancers. It suggests that some miRNAs play a common important role in many cancers while others are more related to specific cancers. Moreover, there are only a few miRNAs whose global influences are extremely larger than those of the others. This indicates their dysfunction may have very crucial impacts on the development of cancers. For instance, miR-21 ranked as the top one in the three cancers, has been confirmed highly involved in cancer proliferation and metastasis (Liu H. et al., 2018; Wang et al., 2019). On the other hand, we find that the most influenced genes by the dysfunctional miRNAs are highly related with cancers, such as CDK2, TP53, HRAS, NFKB1 (Carroll et al., 2000; Normanno et al., 2009; Xu et al., 2016).

FUNCTIONAL ANALYSIS OF THE CRITICAL miRNAs

Intervention of a Few Critical miRNAs May Help to Alleviate the Abnormal Expression of Most Cancer-Related Genes

Because one miRNA may regulate multiple downstream genes and the intervention on it may have different effects on the expression of its target genes. From the perspective of miRNA-based therapy, it is very crucial to figure out how the intervention of one miRNA may affect the abnormal expression of their target genes. If a miRNA promotes the expression of a gene and they are both overexpressed or under-expressed, then the intervention of the miRNA will exert a positive effect on the target gene to alleviate its abnormal expression; If a miRNA represses the expression of a gene and they have an opposite abnormal expression situation, then the intervention of the miRNA will also exert a positive effect on the target gene to alleviate the abnormal expression. Conversely, if a miRNA promotes the expression of a gene and they have the opposite abnormal expression situation, then the intervention of the

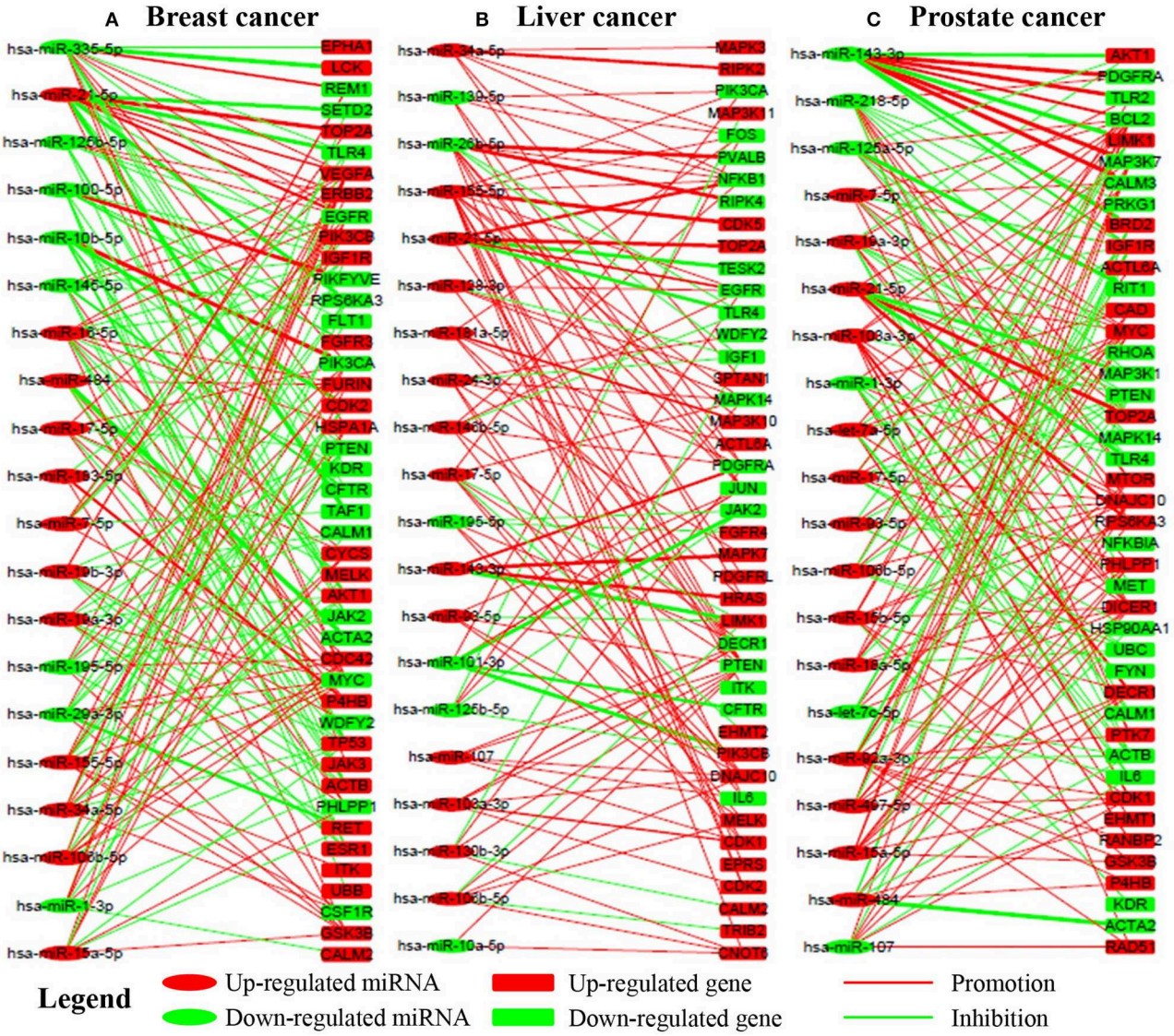


FIGURE 4 | The regualtory networks between miRNAs and genes in breast, liver, and prostate cancer, respectively. Red and green lines indicate upregualtion and downregualtion, respectively. Line width indicates the absoulte value of the local influence strength. Red and green nodes indicate overexpression and underexpression, respectively. **(A–C)** are the subnetworks for breast, liver and prostate cancer respectively.

miRNA will exert a negative effect on the target gene to deteriorate its abnormal expression. On the other hand, if a miRNA represses the expression of a gene and they are both overexpressed/under-expressed, then the intervention of the miRNA will also exert a negative effect on the target gene to deteriorate its abnormal expression.

Figure 6 shows the subnetwork of hsa-miR-21-5p in the three cancers, when intervention on hsa-miR-21-5p, the left-hand genes are those being positively affected while the right-hand ones negatively affected genes. This reveals that the intervention of one miRNA may have complex effects on cancer-related genes. Specifically, the abnormal expression of some genes can be alleviated while the other may be further deteriorated.

Based on the regulation relations and the abnormal expression situations, we summarize the number of positively and the negatively affected miRNA-gene pairs of the top five miRNAs in **Table 1**. The "+"/"-" symbols in **Table 1** represent the number of positively/negatively affected miRNA-gene pairs after intervention. We have the following observations from the table. First, we can see that the interventions of the five miRNAs may affect about 1,000 of abnormal miRNA-gene pairs which indicate they regulate many downstream genes. Second, the number of positively affected genes is extremely larger than that of the negatively affected ones. Third, the absolute relative influence strengths of most of the positively affected pairs are larger than 0.5. The observations indicate that the intervention of the top five

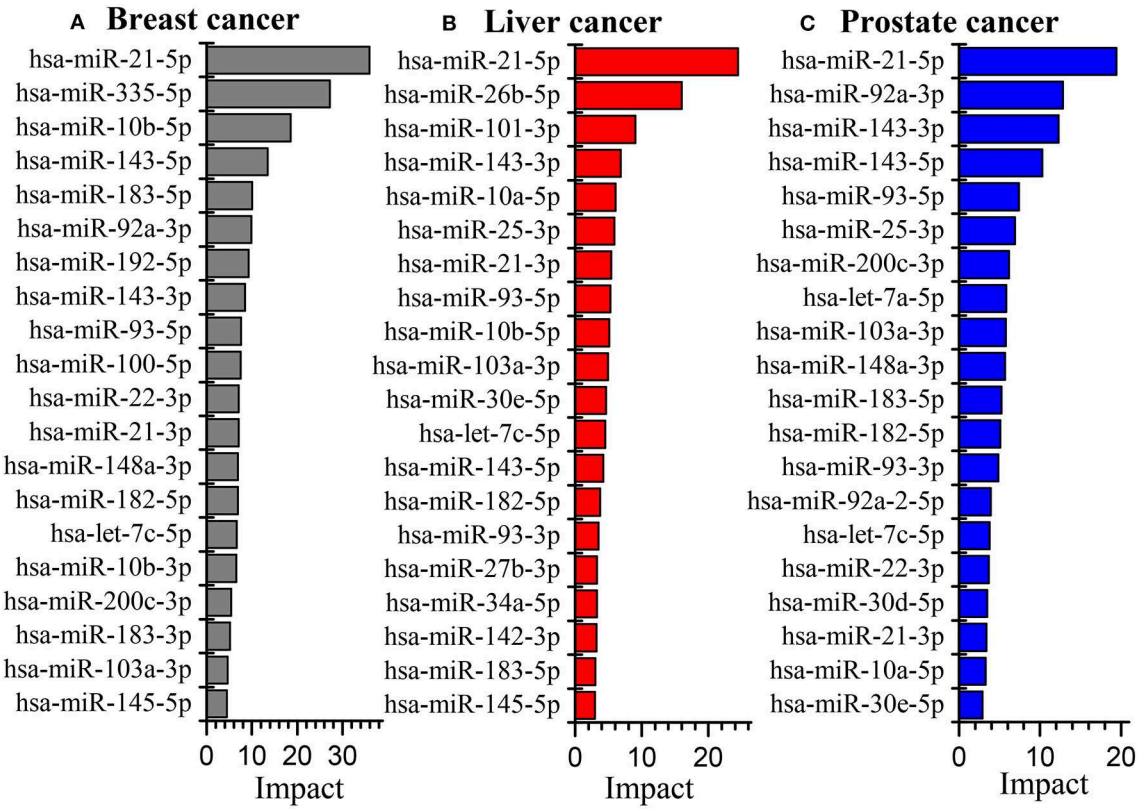


FIGURE 5 | The global influence of the top 20 miRNAs in breast cancer **(A)**, liver cancer **(B)** and prostate cancer **(C)** datasets, respectively.

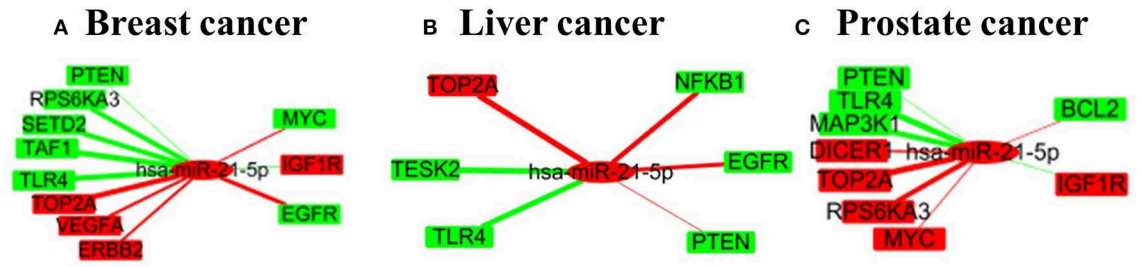


FIGURE 6 | A small subnetwork of miRNA-21-5p and its target genes in three cancers. **(A–C)** are for the subnetworks in breast cancer, liver cancer and prostate cancer respectively.

miRNAs may significantly drive the cancer-related genes to the normal levels. Therefore, they are potential invention biomarkers for miRNA-based therapy.

Most of the Critical miRNAs Involve in Some Important Biological Functions

We conduct functional analysis for top-ranked miRNAs by integrating the co-expression similarity, co-GO similarity, co-literature similarity, and co-similar disease similarity by using miRNA functional analysis tool MISIM (Li et al., 2019). There are 14, 15, and 13 miRNAs respectively annotated by MISIM in the top 20 important miRNAs of the three cancer datasets.

Figure 7A shows the function similarity network of the top 20 miRNAs, where red color lines denote the correlation coefficients larger than 0.5. It suggests most of the miRNAs are highly correlated in their biological functions. **Figure 7B** shows the top 10 enriched biological functions (FDR<7.0E-02) of the miRNAs. These biological functions, such as inflammation, cell proliferation, apoptosis, and cell cycle, have been verified closely related to different cancers (Evan and Vousden, 2001; Taniguchi and Karin, 2018; Xu et al., 2020). This indicates the critical miRNAs might interact in a highly coherent way to drive the biological system from normal to disease state in the three cancers.

As the overlapped critical miRNA in three cancers, hsa-mir-21 was reported to involved in various cancers such as colorectal cancer, breast cancer and lung cancer (Xu et al., 2016; Liu W. et al., 2018). It has been reported that overexpression of miR-21 could promote the cellular proliferation, colony formation, invasion and also inhibit cell death in a wide variety of cancerous cells by regulation of various targets including PTEN, TPM1, and PDCD4 (Najjary et al., 2020).

TABLE 1 | The number of positive and negative effect interactions between the top 5 miRNAs and their targets in three cancers, respectively.

$Abs(I_{ij})$	Breast cancer		Liver cancer		Prostate cancer	
	+	−	+	−	+	−
(0.9–1)	**583**	**42**	**175**	**35**	**344**	**2**
(0.8–0.9)	**190**	**9**	**92**	**16**	**119**	**1**
(0.7–0.8)	**94**	**4**	**76**	**25**	**86**	**1**
(0.6–0.7)	**90**	**13**	**63**	**18**	**58**	**0**
(0.5–0.6)	**84**	**19**	**48**	**15**	**64**	**3**
(0.4–0.5)	77	15	66	28	67	2
(0.3–0.4)	44	19	54	37	71	5
(0.2–0.3)	43	30	52	50	82	4
(0.1–0.2)	52	50	61	59	158	17
Total	1257	201	687	283	1049	35

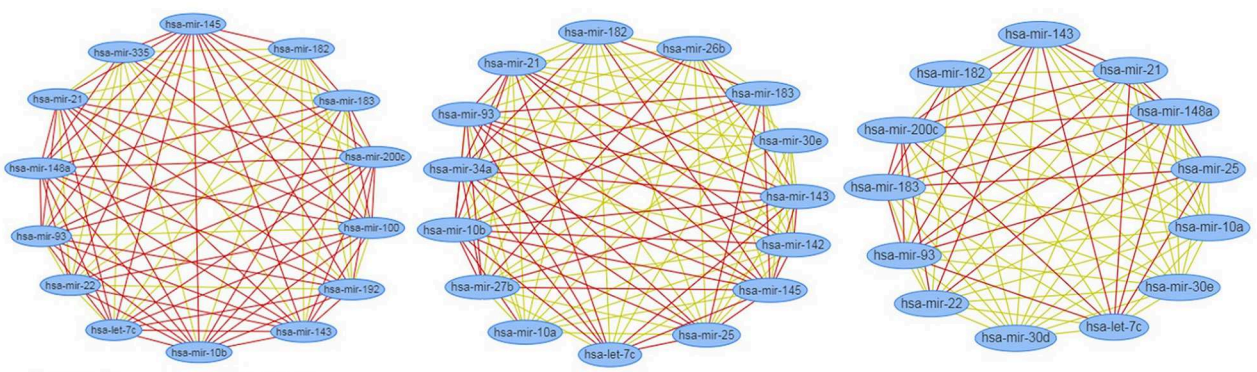
The bold values are the number of interactions with influence strength more than 0.5.

Some Critical miRNAs Also Impact the Resistance/Sensitivity of Drugs

Some non-coding RNAs (ncRNAs) especially miRNAs could promote sensitivity or produce resistance of drugs by regulating their target genes. To evaluate the impacts of the critical miRNAs on the resistance/sensitivity of drugs, we submit the top 20 miRNAs to two state-of-the-art miRNA-drug interaction databases: *ncDR* (Dai et al., 2017) and *mTD* (Chen et al., 2017). The two databases include 1,056, 384, and 127 records for miRNA-drug interactions in breast, liver and prostate cancers respectively as well as curating a lot of resistance/sensitivity related ncRNAs.

We find that there are 10, 6, and 4 miRNAs impacting drug resistance/sensitivity cases in breast cancer, liver and prostate cancer, respectively, among the top 20 miRNAs. The main reason for the small number of miRNAs in the liver and prostate cancers lies in that there are relatively fewer records about them in the two databases. **Figure 8** shows their abnormal expression and corresponding influence on

A



B

Breast cancer			Liver cancer			Prostate cancer		
Term	Count	FDR	Term	Count	FDR	Term	Count	FDR
Inflammation	11	1.5E-07	Inflammation	10	1.8E-06	Aging	8	1.5E-05
Immune Response	8	6.9E-05	Regulation of Stem Cell	10	1.6E-07	Regulation of Stem Cell	8	2.3E-05
Regulation of Stem Cell	8	2.6E-05	Cell Death	9	1.6E-06	Inflammation	8	1.6E-04
Cell Death	7	2.3E-04	Apoptosis	8	1.6E-04	Immune Response	7	4.9E-04
Cell Proliferation	7	2.6E-04	Angiogenesis	7	8.7E-05	Cell Death	7	1.9E-04
EMT	7	3.1E-04	Cell Proliferation	7	2.6E-04	Apoptosis	7	9.7E-04
Aging	7	7.2E-05	Cell Cycle	6	2.0E-03	Cell Proliferation	6	1.9E-03
Angiogenesis	6	7.5E-04	Embryonic Development	6	9.8E-07	EMT	6	2.2E-03
Apoptosis	6	7.1E-03	EMT	6	2.0E-03	Cell Differentiation	5	3.4E-03
Cell Differentiation	6	3.8E-04	Cell Differentiation	6	3.9E-04	Cell Cycle	4	7.0E-02

Note: EMT, Epithelial-to-Mesenchymal Transition.

FIGURE 7 | Biological function analysis of the top 20 miRNAs. **(A)** miRNAs function enrichment analysis; **(B)** functional similarity network analysis of the miRNAs.

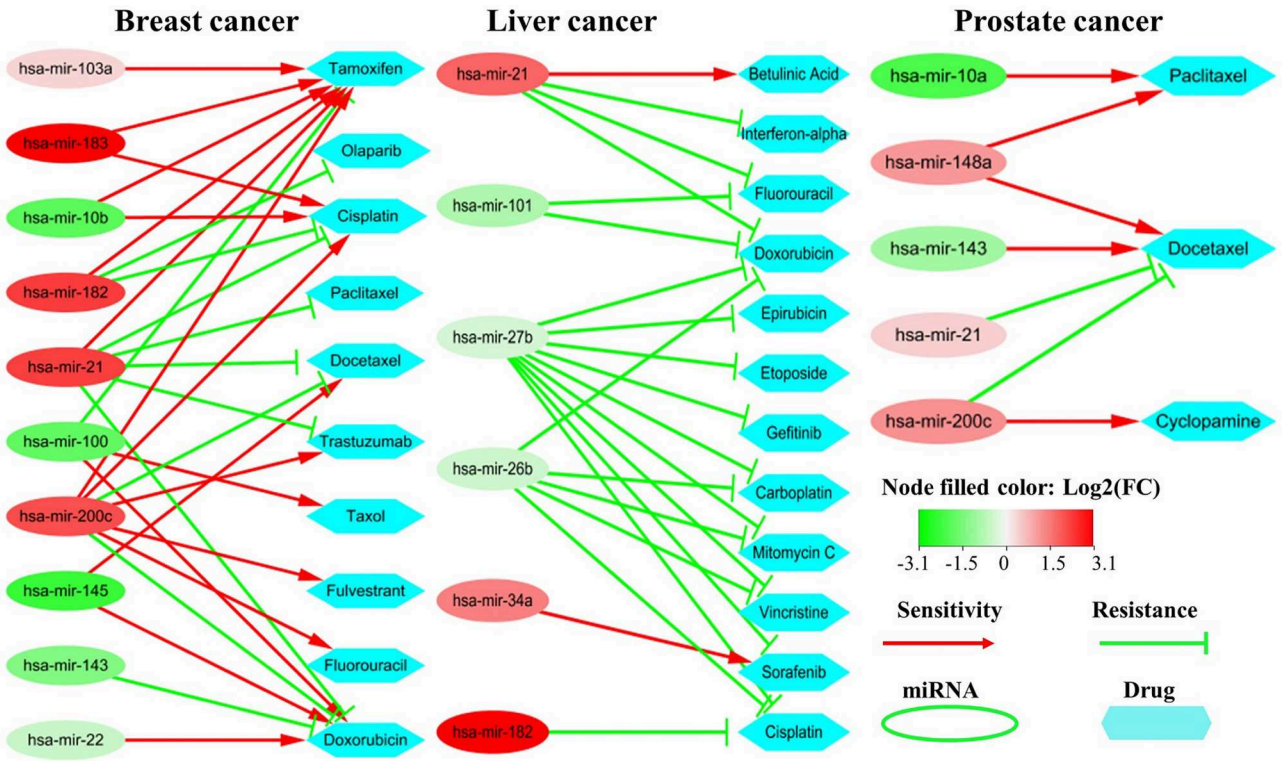


FIGURE 8 | Drug sensitivity/resistance analysis of the top 20 miRNAs. The red (or green) oval represents the miRNA is up-regulated (or down-regulated) in cancer samples. The red line denotes promoting drug sensitivity while the green line denotes inducing drug resistance.

drug sensitivity/resistance. On one hand, one miRNA may influence multiple drugs with different effects. For example, the over-expression of hsa-mir-182 in breast cancer could induce drug resistance to both Olaparib and Cisplatin while promoting the sensitivity of Tamoxifen. On the other hand, some miRNAs may promote a drug sensitivity while others induce its resistance. The complex effects of these miRNAs on cancer-related drugs not only further demonstrate their importance in cancer development, but also provide a new insight for accurate drug selection.

The Expression of Critical miRNAs Is Highly Related to the Survival Time of Prognosis

We also find that critical miRNAs can influence the survival time of prognosis seriously. **Figure 9** shows the Kaplan–Meier curves of the top three miRNAs in breast, liver and prostate cancers, respectively. Most of them are significantly correlated to the overall survival time in both breast and liver cancers except for prostate cancer. One major reason is that most prostate tumors are slow-growing and many of them are not lethal. Furthermore, some important correlations between the miRNAs are supported by wet-lab experiments. For example, Yan et al. demonstrated that overexpression of miR-21 was associated with human breast cancer poor prognosis (Yan et al., 2008). Ji et al. showed that liver cancer patients with low miR-26 expression had shorter overall survival time (Ji et al., 2009, 2013). These observations indicate that the identified critical miRNAs may also serve as potential biomarkers for the survival time of prognosis.

CONCLUSION

MiRNAs have been reported as a kind of important non-coding regulators influencing the expression of more than 60% genes. In this paper, we proposed a novel miRNA influence capturing (miRNAInf) method to characterize the regulatory mechanism of miRNA on their target genes as well as identify critical miRNAs that have dominantly important impacts on target genes. Out results from the breast, prostate and liver cancer datasets further verify that miRNAs may either upregulate or downregulate their target genes instead of mainly repressing them. We identified some critical miRNAs involved in the three cancers by constructing a miRNA-gene regulatory network. Our biological functional analysis shows that those critical miRNAs are not only highly correlated with each other but also involved in many important biological functions such as apoptosis, proliferation, etc. Furthermore, miRNA-gene interaction analysis reveals that the intervention of only a few top crucial miRNAs may potentially alleviate the abnormal expressions of many genes and push the cancer system to a normal situation. It suggests that the identified crucial miRNAs may serve as potential biomarkers for miRNA-based therapy as well as diagnosis. In addition, we find some critical miRNAs may influence the sensitivity/resistance of drugs as well as the survival time of

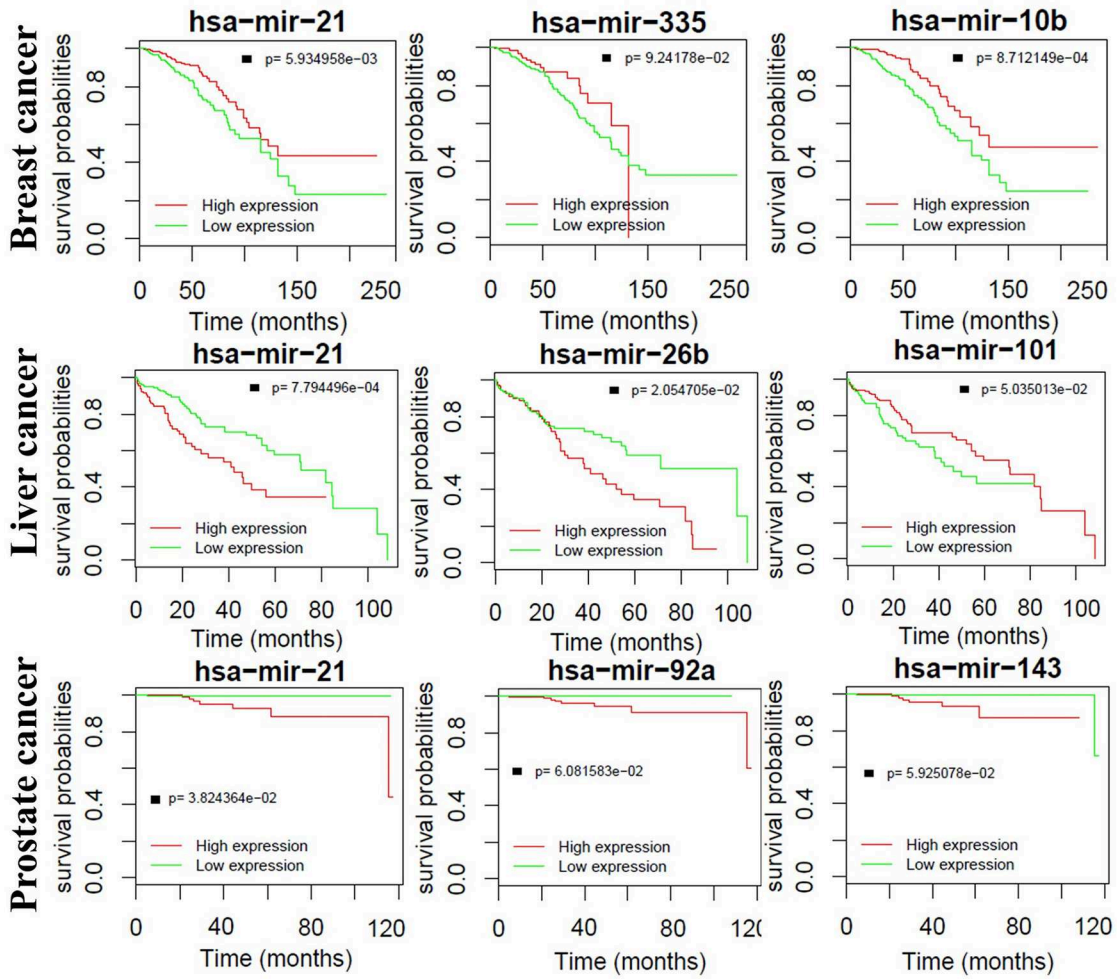


FIGURE 9 | Kaplan–Meier curves of the top three miRNAs in breast, liver and prostate cancers, respectively.

prognosis. Our study provides a strong foundation to support the combination of miRNA-based therapy and cancer drugs to improve the treatment effect in precision medicine. To the best of our knowledge, this study first provides a systematic approach to decipher the roles of miRNAs in the diagnosis and prognosis of complex diseases and will inspire future studies in this field.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**. The source codes for this study can be found at https://github.com/xupeng2017/MiRNAInf.

AUTHOR CONTRIBUTIONS

WL and PX designed the methods. PX wrote the codes. WL, PX, QW, JY, YR, GF, ZK, XS, and HH analyzed the data. WL, HH, and PX wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was funded in part by the National Key R & D Program of China (Grant No. 2019YFA0706402), National Natural Science Foundation of China (grant nos. 61572367, 61972107, and 61972109), Specialized Fund for Science and Technology Platform and Talent Team Project of Guizhou Province (No. QianKeHePingTaiRenCai [2016]5609), Key Supported Disciplines of Guizhou Province Computer Application Technology (No. QianXueWeiHeZi ZDXK [2016]20).

ACKNOWLEDGMENTS

We would like to thank the reviewers for their valuable comments and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00278/full#supplementary-material

REFERENCES

Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005. doi: 10.7554/eLife.05005.028

Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M. Q. (2010). Development of the human cancer microRNA network. *Silence* 1:6. doi: 10.1186/1758-907X-1-6

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297. doi: 10.1016/S0092-8674(04)00045-5

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215–233. doi: 10.1016/j.cell.2009.01.002

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:R90. doi: 10.1186/gb-2010-11-8-r90

Carroll, J. S., Prall, O. W., Musgrove, E. A., and Sutherland, R. L. (2000). A pure estrogen antagonist inhibits cyclin E-Cdk2 activity in MCF-7 breast cancer cells and induces accumulation of p130-E2F4 complexes characteristic of quiescence. *J. Biol. Chem.* 275, 38221–38229. doi: 10.1074/jbc.M004424200

Chen, X., Xie, W.-B., Xiao, P.-P., Zhao, X.-M., and Yan, H. (2017). mTD: a database of microRNAs affecting therapeutic effects of drugs. *J. Genet. Genomics.* 44, 269–271. doi: 10.1016/j.jgg.2017.04.003

Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S. D., Lin, Y. L., Lee, W. H., et al. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* 44, D239–D247. doi: 10.1093/nar/gkv1258

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

Cui, C., Shi, B., Shi, J., Zhou, Y., and Cui, Q. (2019). Defining the importance score of human microRNAs and mutants using their single nucleotide random forest regression and sequence data. *Adv. Theory Simul.* 2:1900083. doi: 10.1002/adts.201900083

Dai, E., Yang, F., Wang, J., Zhou, X., Song, Q., An, W., et al. (2017). ncDR: a comprehensive resource of non-coding RNAs involved in drug resistance. *Bioinformatics* 33, 4010–4011. doi: 10.1093/bioinformatics/btx523

Dela Cruz, F., and Matushansky, I. (2011). MicroRNAs in chromosomal translocation-associated solid tumors: learning from sarcomas. *Discov. Med.* 12, 307–317.

Evan, G. I., and Vousden, K. H. (2001). Proliferation, cell cycle, and apoptosis in cancer. *Nature* 411, 342–348. doi: 10.1038/35077213

Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* 79, 351–379. doi: 10.1146/annurev-biochem-060308-103103

Faraldi, M., Gomarasca, M., Sansoni, V., Perego, S., Banfi, G., and Lombardi, G. (2019). Normalization strategies differently affect circulating miRNA profile associated with the training status. *Sci. Rep.* 9:1584. doi: 10.1038/s41598-019-38505-x

Gotte, M. (2010). MicroRNAs in breast cancer pathogenesis. *Minerva Ginecol.* 62, 559–571.

Hata, A., and Lieberman, J. (2015). Dysregulation of microRNA biogenesis and gene silencing in cancer. *Sci. Signal.* 8:re3. doi: 10.1126/scisignal.2005825

Ji, J., Shi, J., Budhu, A., Yu, Z., Forgues, M., Roessler, S., et al. (2009). MicroRNA expression, survival, and response to interferon in liver cancer. *N. Engl. J. Med.* 361, 1437–1447. doi: 10.1056/NEJMoa0901282

Ji, J., Yu, L., Yu, Z., Forgues, M., Uenishi, T., Kubo, S., et al. (2013). Development of a miR-26 companion diagnostic test for adjuvant interferon-alpha therapy in hepatocellular carcinoma. *Int. J. Biol. Sci.* 9, 303–312. doi: 10.7150/ijbs.6214

Li, J., Zhang, S., Wan, Y., Zhao, Y., Shi, J., Zhou, Y., et al. (2019). MISIM v2.0: a web server for inferring microRNA functional similarity based on microRNA-disease associations. *Nucleic Acids Res.* 47, W536–W541. doi: 10.1093/nar/gkz328

Liu, H., Cheng, L., Cao, D., and Zhang, H. (2018). Suppression of miR-21 expression inhibits cell proliferation and migration of liver cancer cells by targeting Phosphatase and Tensin Homolog (PTEN). *Med. Sci. Monit.* 24, 3571–3577. doi: 10.12659/MSM.907038

Liu, W., Cui, Z., and Zan, X. (2018). Identifying cancer-related microRNAs based on subpathways. *IET Syst. Biol.* 12, 273–278. doi: 10.1049/iet-syb.2018.5025

Lovat, F., Valeri, N., and Croce, C. M. (2011). MicroRNAs in the pathogenesis of cancer. *Semin. Oncol.* 38, 724–733. doi: 10.1053/j.seminoncol.2011.08.006

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

Najjary, S., Mohammadzadeh, R., Mokhtarzadeh, A., Mohammadi, A., Kojabad, A. B., and Baradaran, B. (2020). Role of miR-21 as an authentic oncogene in mediating drug resistance in breast cancer. *Gene* 738:144453. doi: 10.1016/j.gene.2020.144453.

Normanno, N., Tejpar, S., Morgillo, F., de Luca, A., van Cutsem, E., and Ciardiello, F. (2009). Implications for KRAS status and EGFR-targeted therapies in metastatic CRC. *Nat. Rev. Clin. Oncol.* 6, 519–527. doi: 10.1038/nrclinonc.2009.111

Rupaimoole, R., and Slack, F. J. (2017). MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.* 16, 203–222. doi: 10.1038/nrd.2016.246

Slack, F. J., and Chinnaiyan, A. M. (2019). The role of non-coding RNAs in oncology. *Cell* 179, 1033–1055. doi: 10.1016/j.cell.2019.10.017

Song, J., Hu, Y., Li, J., Zheng, H., Wang, J., Guo, L., et al. (2017). Different microRNA profiles reveal the diverse outcomes induced by EV71 and CA16 infection in human umbilical vein endothelial cells using high-throughput sequencing. *PLoS ONE* 12:e0177657. doi: 10.1371/journal.pone.0177657

Sun, B. Z., Zhang, J. Z., Liu, M. H., and Guan, L. Y. (2019). Alkannin inhibits proliferation, migration and invasion of hepatocellular carcinoma cells via regulation of miR-92a. *Biomed. Pharmacotherapy* 114:108782. doi: 10.1016/j.biopha.2019.108782

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). 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

Tan, H., Huang, S., Zhang, Z., Qian, X., Sun, P., and Zhou, X. (2019). Pan-cancer analysis on microRNA-associated gene activation. *EBioMedicine* 43, 82–97. doi: 10.1016/j.ebiom.2019.03.082

Taniguchi, K., and Karin, M. (2018). NF-κB, inflammation, immunity and cancer: coming of age. *Nat. Rev. Immunol.* 18, 309–324. doi: 10.1038/nri.2017.142

Vasudevan, S., and Steitz, J. A. (2007). AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell* 128, 1105–1118. doi: 10.1016/j.cell.2007.01.038

Vasudevan, S., Tong, Y., and Steitz, J. A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science* 318, 1931–1934. doi: 10.1126/science.1149460

Wang, H., Tan, Z., Hu, H., Liu, H., Wu, T., Zheng, C., et al. (2019). microRNA-21 promotes breast cancer proliferation and metastasis by targeting LZTFL1. *BMC Cancer* 19:738. doi: 10.1186/s12885-019-5951-3

Wang, L., Shen, F., Stroehlein, J. R., and Wei, D. (2018). Context-dependent functions of KLF4 in cancers: could alternative splicing isoforms be the key? *Cancer Lett.* 438, 10–16. doi: 10.1016/j.canlet.2018.09.005

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

Xiao, M., Li, J., Li, W., Wang, Y., Wu, F., Xi, Y., et al. (2017). MicroRNAs activate gene transcription epigenetically as an enhancer trigger. *RNA Biol.* 14, 1326–1334. doi: 10.1080/15476286.2015.1112487

Xu, P., Wang, J., Sun, B., and Xiao, Z. (2018). Integrated analysis of miRNA and mRNA expression data identifies multiple miRNAs regulatory networks for the tumorigenesis of colorectal cancer. *Gene* 659, 44–51. doi: 10.1016/j.gene.2018.03.050

Xu, P., Zhao, G., Kou, Z., Fang, G., and Liu, W. (2020). Classification of cancers based on a comprehensive pathway activity inferred by genes and

their interactions. *IEEE Access* 8, 30515–30521. doi: 10.1109/access.2020.2973220

Xu, P., Zhu, Y., Sun, B., and Xiao, Z. (2016). Colorectal cancer characterization and therapeutic target prediction based on microRNA expression profile. *Sci. Rep.* 6:20616. doi: 10.1038/srep20616

Yan, L. X., Huang, X. F., Shao, Q., Huang, M. Y., Deng, L., Wu, Q. L., et al. (2008). MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *RNA* 14, 2348–2360. doi: 10.1261/rna.1034808

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 © 2020 Xu, Wu, Yu, Rao, Kou, Fang, Shi, Liu 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fbioe.2020.00227

iRNA5hmC: The First Predictor to Identify RNA 5-Hydroxymethylcytosine Modifications Using Machine Learning

Yuan Liu[1], Dasheng Chen[1], Ran Su[1], Wei Chen[2,3]* and Leyi Wei[4,5]**

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China, [2] Center for Genomics and Computational Biology, School of Life Sciences, North China University of Science and Technology, Tangshan, China, [3] Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, China, [4] School of Software, Shandong University, Jinan, China, [5] Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan, China

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Lei Chen,
Shanghai Maritime University, China
Xiucai Ye,
University of Tsukuba, Japan

***Correspondence:**
Ran Su
ran.su@tju.edu.cn
Wei Chen
chenweiimu@gmail.com
Leyi Wei
weileyi@sdu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *15 January 2020*
Accepted: *05 March 2020*
Published: *31 March 2020*

Citation:
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

RNA 5-hydroxymethylcytosine (5hmC) modification plays an important role in a series of biological processes. Characterization of its distributions in transcriptome is fundamentally important to reveal the biological functions of 5hmC. Sequencing-based technologies allow the high-throughput identification of 5hmC; however, they are labor-intensive, time-consuming, as well as expensive. Thus, there is an urgent need to develop more effective and efficient computational methods, at least complementary to the high-throughput technologies. In this study, we developed iRNA5hmC, a computational predictive protocol to identify RNA 5hmC sites using machine learning. In this predictor, we introduced a sequence-based feature algorithm consisting of two feature representations, (1) *k*-mer spectrum and (2) positional nucleotide binary vector, to capture the sequential characteristics of 5hmC sites. Afterward, we utilized a two-stage feature space optimization strategy to improve the feature representation ability, and trained a predictive model using support vector machine (SVM). Our feature analysis results showed that feature optimization can help to capture the most discriminative features. As compared to well-known existing feature descriptors, our proposed representations can more accurately separate true 5hmC from non-5hmC sites. To the best of our knowledge, iRNA5hmC is the first RNA 5hmC predictor that enables to make predictions based on RNA primary sequences only, without any need of prior experimental knowledge. Importantly, we have established an easy-to-use webserver which is currently available at http://server.malab.cn/iRNA5hmC. We expect it has potential to be a useful tool for the prediction of 5hmC sites.

Keywords: RNA 5-hydroxymethylcytosine modification, feature representation, machine learning, web server, sequence analysis

KEY POINTS

- iRNA5hmC is the first RNA 5-hydroxymethylcytosine site predictor, which enables to make predictions based on RNA primary sequences without prior experimental knowledge.
- Benchmarking comparison results show that iRNA5hmC outperforms other machine learning algorithms trained with existing sequence-derived feature descriptors.
- Our feature analysis demonstrates that there exists the compositional and positional specificity between true 5hmC sites and non-5hmC sites.
- We have established an easy-to-use webserver that implements the predictor. It is publicly accessible at http://server.malab.cn/iRNA5hmC.

INTRODUCTION

RNA can be decorated by various chemical modifications (Boccaletto et al., 2018). Over the past decades, more than 100 kinds of modifications have been identified in mRNA, tRNA, rRNA, and snRNA, etc. (Shi et al., 2019). These modifications play important roles in a series of biological processes (Roundtree et al., 2017), such as RNA splicing, RNA translation, and RNA decay. In addition, it was also demonstrated that RNA modifications are associated with human diseases (Jonkhout et al., 2017), including cancer, cardiovascular diseases, Bowen–Conradi syndrome, obesity, and diabetes, etc. Hence, determining their distributions in the transcriptomes is important for decoding the biological and physiological functions of RNA modifications.

Thanks to the high-throughput sequencing methods, recent years have witnessed a burst of researches on N^6-methyladenine (m^6A), N^1-methyladenine (m^1A), N^7-methylguanosine (m^7G), and 5-methylcytidine (m^5C), etc. (Conde et al., 2015; Chen et al., 2019; Pian et al., 2019; Yuan et al., 2019). Another kind of RNA modification, called 5-hydroxymethylcytosine (5hmC) is formed by TET-mediated oxidation of m5C (Fu et al., 2014). The 5hmC was originally identified in wheat seedlings (Racz et al., 1978), and was also detected in various tissues of mouse and human (Li and Liu, 2011). Later on, Huber et al. (2015) found that 5hmC is pervasive in all three domains of life across a variety of different species.

Recently, by using the hMeRIP-seq method, Delatte et al. (2016) revealed a transcriptome wide profile of 5hmC in *Drosophila* and found that 5hmC modifications are non-randomly distributed, with an enrichment in coding regions. Meanwhile, they also found that 5hmC modifications are abundant in the *Drosophila* brain. A similar result was also observed by Miao et al. (2016); they found a high level of 5hmC modification enrichment in mouse brain stem, hippocampus, and cerebellum regions. These results suggest that 5hmC modification might play an important role in brain tissue. To further revealing the biological functions of 5hmC, it is necessary to characterize its distribution in the transcriptome of multiple spices. Unfortunately, the distribution of 5hmC remains uncharacterized in most species.

Considering that the high-throughput experimental methods are expensive and time-consuming, it is necessary to develop computational methods for the detection of 5hmC modification sites. Inspired by the successful application of machine learning methods for identifying RNA modifications, in this study, we developed iRNA5hmC, a computational predictor to predict RNA 5hmC sites using machine learning. In this predictor, we used the *k*-mer spectrum and positional nucleotide binary vector to respectively capture the sequence composition and position-specific characteristics of 5hmC sites, utilized a two-stage feature selection strategy to optimize the feature space, and trained the SVM-based predictive model. To the best of our knowledge, iRNA5hmC is the very first machine learning predictor that enables researchers to make RNA 5hmC predictions based on RNA primary sequences only, without any other prior experimental knowledge. Importantly, we have established an easy-to-use webserver to make the proposed predictor more impactful. We expect that it has the potential to be a complementary tool to the high-throughput sequencing methods.

MATERIALS AND METHODS

Datasets

Here, we constructed the first 5hmC dataset for training the predictive model. It consists of positive samples and negative samples. The positive samples were collected based on Delatte et al.'s (2016) work, which contains 662 5hmC site containing sequences with the sequence similarity less than 80%. According to our previous experiences (Chen et al., 2019), the sequences were given the length of 41 nt (nucleotides) with the 5hmC site in the center. The negative samples (non-5hmC site containing sequences) were obtained by choosing 41-nt long sequences with the intermediate cytosines that are not detected as 5hmC by the hMeRIP-seq method. Accordingly, a huge number of negative samples were collected. In order to balance the number of samples between positive and negative dataset in model training, we randomly selected out 662 non-5hmC site containing sequences as the negative samples. The dataset used to train the proposed model is available at http://server.malab.cn/iRNA5hmC.

The Proposed Predictive Framework

The predictive procedure can be concluded as two phases: (1) model training and (2) prediction. In the training phase, the training samples are encoded and integrated by feature representation algorithms. Afterward, the features are optimized to obtain the best feature subset, which are then fed into the SVM algorithm to train predictive model. In prediction phase, given the query sequences that are not characterized, we followed the similar procedure to encode the sequences, and used the trained model to predict whether or not the query sequences are 5hmC sequences. The SVM model gives each query sequence a score to measure how likely it is true 5hmC sequence. If the score is higher than 0.5, it is considered to be the 5hmC sequence; otherwise, it is not.

Feature Representation

In this study, we introduce a feature representation algorithm containing the following two sequence-based feature descriptors: (1) *k*-mer spectrum and (2) nucleotide binary encoding, which are described as follows.

The first feature descriptor is *k*-mer spectrum. There are two reasons for using it. One is that it is a simple and useful feature algorithm to encode character sequences like RNAs and DNAs. On the other hand, more importantly, previous study has demonstrated that DNA 5mC is often found in contexts of CG, CHG, and CHH (H represents either A, C, or T) (Kumar et al., 2018). Therefore, there might be similar for RNA 5hmC modification.

For convenience of discussions, a given RNA sequence can be represented as

$$S = R_1 R_2 \cdots R_i \cdots R_{L-1} R_L \tag{1}$$

where R_1 represents the first nucleotide, R_2 represents the second nucleotide, and so forth. R_i can be any of the four nucleotides {A, C, U, G}. The *k*-mer spectrum computes the occurrence frequencies of all possible sequential patterns with length *k*. Therefore, using this descriptor, the given sequence can be represented as,

$$F^{k\text{-mer}} = \left[f_1^{k\text{-mer}}, f_2^{k\text{-mer}}, \ldots, f_i^{k\text{-mer}}, \ldots, f_{4^k}^{k\text{-mer}}\right] \tag{2}$$

where $f_i^{k\text{-mer}}$ is the occurrence frequency of the *i*-th *k*-mer in *S*. Similarly, we used 2-mer and 3-mer spectrum to encode our RNA sequences. Naturally, *S* is represented as 2-mer and 3-mer vector, respectively:

$$F^{2-\text{mer}} = \left[f\left(AA\right), f\left(AC\right), \ldots, f\left(GG\right)\right] \tag{3}$$

$$F^{3-\text{mer}} = \left[f\left(AAA\right), f\left(AAC\right), \ldots, f\left(GGG\right)\right] \tag{4}$$

The second feature descriptor is nucleotide binary encoding, in which we transform different nucleotides into different numeric vectors by the following rule: the codes of "A," "U," "C," and "G" are "0001," "0010," "0100," and "1000," respectively.

Finally, a given RNA sequence is encoded as a total of 244 features ($41 \times 4 + 4^2 + 4^3 = 244$).

Feature Optimization

Feature optimization is a key step to remove the noisy features and retain the features having the highest degree of separability between two classes, which has been employed to improve the predictive performance in several bioinformatics problems. In this study, we used a two-stage feature selection strategy. In the first step, we compute the feature importance for the 244 features by analysis of variance (ANOVA) (Chen et al., 2016), which calculates the separability degree of each feature to obtain respective *F*-value and yields a feature ranking list regarding their classification importance. The feature with a larger *F*-value indicates much more importance. The ANOVA *F*-value of the θ-th feature definitions is given below:

$$F - \text{value}(\theta) = \frac{S_{B(\theta)}^2}{S_{w(\theta)}^2} \tag{5}$$

where $S_{B(\theta)}^2$ and $S_{w(\theta)}^2$ are the means square between (MSB) and means square within (MSW), respectively. They are defined as follows:

$$S_{B(\theta)}^2 = \frac{1}{df_B}\sum_{i=1}^{K} n_i \left(\frac{\sum_{j=1}^{n_i} f_{ij}\left(\theta\right)}{n_i} - \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} f_{ij}\left(\theta\right)}{\sum_{i=1}^{K} n_i}\right)^2 \tag{6}$$

$$S_w^2\left(\theta\right) = \frac{1}{df_w}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(f_{ij}\left(\theta\right) - \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} f_{ij}\left(\theta\right)}{\sum_{i=1}^{K} n_i}\right)^2 \tag{7}$$

here $df_B = K - 1$ and $df_w = N - K$ are degrees of freedom for MSB and MSW, respectively. *K* and *N* represent the number of groups (for the current case *K* = 2) and total number of samples, respectively; and n_i is the number of sample in the *i*-th group. $f_{ij}(\theta)$ denotes feature value of the θ-th feature of the *j*-th sample in the *i*-th group.

In the second step, we used the sequential forward search (SFS) strategy to determine the optimal feature representations (Whitney, 2006). To be specific, features from the ranked feature list are added ten-by-ten from lower rank (higher index) to higher rank (lower index) each time, and are used to re-construct the SVM-based prediction model on the five-fold cross validation test. Finally, the feature subset with the best performance (in terms of ACC) is recognized as the optimal set. The detail of the feature optimization results is discussed in section "Feature analysis."

Classification Algorithm

Support vector machine is a powerful machine learning algorithm for classification, regression as well as other machine learning tasks. It has been successfully applied to a series of supervised learning problems in computational biology (Bu et al., 2018; Zhang et al., 2018; Li and Liu, 2019; Liu and Li, 2019; Liu et al., 2019). The main principle of SVM is to transform the input data into high-dimensional feature space, and then determine the most suitable hyperplane for separating the samples in one class from another. After that, the hyperplane can be used to predict the class of unknown data. In this study, we implemented the SVM algorithm by using the SVM library in Python (version 2.7.15). We chose the radial basis function (RBF) as the kernel function, which can transform the non-linearly separated feature space into higher-dimensional one that is linearly separable. Moreover, we optimized the parameters by grid search to determine the optimal classification hyperplane for SVM algorithm. The classification algorithm optimization results can be seen in section "Classifier Optimization."

Evaluation Metrics and Methods

Four metrics, namely sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthew's correlation coefficient (MCC), were used to

quantitatively evaluate the performance of the proposed method. Their definitions are given below:

$$\begin{cases} \mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \\ \mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \\ \mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FN+FP}} \\ \mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})\,(\mathrm{TP}+\mathrm{FP})\,(\mathrm{TN}+\mathrm{FP})\,(\mathrm{TN}+\mathrm{FN})}} \end{cases} \quad (8)$$

where TP (true positive) represents the number of correctly predicted positive samples; TN (true negative) represents the number of correctly predicted negative samples; FP (false positive) represents the number of negative samples incorrectly predicted to positive samples; FN (false negative) represents the number of positive samples incorrectly predicted to negative samples.

Moreover, we used the five-fold cross validation method to measure the predictive performance of the predictor (Liu, 2019). The procedure of this validation method involves three steps. Firstly, a dataset is randomly partitioned into five equal-size subsets. Of the five subsets, four are chosen as the training dataset for model training, while the remaining one is retained as the validation data to evaluate the performance of the model. After that, this process is repeated until each subset is used exactly once as the validation data. Lastly, the five results are averaged to obtain a final prediction estimation.

To more intuitively evaluate the predictive performance, we also used two curves: receiver operating characteristic (ROC) curve and Precision-Recall (PR) curve. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR; 1-specificity) under different classification thresholds; while the PR curve plots precision (the fraction of TP in all predicted positives) against recall (sensitivity) at various threshold settings. The PR curve is more sensitive to false positives than the ROC curve, especially evaluated on imbalanced dataset. In addition, the area under the ROC curve (AUC) is utilized to quantitatively measure the quality of the predictive model. The range of AUC is 0.5–1. The higher the AUC is, the better the predictor (Hanley and McNeil, 1982).

RESULTS AND DISCUSSION

Classifier Optimization

To achieve the best performance, we conducted the following experiments to optimize the SVM classifier.

Firstly, we did the parameter optimization. There are two parameters in SVM, including the penalty coefficient (denoted as *c*) and gamma (denoted as *g*). We used the grid search strategy to find the optimal values of log2c and log2g in the range (−2 to 5) and (−5 to 2), respectively. **Figure 1A** shows the visualization of the grid search process in three-dimensional space.

Next, we need to determine which kernel function is most suitable for our dataset. There are three kernel functions in SVM, including RBF, Polynomial, and Sigmoid, for handling different feature space. Therefore, we compared the performance of the three kernels. We can observe in **Figure 1B** that the RBF performs better than the other two kernels, with the highest AUC of 0.70. Consequently, the SVM with RBF kernel is used to train the model in our predictor.

Feature Analysis

To in-depth explore the critical information benefiting for the prediction of 5hmC, we conducted a series of feature analysis experiments, including feature combination, optimization, and contribution analysis.

Feature Combination Analysis

In our predictive framework, three feature descriptors, including 2-mer spectrum, 3-mer spectrum, and nucleotide binary features are concatenated to encode RNA sequences. To evaluate their contributions for 5hmC prediction, we compared the performance of different features and that of their combinations. The results are listed in **Table 1**. As can be seen, amongst the three individual feature descriptors, the 3-mer spectrum performs the best than the other two (2-mer spectrum and binary vector). This indicates that the sequential patterns are more useful for 5hmC prediction. By combining 2-mer and 3-mer spectrum, the performance is slightly improved. Particularly, adding binary vector to the combination of 2-mer and 3-mer spectrum, the performance decreases dramatically to 56.1% and 0.122 in terms of ACC and MCC, respectively, which is almost the same with the performance by using binary vector only. The possible reason is that integrating different types of feature space results in mutual information that is not useful for the performance.

Feature Optimization Analysis

To obtain the most discriminative features, we further did the two-stage feature optimization to the integrated feature space. The procedure of the optimization strategy can be seen in section "Methods and Materials." **Figure 2A** illustrates the ACC curve of the predictive model by gradually adding features (from the feature rank list) under the SFS process. As shown in **Figure 2A**, when the feature number reaches to 26, the model achieves the maximum ACC. After reaching the peak, the performance leads to a significant drop as adding more features (see **Figure 2A**). This suggests most of the low-ranked features (binary vector) are relatively irrelevant with the high-ranked features, and even result in a decrease in the performance. The significant improvement by the optimal features is observed, for which the overall performances in terms of ACC and MCC were increased approximately 9.38% and 0.188 after feature optimization. These results demonstrate that feature optimization can effectively enhance the feature representation ability, thereby contributing to the improved performance.

Next, we further compared the spatial distribution of the original feature space and the optimal feature space. For intuitive comparison, we used a visualization tool t-SNE (Maaten and Hinton, 2008) that enable to reduce the feature space to a two-dimensional space. **Figures 2B,C** depict the t-SNE visualization of the original and optimal feature space, respectively. As can be seen from **Figure 2B**, the positive (true 5hmC sites) and negative (non-5hmC sites) samples in the original feature space are mixed up, indicating that the original feature space cannot separate

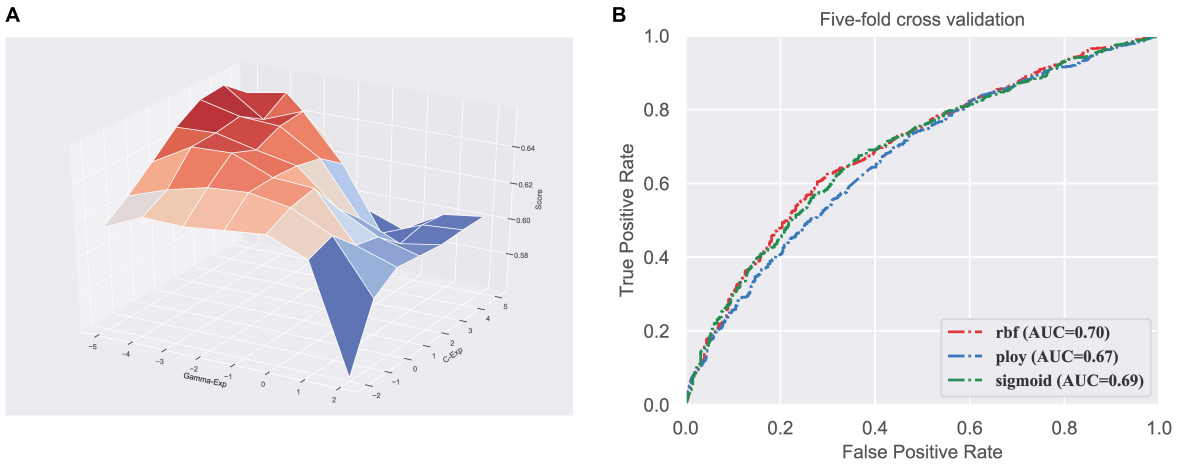


FIGURE 1 | Parameter and kernel optimization of the SVM. **(A)** Visualization of classifier parameter optimization based on grid search; **(B)** ROC curves of different kernels in SVM.

true 5hmC sites from non-5hmC sites well. In contrast, after feature optimization (see **Figure 2C**), the positives and negative samples in feature space are distributed in relatively clear clusters. This demonstrates that feature optimization is able to remove some irrelevant features and learn the most representatives of true 5hmC sites.

Feature Contribution Analysis

To specify which features are important for the prediction of 5hmC, we further analyzed the importance of different features in our feature set. The details regarding how to calculate the feature importance can be referred to section "Feature Optimization." **Figure 2D** illustrates the importance scores (*F*-value) of the top 20 features, and the detail of all the features can be found in **Supplementary Material**. As shown in **Figure 2D**, amongst the top 20 features, most of the features are *k*-mer spectrum (3-mer and 2-mer) while only 4 of the 20 are binary features, indicating that there exist significant compositional differences between the positive and negative samples. In particular, the sequential patterns "GGG" and "GG" are the most important features, indicating that the compositions of the *guanine* (G) nucleotide are discriminative features for the prediction of 5hmC.

TABLE 1 | Five-fold cross validation results of different features and their combinations.

Features	ACC (%)	SN (%)	SP (%)	MCC
A	62.3	61.8	62.8	0.246
B	64.0	63.4	64.5	0.279
C	53.3	53.5	53.2	0.066
A + B	64.0	62.7	65.3	0.280
A + C	55.4	55.9	54.8	0.107
B + C	55.8	57.1	54.5	0.116
A + B + C	56.1	57.6	54.7	0.122

A, 2-mer spectrum; B, 3-mer spectrum; C, binary vector; +, operation of combining.

This observation is different from the fact that DNA 5mC is often found in contexts of CG or C × G (Kumar et al., 2018). We further used Two Sample Logos (TSL) (Vacic et al., 2006), a web-based application to calculate and visualize differences between two sets (the positive and negative) of aligned samples of nucleotides. **Figure 2E** depicts the TSL visualization of the positive and negative samples in our dataset. We observed that the enrichment of nucleotides is significantly different in specific positions along the sequences between the positive and negative samples. For example, the *adenine* (A) nucleotide is enriched at 38th position in the positive set while not in the negative set. This demonstrates that the compositional features might have the positional preference. Therefore, exploring positional features is probably helpful for the further performance improvement.

Comparison of Our Feature Set With Existing Feature Algorithms

In this section, we compared the proposed features and four sequence-based feature descriptors, including PCP (physical–chemical properties), MMI (multivariate mutual information), PseDNC (pseudo dinucleotide composition), and PseEIIP (electron-ion interaction pseudopotentials of trinucleotide). The compared feature descriptors explore sequential information from different aspects. For example, PCP uses the physical–chemical properties of dinucleotides and explores the correlation between any two nucleotides using auto-covariance and cross covariance transformations (Liu et al., 2015; Wei et al., 2019). MMI calculates the multivariate mutual information of nucleotides (Wei et al., 2019). Pse-DNC can capture the local and global characteristic patterns by integrating the sequence-order information with PCP (Chen et al., 2014). More details of the feature descriptors can be referred to (Wei et al., 2019). We evaluated all the feature descriptors including our feature set on the same data set with five-fold cross validation. Since our feature set is optimized using the feature optimization strategy, for the purpose of fair comparison, we also used the same strategy to

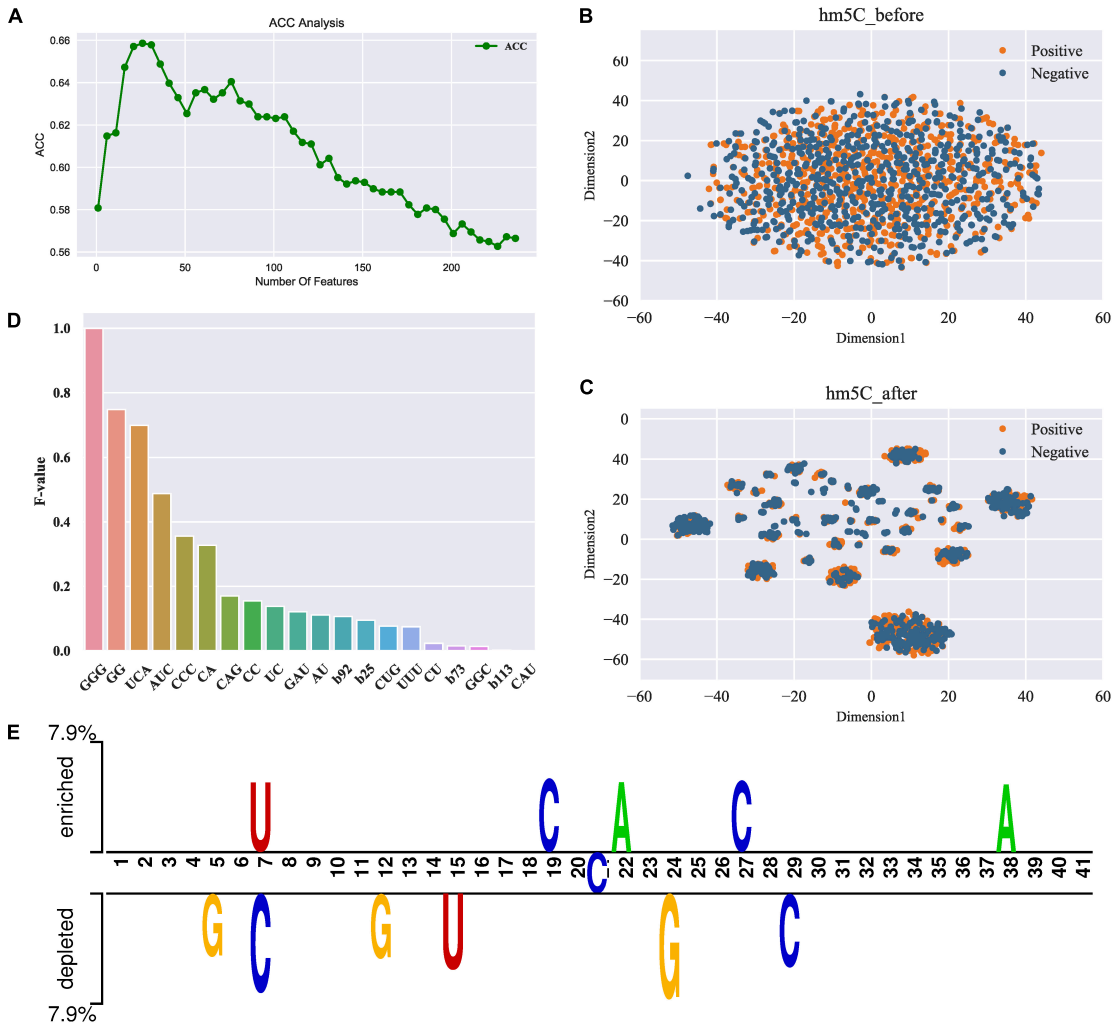


FIGURE 2 | Feature analysis results. **(A)** ACC curve of the feature selection; **(B,C)** represent the distribution visualization of the samples (positive and negative) in feature space before and after feature optimization, respectively; **(D)** *F*-values of the top 20 most important features. Note that the *x*-axis represents the specific features and the *y*-axis represents the *F*-value. Note that b92 denotes the 92th feature of the binary vector, b25 denotes the 25th feature, and so forth; **(E)** TSL (Two Sample Logos) visualization of the positives and negatives in the dataset used in this study.

optimize the four compared feature descriptors. The obtained results by using different features were reported in **Table 2**.

As seen in **Table 2**, our feature set performs better than other sequence-based feature descriptors in terms of ACC and MCC, with exceptions of SN and SP. The ACC and MCC of our feature set is 65.48% and 0.31, respectively, which are 1.2% and 0.023 higher than that of the runner-up feature descriptor – PseEIIP, with the ACC of 64.27% and MCC of 0.2872. It is worth

TABLE 2 | Five-fold cross validation results of the proposed feature set with other sequence-based feature descriptors.

Features	ACC (%)	SN (%)	SP (%)	MCC
PCP	63.97	68.73	59.21	0.2807
MMI	61.56	63.14	59.97	0.2312
PseDNC	62.84	61.33	64.35	0.2569
PseEIIP	64.27	69.64	58.91	0.2872
Our feature set	**65.48**	67.67	63.29	**0.3100**

The bold value indicates the highest value of this column.

TABLE 3 | Comparative results of SVM and four well-known classifiers on the dataset used in this study.

Classifiers	ACC (%)	SN (%)	SP (%)	MCC
GBDT	63.60	63.90	63.29	0.2719
KNN	58.46	56.95	59.97	0.1693
NB	63.37	63.00	63.75	0.2674
RF	60.27	62.08	58.46	0.2056
SVM (this study)	**65.48**	67.67	63.29	**0.3100**

The bold value indicates the highest value of this column.

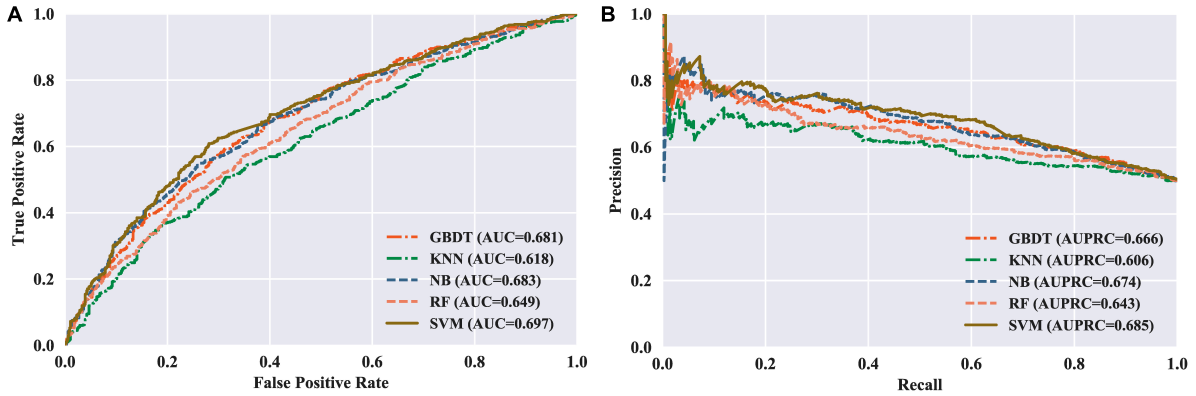


FIGURE 3 | Performance of different classifiers evaluated with five-fold cross validation. **(A)** ROC curves of different classifiers. **(B)** PR curves of different classifiers.

noting that our SN and SP are 67.67% and 63.29%, slightly worse than the best descriptor – PseEIIP in SN and PseDNC in SP, respectively. Although our SN and SP are not the best, they are more balanced as compared to PseEIIP and PseDNC, thus contributing to the highest overall performance. This indicates that our feature set is more effective to distinguish true 5hmC sites from non-5hmC sites. In addition, since the majority of our feature set is *k*-mer spectrum features, this also demonstrates that the sequential patterns is capable of better capturing the characteristics of 5hmC sites as compared to other information like PCP and nucleotide mutual information, and so on.

Comparison With Different Classification Algorithms

To measure the effectiveness of SVM, we compared its performance with multiple well-known classifiers, like gradient boosting decision tree (GBDT) (Liao et al., 2018), k-nearest neighbor (KNN), logistic regression (LR), naive Bayes (NB) (Feng et al., 2013), and random forest (RF) (Lv et al., 2019; Ru et al., 2019; Wei et al., 2017). For fair comparison, we trained the classifiers on the same dataset with our feature set, and then fine-tuned the classifiers one by one to achieve the optimal performance. The models are also evaluated by five-fold cross validation, and the evaluation results are presented in **Table 3**. We can see that the SVM achieves ACC of 65.48%, SN of 67.67%, SP of 63.29%, and MCC of 0.31, respectively, outperforming the other four classifiers in two out of the four metrics: MCC and ACC. To be specific, our ACC and MCC are higher than that of the runner-up GBDT by 1.88% and 0.0381, respectively. Additionally, we further intuitively compared the performance of different classifiers using ROC and PR curves as shown in **Figures 3A,B**, respectively. The results demonstrate that the SVM classifier has the better discriminative power to distinguish the 5hmC sites from non-5hmC sites than the other four classifiers in this study.

Webserver Implementation

For the convenience of researchers, we established an easy-to-use webserver that implements our predictor, which is freely available at http://server.malab.cn/iRNA5hmC. Below, we give researchers a step-by-step guideline on how to use the webserver to get the desired prediction results. Firstly, users need to submit their query RNA sequences into the input box. Note that the input sequences should be in FASTA format. After that, users can specify the prediction confidence from 0 to 1. Otherwise, under default setting, the query sequence is predicted as true 5hmC sequence if the prediction confidence is >0.5. Afterward, clicking on the "Submit" button, users can obtain the desired results on the screen of the computer.

CONCLUSION

In this study, we have proposed a computational predictor namely iRNA5hmC to predict RNA 5hmC sites using machine learning. To the best of our knowledge, this is the first RNA 5hmC predictor that enables to make predictions based on RNA primary sequences only, without any other prior experimental knowledge. In particular, we have established an easy-to-use webserver for researchers to make the proposed predictor more impactful and have the potential to be complementary tool to the high-throughput sequencing methods. However, we have to see there still has some aspects, such as the relatively low predictive performance, and small-size dataset, which need to be improved in our future work.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the http://server.malab.cn/iRNA5hmC.

AUTHOR CONTRIBUTIONS

LW, WC, and RS conceived and designed the experiments. WC acquired the experiment data. DC and YL performed the study. DC and RS carried out the data analysis. YL, RS, and LW wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

The work was supported by the National Natural Science Foundation of China (Nos. 31771471, 61701340, and 61702361).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00227/full#supplementary-material

REFERENCES

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Bu, H. D., Hao, J. Q., Guan, J. H., and Zhou, S. G. (2018). Predicting enhancers from multiple cell lines and tissues across different developmental stages based on SVM method. *Curr. Bioinform.* 13, 655–660. doi: 10.2174/1574893613666180726163429

Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K.-C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 7, 16895–16909. doi: 10.18632/oncotarget.7815

Chen, W., Feng, P., Song, X., Lv, H., and Lin, H. (2019). iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol. Therapy Nucleic Acids* 18, 269–274. doi: 10.1016/j.omtn.2019.08.022

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.-C. (2014). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* 31, 119–120. doi: 10.1093/bioinformatics/btu602

Conde, J., Yoon, J.-H., Choudhury, J. R., Prakash, L., and Prakash, S. (2015). Genetic control of replication through N1-methyladenine in human cells. *J. Biol. Chem.* 290, 29794–29800. doi: 10.1074/jbc.M115.693010

Delatte, B., Wang, F., Ngoc, L. V., Collignon, E., Bonvin, E., Deplus, R., et al. (2016). RNA biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* 351, 282–285. doi: 10.1126/science.aac5253

Feng, P.-M., Lin, H., and Chen, W. (2013). Identification of antioxidants from sequence information using Naive Bayes. *Comput. Math. Methods Med.* 2013:567529. doi: 10.1155/2013/567529

Fu, L., Guerrero, C. R., Zhong, N., Amato, N. J., Liu, Y., Liu, S., et al. (2014). Tet-mediated formation of 5-hydroxymethylcytosine in RNA. *J. Am. Chem. Soc.* 136, 11582–11585. doi: 10.1021/ja505305z

Hanley, J. A., and McNeil, B. J. J. R. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Huber, S. M., van Delft, P., Mendil, L., Bachman, M., Smollett, K., Werner, F., et al. (2015). Formation and abundance of 5-hydroxymethylcytosine in RNA. *Chembiochem* 16, 752–755. doi: 10.1002/cbic.201500013

Jonkhout, N., Tran, J., Smith, M. A., Schonrock, N., Mattick, J. S., and Novoa, E. M. (2017). The RNA modification landscape in human disease. *RNA* 23, 1754–1769. doi: 10.1261/rna.063503.117

Kumar, S., Chinnusamy, V., and Mohapatra, T. (2018). Epigenetics of modified DNA bases: 5-methylcytosine and beyond. *Front. Genet.* 9:640. doi: 10.3389/fgene.2018.00640

Li, C.-C., and Liu, B. (2019). MotifCNN-fold: protein fold recognition based on Fold-specific features extracted by motif-based convolutional neural networks. *Brief. Bioinform.* doi: 10.1093/bib/bbz133 [Epub ahead of print].

Li, W., and Liu, M. (2011). Distribution of 5-hydroxymethylcytosine in different human tissues. *J. Nucleic Acids* 2011:870726. doi: 10.4061/2011/870726

Liao, Z. J., Wan, S. X., He, Y., and Zou, Q. (2018). Classification of Small GTPases with hybrid protein features and advanced machine learning techniques. *Curr. Bioinform.* 13, 492–500. doi: 10.2174/1574893612666171121162552

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

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:e127. doi: 10.1093/nar/gkz740

Liu, B., and Li, K. (2019). iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids* 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Miao, Z., Xin, N., Wei, B., Hua, X., Zhang, G., Leng, C., et al. (2016). 5-hydroxymethylcytosine is detected in RNA from mouse brain tissues. *Brain Res.* 1642, 546–552. doi: 10.1016/j.brainres.2016.04.055

Pian, C., Zhang, G., Li, F., and Fan, X. (2019). MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* 36, 388–392. doi: 10.1093/bioinformatics/btz556

Racz, I., Kiraly, I., and Lasztily, D. (1978). Effect of light on the nucleotide composition of rRNA of wheat seedlings. *Planta* 142, 263–267. doi: 10.1007/BF00385075

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

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

Shi, H., Wei, J., and He, C. (2019). Where, when, and how: context-dependent functions of RNA methylation writers, readers, and erasers. *Mol. cell* 74, 640–650. doi: 10.1016/j.molcel.2019.04.025

Vacic, V., Iakoucheva, L. M., and Radivojac, P. (2006). Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22, 1536–1537. doi: 10.1093/bioinformatics/btl151

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z., and Zou, Q. (2017). CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Whitney, A. W. (2006). A direct method of nonparametric measurement selection. *IEEE Trans. Comput. C* 20, 1100–1103. doi: 10.1109/t-c.1971.223410

Yuan, F., Bi, Y., Siejka-Zielinska, P., Zhou, Y.-L., Zhang, X.-X., and Song, C.-X. (2019). Bisulfite-free and base-resolution analysis of 5-methylcytidine and 5-hydroxymethylcytidine in RNA with peroxotungstate. *Chem. Commun.* 55, 2328–2331. doi: 10.1039/c9cc00274j

Zhang, N., Yu, S., Guo, Y., Wang, L., Wang, P., and Feng, Y. (2018). Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinform.* 13, 50–56. doi: 10.2174/1574893611666160608102537

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 © 2020 Liu, Chen, Su, Chen and Wei. 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 March 2020
doi: 10.3389/fcell.2020.00184

Characterization of Dysregulated lncRNA-Associated ceRNA Network Reveals Novel lncRNAs With ceRNA Activity as Epigenetic Diagnostic Biomarkers for Osteoporosis Risk

*Meijie Zhang†, Luyang Cheng† and Yina Zhang**

Department of Geriatrics, The Second Affiliated Hospital of Harbin Medical University, Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Jie Zhang,
Tongji University, China
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Yina Zhang
yinazhlu@163.com

†*These authors share first authorship*

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

Received: *03 February 2020*
Accepted: *04 March 2020*
Published: *31 March 2020*

Citation:
Zhang M, Cheng L and Zhang Y (2020) Characterization of Dysregulated lncRNA-Associated ceRNA Network Reveals Novel lncRNAs With ceRNA Activity as Epigenetic Diagnostic Biomarkers for Osteoporosis Risk. Front. Cell Dev. Biol. 8:184. doi: 10.3389/fcell.2020.00184

The altered expression of long non-coding RNAs (lncRNAs) has been implicated in the development and human diseases. However, functional roles and regulatory mechanisms of lncRNA as competing endogenous RNAs (ceRNAs) in osteoporosis and their potential clinical implication for osteoporosis risk are largely unexplored. In this study, we performed integrated analysis for paired expression profiles and regulatory relationships of dysregulated lncRNAs, mRNAs, and miRNAs based on "ceRNA hypothesis," and constructed an osteoporosis-related dysregulated miRNA-mediated lncRNA–mRNA ceRNA network (DysCeNet) composed of 105 nodes (including eight miRNAs, 24 mRNAs, and 73 lncRNAs) and 515 edges. Functional analysis suggested that the DysCeNet was involved in known osteoporosis or bone metabolism-related biological processes and pathways. Then, we performed random forest-based feature selection for 73 lncRNAs with ceRNA activity and identified 25 of 73 lncRNAs as potential diagnostic biomarkers. A random forest-based classifier composed of 25 lncRNA biomarkers (RF-25lncRNA) was developed for predicting osteoporosis risk. Performance evaluation with the leave-one-out cross-validation (LOOCV) procedure showed that the RF-25lncRNA achieved a good performance in distinguishing high- and low-bone mineral density (BMD) subjects in different osteoporosis datasets. Our study for the first time revealed a global view of lncRNA-associated ceRNA regulation in osteoporosis and provided novel lncRNAs with ceRNA activity as candidate epigenetic diagnostic biomarkers for early detection of osteoporosis risk.

Keywords: competing endogenous RNAs, long non-coding RNAs, osteoporosis, biomark, epigenetics

INTRODUCTION

Osteoporosis is a progressive systemic skeletal disease with low bone density and deterioration of bone architecture (Garnero, 2017; Tu et al., 2018). It is estimated that over 14 million men and women in the United States will have osteoporosis by 2020 (Burge et al., 2007). People with osteoporosis substantially have an increased risk of bone fragility and fracture, leading to increase

pain, disability, nursing home placement, total health care costs, and mortality (Tu et al., 2018). Early detection and intervention for osteoporosis risk are an effective way to delay the development of disease and improve the quality of life of patients. Therefore, there is an urgent need to identify novel molecular biomarkers in the clinical assessment of osteoporosis risk.

With the development and application of high-throughput omics technologies, it has been shown that osteoporosis has genetic and molecular heterogeneity like many other common complex diseases, and disruption in some molecular pathways can disturb the equilibrium of bone turnover and thereby contribute to osteoporosis (Al Anouti et al., 2019). Long non-coding RNAs (lncRNAs) are a newly discovered class of non-coding RNAs (ncRNAs) that are longer than 200 bp (Pelechano and Steinmetz, 2013) and have attracted much attention in recent medical studies. A large number of studies have demonstrated that lncRNAs is an important player of the genomic regulatory network and is involved in a wide variety of biological progress (Mercer and Mattick, 2013; Guo et al., 2016; Quinn and Chang, 2016; Bunch, 2018; Sun et al., 2020). lncRNAs have been attributed to various functions in development, differentiation, and human disease by negatively or positively regulating gene expression at the transcriptional, post-transcriptional, and epigenetic levels (Kornienko et al., 2013; Maass et al., 2014). Growing functional roles of lncRNAs has been highlighted in osteoporosis during recent studies. For example, *XIST*, a well-known major effector of the X-inactivation process, was recently reported to promote osteoporosis through inhibiting bone marrow mesenchymal stem cell differentiation (Chen et al., 2019). Mei et al. (2019) found that lncRNA *ZBTB40-IT1* played a critical role in bone metabolism and can be modulated by osteoporosis GWAS risk SNPs suppresses osteogenesis. It has become increasingly clear that lncRNAs can act as competing endogenous RNAs (ceRNAs) to interact other RNA molecules by competing for binding to shared microRNAs, which has been implicated in development and human disease including osteogenesis (Tay et al., 2014; Huang et al., 2019a,b; Silva et al., 2019). However, genome-wide exploration for miRNA-mediated lncRNA-associated ceRNA mechanism in osteoporosis and their potential clinical implication for osteoporosis risk remained largely unknown.

In this study, we tried to construct a global miRNA-mediated lncRNA–mRNA ceRNA network in the development of osteoporosis by integrating paired expression profiles and regulatory relationship of lncRNAs, mRNAs, and miRNAs based on "ceRNA hypothesis," and further to uncover novel lncRNAs with ceRNA activity as epigenetic diagnostic biomarkers for identifying people at high risk for developing osteoporosis.

MATERIALS AND METHODS

Patients and Samples

Ten samples [including five high-bone mineral density (BMD) subjects and five low-BMD subjects] with corresponding transcriptome gene expression microarray data (Affymetrix Human Exon 1.0 ST Array) and epigenomic miRNA microarray data (Affymetrix Multispecies miRNA-2 Array) were obtained from the Gene Expression Omnibus (GEO) database (the accession number is GSE62589[1]). Two other osteoporosis datasets with transcriptome gene expression microarray data were obtained from the GEO database, including GSE56814 dataset (16 high-BMD subjects and 15 low-BMD subjects) (the accession number is GSE56814[2]) and GSE13850 dataset (10 high-BMD subjects and 10 low-BMD subjects) (the accession number is GSE13850[3]).

Acquisition and Analysis of Expression Profiles

Raw transcriptome gene expression microarray data (CEL files) profiled on Affymetrix GeneChip Human Exon 1.0 ST Array (HuEx-1_0-st) and Affymetrix Human Genome U133A Array (HG-U133A) were obtained from the GSE63402 and GSE13850. These raw data were processed and normalized using the Robust Multichip Average (RMA) algorithm of R package "oligo" for background subtraction, quantile normalization, and summarization. Then, all probes of microarray were mapped into protein-coding genes using the R package "biomaRt" (Durinck et al., 2009). LncRNA expression profiles were obtained using repurposing strategy by mapping array probes into the human genome (GRCh 38) and lncRNA annotations from the GENCODE database[4] (Zhou et al., 2016) using SeqMap tool (Jiang and Wong, 2008). Human mature miRNAs were retrieved from the miRNA microarray. Finally, expression profiles of 20,068 mRNA, 7821 lncRNAs, and 1100 miRNA were obtained for further analysis.

Differential expression analysis of lncRNAs, miRNAs, and mRNAs between high- and low-BMD subjects was performed using the R package "limma" (Ritchie et al., 2015). Those lncRNAs, miRNAs, and mRNAs with $p < 0.05$ were considered as differentially expressed genes. Hierarchical clustering was performed to investigate the expression patterns between high- and low-BMD subjects.

Construction of Dysregulated lncRNA-Associated ceRNA Network

The experimentally validated miRNA–mRNA and miRNA–lncRNA interaction data were collected from the TarBase database[5] (Li et al., 2014). The dysregulated lncRNA-associated ceRNA network (DysCeNet) in osteoporosis was constructed based on the "ceRNA hypothesis" as follows: (i) Pearson correlation coefficient (PCC) was calculated to measure the expression correlation between differentially expressed mRNAs and lncRNAs from matched mRNA and lncRNA expression profiles. Those dysregulated lncRNA–mRNA pairs with PCC > 0.5 were considered as candidate co-dysregulated lncRNA–mRNA ceRNA crosstalk; (ii) expression correlation between differentially expressed miRNAs and differentially

[1]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62589
[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56814
[3]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13850
[4]https://www.gencodegenes.org/
[5]http://starbase.sysu.edu.cn/

expressed mRNAs, and between differentially expressed miRNAs and differentially expressed lncRNAs was evaluated using PCC; (iii) for a candidate co-dysregulated lncRNA–mRNA ceRNA crosstalk, both mRNAs and lncRNAs in this lncRNA–mRNA ceRNA crosstalk are targeted and co-expressed negatively with a common miRNA; this candidate co-dysregulated lncRNA–mRNA ceRNA crosstalk was selected as dysregulated miRNA-mediated lncRNA–mRNA ceRNA crosstalk; (iv) all miRNA-mediated lncRNA–mRNA ceRNA crosstalks were integrated to form a global DysCeNet.

Identification of lncRNA Biomarkers With ceRNA Activity Using a Machine Learning Method

To identify potential lncRNA biomarkers with ceRNA activity, a random forest approach and leave-one-out cross-validation (LOOCV) were used to select optimal lncRNAs biomarkers using the R package "randomForest" and out-of-bag (OOB) error, which measure the performance of the model on the training set (Lv et al., 2019; Tan et al., 2019). The OOB error will produce an unbiased estimate for the classification error, while the bagging method will decrease the chance of overfitting (Toth et al., 2019). Then, a random forest-based classifier was built using the optimal lncRNA biomarkers, and a receiver operating characteristic (ROC) curve and the area under ROC curve (AUC) was used to measure the diagnostic performance of the lncRNA classifier (Lai et al., 2019).

Functional Enrichment Analysis

Functional enrichment analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) for mRNAs in the DysCeNet was conducted to predict potential biological functions of lncRNAs in the DysCeNet using the R package "clusterprofiler" (Yu et al., 2012). Those significantly enriched GO terms with $p < 0.05$ with mutually overlapping gene sets were clustered together using the Enrichment Map plugin in Cytoscape environment (Merico et al., 2010).

RESULTS

Identification of Differentially Expressed mRNAs, miRNAs, and lncRNAs Associated With Osteoporosis

To identify potential risk mRNAs, miRNAs, and lncRNAs in osteoporosis, we performed a comparative analysis for expression profiles of mRNAs, miRNAs, and lncRNAs between high- and low-BMD subjects. A total of 68 mRNAs, 11 miRNAs, and 95 lncRNAs were identified as differentially expressed ($p < 0.05$) in high-BMD subjects compared with low-BMD

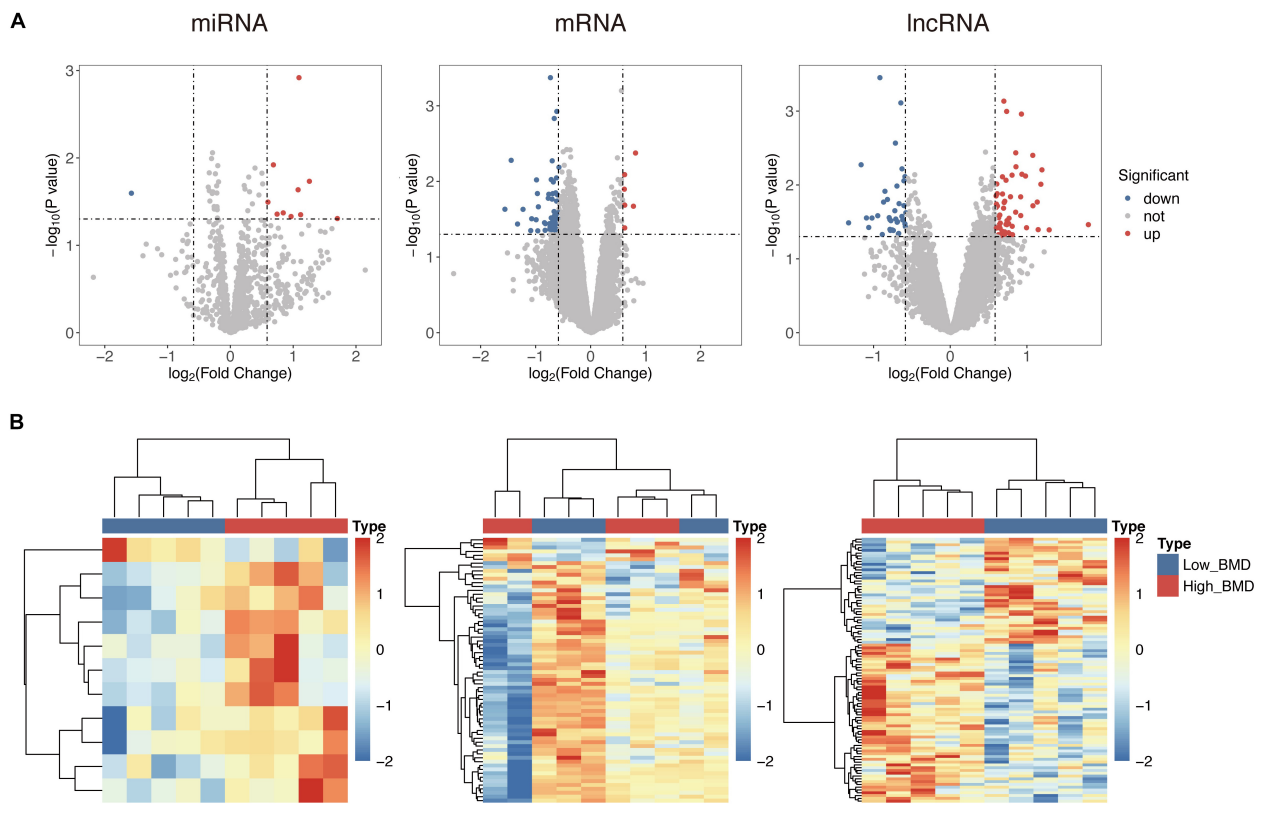


FIGURE 1 | Differential expressed analysis between high- and low-BMD subjects. **(A)** Volcano plot of the distribution of differentially expressed mRNAs, miRNAs, and lncRNAs. **(B)** Hierarchical clustering heatmap and dendrogram of 10 samples based on differentially expressed mRNAs, miRNAs, and lncRNAs.

subjects (**Supplementary Table S1**). Among them, there were 73 unregulated genes (including six mRNAs, 10 miRNAs, and 57 lncRNAs) and 101 downregulated genes (including 62 mRNAs, one miRNA, and 38 lncRNAs) in high-BMD subjects compared with low-BMD subjects (**Figure 1A**). Hierarchical clustering analysis showed that expression patterns of these differentially expressed mRNAs, miRNAs, and lncRNAs were capable of distinguishing high-BMD subjects from low-BMD subjects (**Figure 1B**).

Construction and Analysis of Dysregulated lncRNA-Associated ceRNA Network

To construct an osteoporosis-related dysregulated miRNA-mediated lncRNA–mRNA ceRNA network, we performed an integrated analysis for paired expression profiles and regulatory relationships of dysregulated lncRNAs, mRNAs, and miRNAs of 10 samples in the GSE62589 as described in Section "Materials and Methods." Finally, an osteoporosis-related dysregulated miRNA-mediated lncRNA–mRNA ceRNA network was constructed and was composed of 105 nodes and 515 edges (including eight miRNAs, 24 mRNAs, and 73 lncRNAs) (**Figure 2A**) (**Supplementary Table S2**). To further explore the functional implication of the DysCeNet, we performed functional enrichment analysis for mRNAs in the DysCeNet and found that mRNAs in the DysCeNet were significantly enriched in blood vessel development, NIK/NF-kappaB signaling, bone mineralization involved in bone maturation, osteoclast differentiation, and cell aging.

Identification of Potential lncRNA Biomarkers for the Osteoporosis Risk

To identify potential lncRNA biomarkers for the osteoporosis risk, we performed feature selection for lncRNAs in the DysCeNet using a random forest model. Finally, 25 lncRNAs of 73 lncRNAs in the DysCeNet were identified as potential biomarkers for the osteoporosis risk according to their discriminative power using OOB error (**Table 1**). Of them, 10 lncRNA biomarkers are up-regulated and 15 lncRNA biomarkers are down-regulated in high-BMD subjects compared

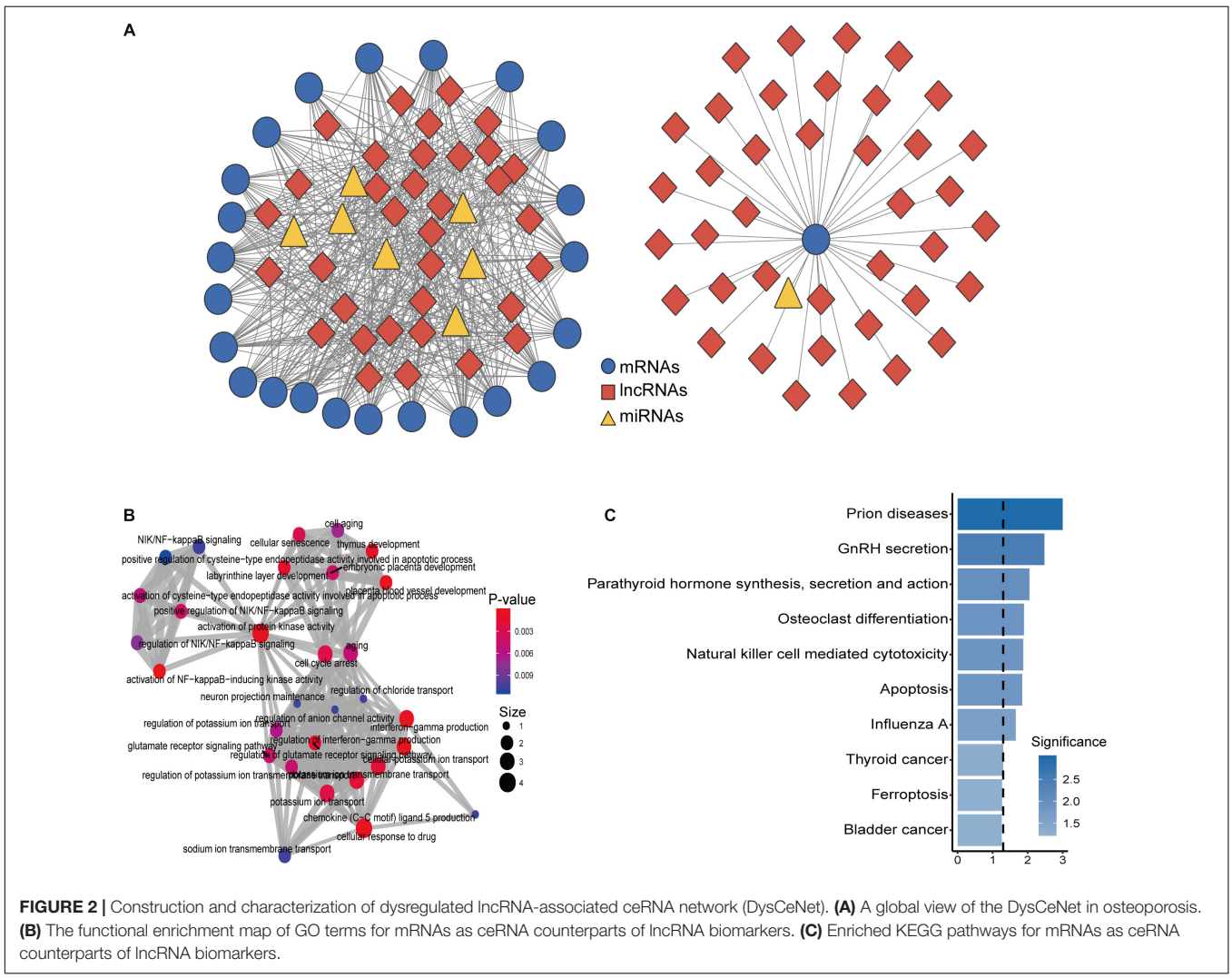


FIGURE 2 | Construction and characterization of dysregulated lncRNA-associated ceRNA network (DysCeNet). **(A)** A global view of the DysCeNet in osteoporosis. **(B)** The functional enrichment map of GO terms for mRNAs as ceRNA counterparts of lncRNA biomarkers. **(C)** Enriched KEGG pathways for mRNAs as ceRNA counterparts of lncRNA biomarkers.

TABLE 1 | Genomic information of 25 lncRNA biomarkers with ceRNA activity for osteoporosis risk.

Ensembl id	Gene name	Location	*p*-value[a]
ENSG00000228290	RP11-132M7.3	Chr 6: 85,399,143-85,419,252 (+)	0.0008
ENSG00000258610	RP11-488C13.7	Chr 14: 77,245,295-77,248,592 (+)	0.0004
ENSG00000242009	RP11-359D24.1	Chr 3: 81,889,698-81,972,012 (+)	0.0078
ENSG00000254562	RP11-64I17.1	Chr 11: 38,668,026-38,695,791 (+)	0.0011
ENSG00000224239	AC090044.2	Chr 3: 898,807-899,774 (+)	0.0063
ENSG00000279785	AC008267.8	Chr 7: 66,474,556-66,475,749 (+)	0.0007
ENSG00000264108	RP11-357H3.1	Chr 18: 73,720,073-73,727,816 (–)	0.0027
ENSG00000227885	RP11-79N23.1	Chr 6: 53,605,547-53,611,820 (–)	0.0088
ENSG00000272599	RP11-152N13.16	Chr 10: 74,884,331-74,885,163 (–)	0.0053
ENSG00000236145	AC079988.3	Chr 2: 121,935,903-121,940,156 (+)	0.0061
ENSG00000255471	RP11-736K20.5	Chr 11: 86,603,256-86,636,079 (–)	0.0296
ENSG00000203496	AL133216.1	Chr 10: 38,453,181-38,466,176 (+)	0.0444
ENSG00000250893	AC098869.2	Chr 4: 40,426,119-40,427,585 (+)	0.0299
ENSG00000278990	AL132796.3	Chr 14: 98,916,902-98,917,801 (–)	0.0098
ENSG00000235890	TSPEAR-AS1	Chr 21: 44,506,807-44,516,575 (+)	0.0414
ENSG00000235823	OLMALINC	Chr 10: 100,373,099-100,454,043 (+)	0.0207
ENSG00000253210	AC040970.1	Chr 8: 141,124,717-141,130,064 (–)	0.0281
ENSG00000228705	LINC00659	Chr 20: 62,774,121-62,776,996 (–)	0.0193
ENSG00000236054	Z82246.1	Chr 22: 32,583,300-32,584,204 (+)	0.0097
ENSG00000251329	LINC02353	Chr 4: 32,351,038-32,353,220 (+)	0.0171
ENSG00000234584	AC019186.1	Chr 2: 157,725,708-157,736,005 (–)	0.0104
ENSG00000237017	AC245052.4	Chr 19: 54,119,511-54,125,343 (–)	0.0191
ENSG00000230753	ZNF341-AS1	Chr 20: 33,787,373-33,811,109 (–)	0.0077
ENSG00000240591	AL096701.3	Chr 22: 31,452,881-31,464,304 (+)	0.0179
ENSG00000226649	AC019118.1	Chr 2: 3,156,756-3,157,797 (+)	0.0087

[a]*p-value was obtained from differential expression analysis.*

with low-BMD subjects (**Figure 3A**). To test whether these 25 lncRNA biomarkers could efficiently distinguish high- and low-BMD subjects, we performed hierarchical clustering for 10 samples in the GSE62589 according to the expression pattern of 25 lncRNA biomarkers. As shown in **Figure 3B**, all 10 samples were classified into two clusters according to the expression pattern of seven lncRNA biomarkers with 100% accuracy. The results of hierarchical clustering demonstrated the potential of 25 lncRNAs as diagnostic biomarkers for osteoporosis risk. Therefore, a random forest-based classifier composed of 25 lncRNA biomarkers was developed. The performance of the RF-25lncRNA for predicting osteoporosis risk was evaluated in the GSE62589 dataset using the LOOCV procedure, in which nine samples were used as the training set and the remaining one was served as the test sample. Results of performance evaluation showed that the RF-25lncRNA achieves a perfect predictive performance in distinguishing high- and low-BMD subjects with an AUC of 1.0 (**Figure 3C**).

Further Validation of lncRNA Biomarkers for the Osteoporosis Risk in Two Other Independent Datasets

To test the robustness of the lncRNA biomarkers for the osteoporosis risk, 25 lncRNA biomarkers were applied to the independent GSE56814 dataset. The RF-25lncRNA correctly classified 11 of 15 low-BMD subjects and nine of 16 high-BMD subjects, achieving an AUC of 0.733 (95% CI: 0.553–0.913) with an accuracy of 64.5% (**Figure 4A**).

The performance of the lncRNA biomarkers was further tested in another independent GSE13850 dataset. However, transcriptome gene expression data of the GSE13850 was profiled on the HG-U133A platform, and only two lncRNAs (*RP11-488C13.7* and *RP11-152N13.16*) of 25 lncRNA biomarkers in the RF-25lncRNA were covered on the HG-U133A platform. When two lncRNA biomarkers (*RP11-488C13.7* and *RP11-152N13.16*) were applied to the 20 samples of the GSE13850, two lncRNA biomarkers (*RP11-488C13.7* and *RP11-152N13.16*) correctly classified seven of 10 low-BMD subjects and seven of 10 high-BMD subjects, achieving an AUC of 0.66 (95% CI: 0.396–0.924) with an accuracy of 70% (**Figure 4B**).

DISCUSSION

Early detection and intervention for osteoporosis are crucial to prevent fragility fractures and delay disease progression. Traditional markers for osteoporosis risk are BMD, vitamin D, alkaline phosphatase, and so on (Parveen et al., 2019). Increasing evidence has suggested that altered molecular profiles contributed to the osteoporosis and outcome, which provided novel insights into molecular basis of osteoporosis and also highlighted the potential of molecular factors as markers for

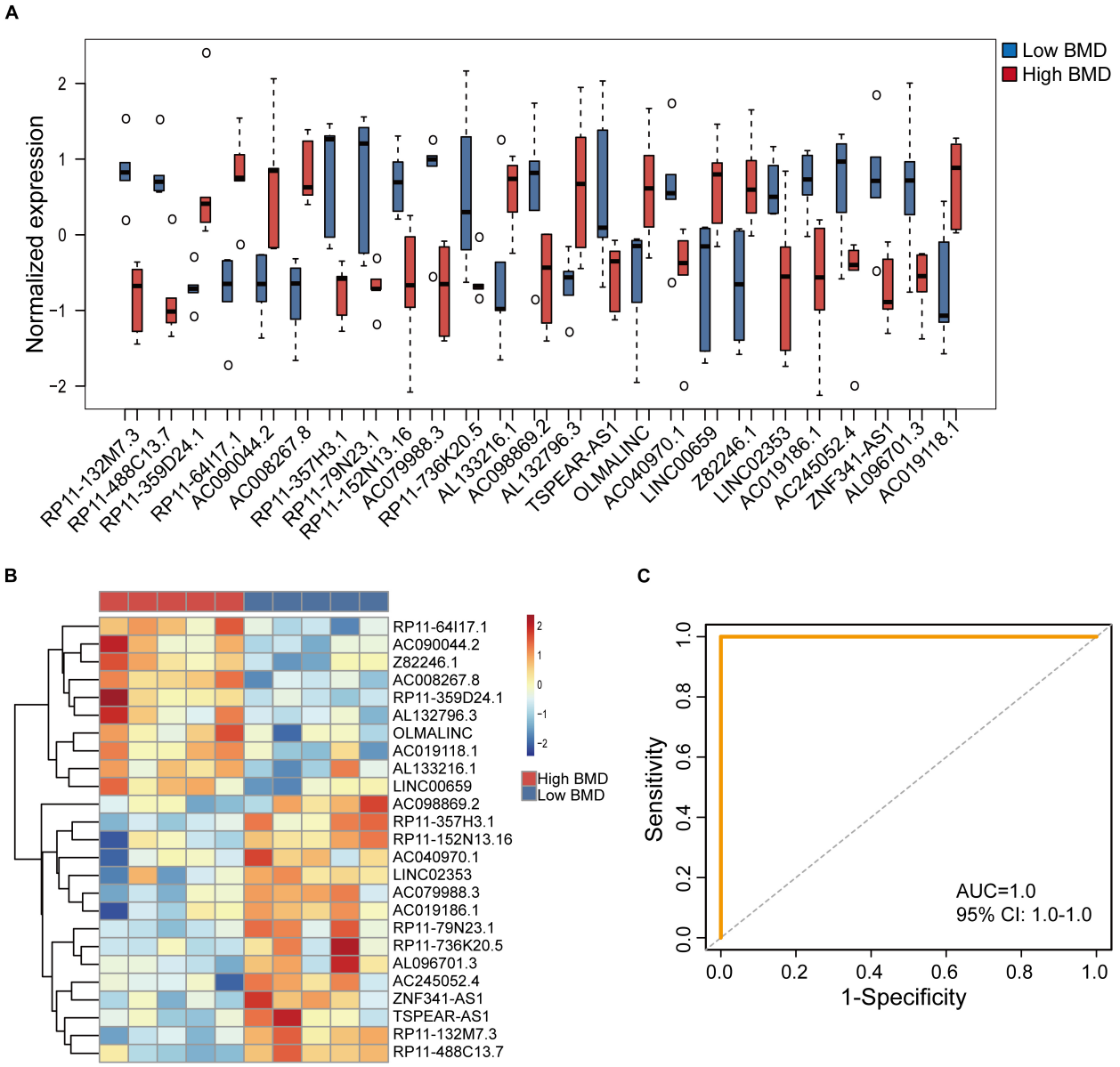


FIGURE 3 | Performance evaluation of identified 25 lncRNA biomarkers in the discovery dataset. **(A)** Expression pattern of 2525 lncRNA biomarkers in high-BMD and low-BMD subjects. **(B)** Hierarchical clustering heatmap and dendrogram of 10 samples based on expression patterns of 25 lncRNA biomarkers. **(C)** The ROC curves of RF-25lncRNA in the GSE62589 dataset.

osteoporosis diagnosis and prognosis (Yu et al., 2013; Makitie et al., 2018; Gong et al., 2019; Ukon et al., 2019). Biomarker identification has been proven to be an effective way to recognize people at high risk for developing osteoporosis and have attracted much attention in the clinical decision-making for osteoporosis management. Previous studies have focused on mRNA profiles and miRNA profiles and identified some candidate mRNA- or miRNA-based biomarkers. For example, circulating miRNAs, *hsa-miR-122-5p* and *hsa-miR-4516*, have been found to be diagnostic biomarkers for osteoporosis risk (Mandourah et al., 2018). Another study by Seeliger et al. (2014) also identified five freely circulating miRNAs associated with osteoporosis fractures. There is increasing evidence that lncRNAs also play important roles in the pathogenesis of osteoporosis (Peng et al., 2018; Feng et al., 2019; Liu et al., 2019). In this study, we obtained lncRNA profiles of osteoporosis patients by repurposing array probes on publicly available microarray data and performed genome-wide analysis for expression patterns of lncRNAs, miRNAs, and mRNAs between high- and low-BMD subjects. A total of 68 mRNAs, 11 miRNAs, and 95 lncRNAs were found to be associated with osteoporosis, which provided a candidate resource for

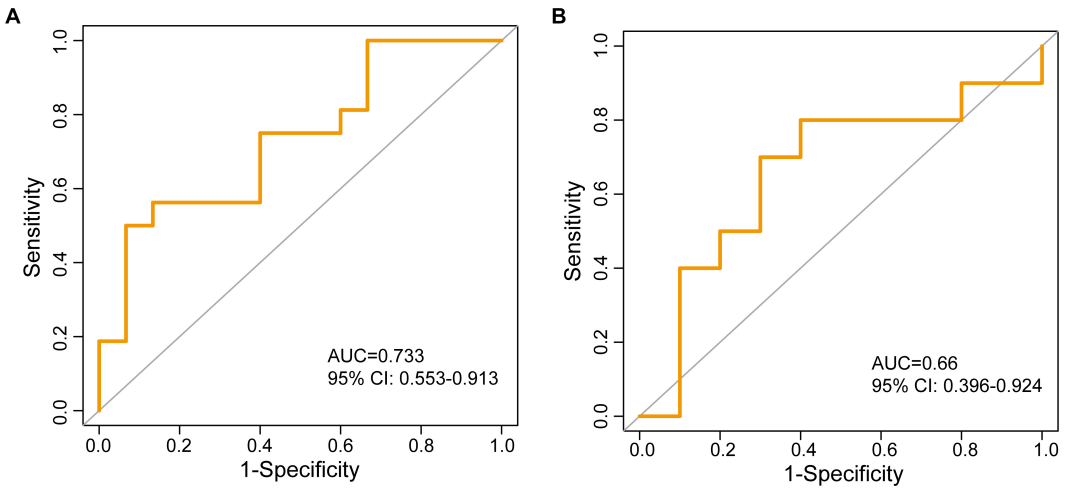


FIGURE 4 | Independent validation of identified 25 lncRNA biomarkers in two independent testing datasets. **(A)** The ROC curves of RF-25lncRNA in the GSE56814 dataset. **(B)** The ROC curves of two lncRNA biomarkers (*RP11-488C13.7* and *RP11-152N13.16*) in the GSE13850 dataset.

experimental scientist for studying the molecular mechanism of osteoporosis.

Through many candidate osteoporosis-related mRNAs, miRNAs, and lncRNAs were identified in our study, regulatory relationships and mechanisms among these different types of RNA molecules in the development of osteoporosis are still unknown. It is well known that different RNA molecules can act as ceRNAs to communicate with and co-regulate each other by competing for binding to shared miRNAs (Tay et al., 2014). LncRNAs have been reported as key components of the ceRNA-mediated regulatory network, and aberrant miRNA-mediated lncRNA–mRNA ceRNA crosstalk has been implicated in many human complex diseases, including cancers. However, functional roles and regulatory mechanisms of lncRNA as ceRNAs in the development of osteoporosis and their potential implication for osteoporosis are largely unexplored. To explore the ceRNA activity of lncRNAs in the development of osteoporosis, we performed integrated analysis for paired expression profiles of dysregulated 68 mRNAs, 11 miRNAs, and 95 lncRNAs and miRNA–target regulatory relationship based on "ceRNA hypothesis," and constructed an osteoporosis-related miRNA-mediated dysregulated lncRNA–mRNA ceRNA network (DysCeNet) composed of 105 nodes and 515 edges (including eight miRNAs, 24 mRNAs, and 73 lncRNAs). Network analysis suggested that a large proportion of deregulated lncRNAs (76.8%, 73/95) in osteoporosis function as ceRNA and communicated with 24 mRNAs by competing for eight common miRNAs (**Figure 2A**), which implied that extensive variation in miRNAs and lncRNAs disrupted the miRNA-mediated lncRNA–mRNA ceRNA regulatory network contributing to osteoporosis at the post-transcriptional level. Functional analysis through functional enrichment analysis for mRNAs in the DysCeNet found that mRNAs as ceRNA counterparts of lncRNA biomarkers were involved in known osteoporosis or bone metabolism-related biological progression and pathways, including blood vessel development, NIK/NF-kappaB signaling, bone mineralization involved in bone maturation, osteoclast differentiation, and cell aging (**Figures 2B,C**). For example, blood vessels in the bone play vital roles for the formation of new bone and promoting blood vessel growth could reverse the weakening of bones and treat osteoporosis (Sivaraj and Adams, 2016). NIK/NF-kappaB signaling has been shown to play an important role in the positive and negative regulation of cytokine-mediated osteoclast formation and activation (Boyce et al., 2015). These results suggested that dysregulated expression of lncRNA ceRNAs and the resultant perturbation in miRNA-mediated lncRNA–mRNA crosstalk in the DysCeNet are involved in osteoporosis-biological processes and contributed to the osteoporosis.

A large number of studies have indicated the superior potential of lncRNAs as diagnostic and prognostic biomarkers compared with protein-coding genes due to the fact that lncRNAs were expressed in much more cell- type-, tissue-, and disease-specific patterns that are closely more associated with their function (Hauptman and Glavaè, 2013; Chen et al., 2018). lncRNA biomarkers have been widely investigated and identified in various cancers during the past years (Zhou et al., 2018a,b, 2019; Bao et al., 2019). To explore the potential application of lncRNAs with ceRNA activity as diagnostic biomarkers for osteoporosis risk, we performed random forest-based feature selection for 73 lncRNAs with ceRNA activity and identified 25 of 73 lncRNAs as potential diagnostic biomarkers. To accelerate the clinical application, we also developed a random forest-based classifier (RF-25lncRNA) composed of seven lncRNA biomarkers and test the performance of the RF-25lncRNA in different osteoporosis datasets. Performance evaluation of the LOOCV procedure showed that the RF-25lncRNA achieved a good performance in distinguishing high- and low-BMD subjects in three osteoporosis datasets. These results demonstrated that

seven lncRNA with ceRNA activity may become reliable and powerful epigenetic diagnostic biomarkers for early detection of osteoporosis risk.

There are some limitations to this study. First, the performance of newly identified lncRNA biomarkers was validated only in three osteoporosis datasets because of no other publicly available osteoporosis datasets. Second, the biological functions of newly identified lncRNA biomarkers are unknown, although they were found to have ceRNA activity in our study. Therefore, more laboratory and clinical researches were needed.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Three osteoporosis datasets were obtained from the Gene Expression Omnibus (GEO) database (GSE62589, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62589, GSE56814, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56814, and GSE13850, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13850).

AUTHOR CONTRIBUTIONS

YZ conceived and designed the experiments. MZ and LC analyzed the data. MZ wrote 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.2020.00184/full#supplementary-material

TABLE S1 | List of differentially expressed miRNAs, mRNAs and lncRNAs between high-BMD subjects and low-BMD subjects.

TABLE S2 | An osteoporosis-related dysregulated miRNA-mediated lncRNA-mRNA ceRNA network.

REFERENCES

Al Anouti, F., Taha, Z., Shamim, S., Khalaf, K., Al Kaabi, L., and Alsafar, H. J. (2019). An insight into the paradigms of osteoporosis: from genetics to biomechanics. *Bone Rep.* 11:100216. doi: 10.1016/j.bonr.2019.100216

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief Bioinform.* [Epub ahead of print]

Boyce, B. F., Xiu, Y., Li, J., Xing, L., and Yao, Z. J. E. (2015). NF-(B-mediated regulation of osteoclastogenesis. *Endocrinol. Metab.* 30, 35–44.

Bunch, H. (2018). Gene regulation of mammalian long non-coding RNA. *Mol. Genet. Genomics* 293, 1–15. doi: 10.1007/s00438-017-1370-9

Burge, R., Dawson-Hughes, B., Solomon, D. H., Wong, J. B., King, A., and Tosteson, A. (2007). Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025. *J. Bone. Miner. Res.* 22, 465–475. doi: 10.1359/jbmr.061113

Chen, L., Zhang, Y.-H., Pan, X., Liu, M., Wang, S., Huang, T., et al. (2018). Tissue expression difference between mRNAs and lncRNAs. *Eur. Rev. Med. Pharmacol. Sci.* 19:3416. doi: 10.3390/ijms19113416

Chen, X., Yang, L., Ge, D., Wang, W., Yin, Z., Yan, J., et al. (2019). Long non-coding RNA XIST promotes osteoporosis through inhibiting bone marrow mesenchymal stem cell differentiation. *Exp. Ther. Med.* 17, 803–811. doi: 10.3892/etm.2018.7033

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

Feng, J., Wang, J., and Li, C. H. (2019). LncRNA GAS5 overexpression alleviates the development of osteoporosis through promoting osteogenic differentiation of MSCs via targeting microRNA-498 to regulate RUNX2. *Eur. Rev. Med. Pharmacol. Sci.* 23, 7757–7765. doi: 10.26355/eurrev_201909_18985

Garnero, P. (2017). The utility of biomarkers in osteoporosis management. *Mol. Diagn. Ther.* 21, 401–418. doi: 10.1007/s40291-017-0272-1

Gong, R., Ren, S., Chen, M., Wang, Y., Zhang, G., Shi, L., et al. (2019). Bioinformatics analysis reveals the altered gene expression of patients with postmenopausal osteoporosis using liuweidihuang pills treatment. *Biomed. Res. Int.* 2019:1907906. doi: 10.1155/2019/1907906

Guo, X., Gao, L., Wang, Y., Chiu, D. K., Wang, T., and Deng, Y. (2016). Advances in long non-coding RNAs: identification, structure prediction and function annotation. *Brief Funct. Genomics* 15, 38–46. doi: 10.1093/bfgp/elv022

Hauptman, N., and Glavaè, D. (2013). Long non-coding RNA in cancer. *Noncoding RNA Res.* 14, 4655–4669.

Huang, R., Wu, J., Zheng, Z., Wang, G., Song, D., Yan, P., et al. (2019a). The construction and analysis of ceRNA network and Patterns of immune infiltration in Mesothelioma with bone metastasis. *Front. Bioeng. Biotechnol.* 7:257. doi: 10.3389/fbioe.2019.00257

Huang, R., Zeng, Z., Li, G., Song, D., Yan, P., Yin, H., et al. (2019b). The construction and comprehensive analysis of ceRNA networks and tumor-infiltrating immune cells in bone metastatic melanoma. *Front. Genet.* 10:828. doi: 10.3389/fgene.2019.00828

Jiang, H., and Wong, W. H. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 24, 2395–2396. doi: 10.1093/bioinformatics/btn429

Kornienko, A. E., Guenzl, P. M., Barlow, D. P., and Pauler, F. M. (2013). Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.* 11:59. doi: 10.1186/1741-7007-11-59

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

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

Liu, S., Huang, H., Chai, S., Wei, H., Huang, J., and Wan, L. (2019). Expression profile analysis of long non-coding RNA in skeletal muscle of osteoporosis by microarray and bioinformatics. *J. Biol. Eng.* 13:50. doi: 10.1186/s13036-019-0180-5

Lv, H., Zhang, Z.-M., Li, S.-H., Tan, J.-X., Chen, W., and Lin, H. (2019). Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief Bioinform.* [Epub ahead of print],

Maass, P. G., Luft, F. C., and Bähring, S. (2014). Long non-coding RNA in health and disease. *J. Mol. Med.* 92, 337–346.

Makitie, R. E., Hackl, M., Niinimaki, R., Kakko, S., Grillari, J., and Makitie, O. (2018). Altered MicroRNA profile in osteoporosis caused by impaired WNT signaling. *J. Clin. Endocrinol. Metab.* 103, 1985–1996. doi: 10.1210/jc.2017-02585

Mandourah, A. Y., Ranganath, L., Barraclough, R., Vinjamuri, S., Hof, R. V. T., Hamill, S., et al. (2018). Circulating microRNAs as potential diagnostic biomarkers for osteoporosis. *Sci. Rep.* 8, 1–10.

Mei, B., Wang, Y., Ye, W., Huang, H., Zhou, Q., Chen, Y., et al. (2019). LncRNA ZBTB40-IT1 modulated by osteoporosis GWAS risk SNPs suppresses osteogenesis. *Biol. Med.* 138, 151–166. doi: 10.1007/s00439-019-01969-y

Mercer, T. R., and Mattick, J. S. (2013). Structure and function of long non-coding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* 20, 300–307. doi: 10.1038/nsmb.2480

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G. D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* 5:e13984. doi: 10.1371/journal.pone.0013984

Parveen, B., Parveen, A., and Vohora, D. (2019). Biomarkers of osteoporosis: an update. *Endocr. Metab. Immune Disord. Drug Targets* 19, 895–912. doi: 10.2174/1871530319666190204165207

Pelechano, V., and Steinmetz, L. M. (2013). Gene regulation by antisense transcription. *Nat. Rev. Genet.* 14, 880–893. doi: 10.1038/nrg3594

Peng, S., Cao, L., He, S., Zhong, Y., Ma, H., Zhang, Y., et al. (2018). An overview of long non-coding RNAs involved in bone regeneration from mesenchymal stem cells. *Stem Cells Int.* 2018:8273648. doi: 10.1155/2018/8273648

Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17, 47–62. doi: 10.1038/nrg.2015.10

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

Seeliger, C., Karpinski, K., Haug, A. T., Vester, H., Schmitt, A., Bauer, J. S., et al. (2014). Five freely circulating miRNAs and bone tissue miRNAs are associated with osteoporotic fractures. *J. Bone Miner. Res.* 29, 1718–1728. doi: 10.1002/jbmr.2175

Silva, A. M., Moura, S. R., Teixeira, J. H., Barbosa, M. A., Santos, S. G., and Almeida, M. I. (2019). Long non-coding RNAs: a missing link in osteoporosis. *Bone Res.* 7:10. doi: 10.1038/s41413-019-0048-9

Sivaraj, K. K., and Adams, R. H. (2016). Blood vessel formation and function in bone. *Development* 143, 2706–2715. doi: 10.1242/dev.136861

Sun, J., Zhang, Z., Bao, S., Yan, C., Hou, P., Wu, N., et al. (2020). Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *J. Immunother. Cancer* 8:e000110. doi: 10.1136/jitc-2019-000110

Tan, J.-X., Li, S.-H., Zhang, Z.-M., Chen, C.-X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505, 344–352. doi: 10.1038/nature12986

Toth, R., Schiffmann, H., Hube-Magg, C., Buscheck, F., Hoflmayer, D., Weidemann, S., et al. (2019). Random forest-based modelling to detect biomarkers for prostate cancer progression. *Clin. Epigenetics* 11:148. doi: 10.1186/s13148-019-0736-8

Tu, K. N., Lie, J. D., Wan, C. K. V., Cameron, M., Austel, A. G., Nguyen, J. K., et al. (2018). Osteoporosis: a review of treatment options. *PT* 43, 92–104.

Ukon, Y., Makino, T., Kodama, J., Tsukazaki, H., Tateiwa, D., Yoshikawa, H., et al. (2019). Molecular-based treatment strategies for osteoporosis: a literature review. *Int. J. Mol. Sci.* 20:E2557. doi: 10.3390/ijms20102557

Yu, G., Wang, L., Li, Y., Ma, Z., and Li, Y. (2013). Identification of drug candidate for osteoporosis by computational bioinformatics analysis of gene expression profile. *Eur. J. Med. Res.* 18:5. doi: 10.1186/2047-783X-18-5

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

Zhou, M., Diao, Z., Yue, X., Chen, Y., Zhao, H., Cheng, L., et al. (2016). Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer. *Oncotarget* 7, 56383–56394. doi: 10.18632/oncotarget.10891

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018a). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018b). An immune-related Six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. *Mol. Neurobiol.* 55, 3684–3697. doi: 10.1007/s12035-017-0572-9

Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2019). Analysis of long non-coding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Brief Bioinform.* 20, 598–608. doi: 10.1093/bib/bby021

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 © 2020 Zhang, 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 02 April 2020
doi: 10.3389/fbioe.2020.00268

Identification of Pan-Cancer Prognostic Biomarkers Through Integration of Multi-Omics Data

Ning Zhao[1], Maozu Guo[1,2,3], Kuanquan Wang[1,4], Chunlong Zhang[5] and Xiaoyan Liu[4]*

[1] School of Life Sciences and Technology, Harbin Institute of Technology, Harbin, China, [2] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, [3] Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing, China, [4] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [5] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Prognostic biomarkers dedicating to treat cancer are very difficult to identify. Although high-throughput sequencing technology allows us to mine prognostic biomarkers much deeper by analyzing omics data, there is lack of effective methods to comprehensively utilize multi-omics data. In this work, we integrated multi-omics data [DNA methylation (DM), gene expression (GE), somatic copy number alternation, and microRNA expression (ME)] and proposed a method to rank genes by desiring a "Score." Applying the method, cancer-specific prognostic biomarkers for 13 cancers were obtained. The prognostic powers of the biomarkers were further assessed by C-indexes (ranged from 0.76 to 0.96). Moreover, by comparing the 13 survival-related gene lists, seven genes (*SLK*, *API5*, *BTBD2*, *PTAR1*, *VPS37A*, *EIF2B1*, and *ZRANB1*) were found to be associated with prognosis in a variety of cancers. In particular, *SLK* was more likely to be cancer-related due to its high missense mutation rate and associated with cell adhesion. Furthermore, after network analysis, *EPRS*, *HNRNPA2B1*, *BPTF*, *LRRK1*, and *PUM1* were demonstrated to have a broad correlation with cancers. In summary, our method has a better integration of multi-omics data that can be extended to the researches of other diseases. And the prognostic biomarkers had a better prognostic power than previous methods. Our results could provide a reference for translational medicine researchers and clinicians.

Keywords: multi-omics, pan-cancer, survival, biomarker, prognosis

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Richard R. Rodrigues,
Oregon State University, United States
Alfred Grant Schissler,
University of Nevada, Reno, United States

***Correspondence:**
Maozu Guo
guomaozu@bucea.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *02 January 2020*
Accepted: *13 March 2020*
Published: *02 April 2020*

Citation:
Zhao N, Guo M, Wang K, Zhang C and Liu X (2020) Identification of Pan-Cancer Prognostic Biomarkers Through Integration of Multi-Omics Data. Front. Bioeng. Biotechnol. 8:268. doi: 10.3389/fbioe.2020.00268

INTRODUCTION

Cancer is a major public health problem worldwide (Siegel et al., 2020) and the occurrence of cancer is caused by many factors. It is not only controlled by genetics and epigenetics, but also influenced by many other regulatory factors, such as miRNAs. A variety of regulatory factors contribute to the heterogeneity of cancer (Marusyk et al., 2012; Swanton, 2012; Burrell et al., 2013), which leads to a low cure rate and poor prognosis. Survival prediction provided a crucial evidence for the process of cancer diagnosis and treatment. Prognostic biomarkers are used to predict likelihood of recurrence or progression in patients with cancer (Cagney et al., 2018). However, it is still hard to identify the prognostic biomarkers of cancer accurately.

Omics data play a key role in predicting prognostic biomarkers. At present, many researchers have identified prognostic biomarkers based on differential analysis of DNA methylation (DM) or other omics data, involving gene expression (GE), somatic copy-number alteration (SCNA) and microRNA expression (ME). Dalerba et al. (2016) found that *CDX2* was a prognostic biomarker in stage II and stage III colon cancer by analyzing GE data. Zhao et al. (2017) identified eight differentially methylated CpGs as new prognostic biomarkers for prostate cancer by analyzing DM data. Morikawa et al. (2018) discovered that SCNAs in 8p11.21-22, 12p13.31, 20q13.2, 3q26.1, 4q13.2, and 22q11.23 were critical for the development and survival of ovarian clear cell carcinoma. Lindahl et al. (2018) developed a prognostic 3-miRNA classifier (miR-106b-5p, miR-148a-3p, and miR-338-3p) in early-stage mycosis fungoides. The advantage of omics data for identifying cancer-related prognostic biomarkers can be clearly seen in the studies mentioned above. However, each of these studies used only one type of omics data, which did not make full use of omics data.

The regulation of GE is a complex process. Generally, the DNA hypermethylation in promoter region of genes could cause transcriptional silencing (Baylin, 2005) and DNA hypomethylation was associated with the activation of GE (Berdasco and Esteller, 2010). Besides, the copy number correlated positively with expression levels for genes (Fehrmann et al., 2015). Moreover, miRNAs complementary bound to messenger RNAs (mRNAs) and formed RNA-induced silencing complex (RISC) to downregulate GE levels (Bartel, 2004). The researches of cancer focusing on one-dimensional omics data may only provide limited information for the etiology of oncogenesis and tumor progression. In the past few years, more and more researches applied multi-omics data. Xu et al. (2019) proposed a method, named high-order path elucidated similarity (HOPES), to identify cancer subtypes by simultaneous interrogation multi-omics data. They utilized their method on GE, DM, and ME data of five TCGA cancers to identify subtypes and further validated reliability and clinical role of them. Vasaikar et al. (2018) developed a powerful database, named LinkedOmics, for analysis of omics data in cancer. LinkedOmics contained multi-omics data of 32 cancer types and allowed for flexible exploration and comparison of associations between multiple types of attributes within and across tumor types. The positive results of these researches confirmed the feasibility of integrating multi-omics data. Both of these work used multi-omics data for cancer research. However, they did not focus on prognostic markers, so we cannot further compare them numerically with our method.

Similarly, integrating omics data indicated the potential benefits for discovering underlying prognostic markers in cancer (Huang et al., 2017). Using multi-omics data acquiring from the same set of samples has the potential capacity to expose more accurate biomarkers for patients' survival than examining by one single-omics data (Rappoport and Shamir, 2018). Yuan et al. (2014) used somatic copy-number alteration (SCNA), DM, GE, ME, and protein expression data to predict survival status of patients. They found that incorporating molecular data with clinical variables improved the accuracy of survival prediction for cancers. This work provided a starting point and resources for the subsequent researches. Zhang et al. (2016) utilized GE, SCNA, ME, and DM data to uncover protein–protein subnetworks associated with prognosis. This work built a multi-dimensional subnetwork atlas for cancer prognosis to investigate the potential impact of multiple genetics and epigenetics better. Chaudhary et al. (2018) presented a deep learning based model on liver hepatocellular carcinoma (LIHC) that robustly differentiates survival subpopulations of patients using GE, DM, and ME data. They validated this multi-omics model on five external datasets of various omics types and all have good performance. Zhu et al. (2017) presented a kernel machine learning method to systematically quantify the prognostic values of clinical information, GE, SCNA, DM, and ME across 14 cancer types. This study aimed to compare the advantages and disadvantages of using different omics data to evaluate patients' survival. Based on their result, GE and ME data were demonstrated to be the best data for the prognosis of cancers. Mishra et al. (2019) used DM, GE, ME, and long non-coding RNA (lncRNA) expression data to identify potential prognostic markers of pancreatic ductal adenocarcinoma. They identified several genes, miRNA, lncRNA, and CpG sites as probable prognostic biomarkers. All methods mentioned above used multi-omics data to perform prediction of patients' survival. However, most of them did not integrate multi-omics data comprehensively but only utilized multi-omics data to explore mechanism of cancer separately. Moreover most of them they did not provide specifically prognostic biomarkers for other clinical researches or just aimed at limited kinds of cancers.

The Cancer Genome Atlas (TCGA) provides multiple omics data for different cancers (Cancer Genome Atlas Research Network, 2011, 2012, 2013; Cancer Genome Atlas Research Network et al., 2016), which allows for analyzing multi-omics data coming from the same samples. So far, there already exist a variety of methods for predicting patients' survival status using TCGA omics data.

In this work, we put forward our own method to identify prognostic biomarkers and identified prognostic gene lists for 13 types of cancers. This work provided theoretical foundation and reliably prognostic biomarkers for other researches focusing on diagnosis, prognosis, and treatment of cancers.

MATERIALS AND METHODS

Data

Multi-omics data were downloaded from TCGA. The scale and platform of each cancer data are shown in **Table 1**. We selected the cancers which had HM450K DM data, RNA-seq data (GE), miRNA-seq data (ME), and SNP 6.0 copy number data (SCNA) simultaneously and whose sample size was greater than 200. Samples with sample type codes of "01" were retained, which represented "Primary Solid Tumor." After being filtered, there were 13 types of cancers available. For SCNA data, a matrix was obtained after being processed by Gistic 2.0 (Mermel et al., 2011). Next, all omics data matrixes except ME were converted into gene matrixes based on the annotation information from TCGA. Genes with missing values in $> 5\%$ of the samples were removed

TABLE 1 | The sample size of 13 types of cancers.

Cancer	Clinical	DM (450K)	SCNA (nocnv)	GE (FPKM-UQ)	ME (isoform)	Total size
Bladder urothelial carcinoma [BLCA]	291	412	410	408	409	283
Breast invasive carcinoma [BRCA]	735	783	1092	1091	1078	497
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	177	307	295	304	307	166
Colon adenocarcinoma [COAD]	186	296	452	456	444	164
Head and neck squamous cell carcinoma [HNSC]	416	528	522	500	523	380
Kidney renal clear cell carcinoma [KIRC]	452	319	532	530	516	251
Kidney renal papillary cell carcinoma [KIRP]	183	275	289	288	291	172
Brain lower grade glioma [LGG]	347	516	515	511	512	340
Liver hepatocellular carcinoma [LIHC]	259	377	375	371	372	250
Lung adenocarcinoma [LUAD]	292	460	520	515	515	229
Lung squamous cell carcinoma [LUSC]	299	370	503	501	478	190
Sarcoma [SARC]	204	261	260	259	259	200
Stomach adenocarcinoma [STAD]	201	395	441	375	436	157

The sample size of clinical, DNA methylation (DM), somatic copy-number alteration (SCNA), gene expression (GE), and microRNA expression(ME) of 13 types of cancers. The platforms for each data were written in the parentheses below. Abbreviations for cancer names were written in square brackets. The total size of each cancer was the number of samples which have all the five types of data simultaneously.

in each matrix. Moreover, for GE and ME data, we retained the genes or miRNAs with values greater than 0 in > 50% of the samples and with values greater than 1 in > 10% of the samples, respectively. After converting if one gene had multiple signals in one sample, we calculated the average of the values as the final signal. For ME, miRNAs were specifically bound to mRNAs by complementary base pairing, therefore the corresponding relationships between miRNAs and genes were obtained through the miRNA–mRNA interactions which were downloaded from the Starbase database (Yang et al., 2011). Interactions with no less than five supporting experiments and anti-correlation in no less than one cancer type were selected. Since multiple miRNAs were bound to the same gene, the average value of the miRNAs was assigned to the gene.

Because of different scales for the omics data, the data were normalized based on the following rules. First, each omics data were organized into a matrix of the same genes and samples, separately. Second, the method z-score was used to transform a matrix into standardized one with the mean and standard deviation of 0 and 1, respectively. Finally, we uniformly kept the fourth decimal place for better integration of the standardized data.

Screening of Candidate Survival-Related Genes

Univariate Cox proportional hazards regression model (Cox, 1986) was used to identify candidate survival-related genes from each omics data through the formula:

$$\mathrm{h(X, t)} = \mathrm{h_0(t)exp(\beta X)} \quad (1)$$

where the explanatory variable X was the omics data (DM, GE, copy number variation, or miRNA expression) of a gene, and the response variable t was the survival time (Aalen, 1989). The proportional hazards regression model was calculated through the R package "survival." β greater than zero meant the gene was a risk factor base on the corresponding omics data. Then using voting strategy, if a gene had a *p*-value of likelihood ratio test less than 0.05 (Yuan et al., 2014), the gene was denoted as "1". Otherwise, it was denoted as "0." Finally, a gene defined as a candidate survival-related gene should be marked as "1" in no less than two of the four omics data types (**Figure 1A**).

Identification of Prognostic Biomarkers

As shown in **Figure 1B**, prognostic biomarkers were further identified in the set of candidate survival-related genes. For each gene, a matrix $\mathrm{M} = [Omics_{GE}, Omics_{SCNA}, Omics_{DM}, Omics_{ME}]$ merged by the vectors of the four omics data of the gene was obtained. Then, the multivariate Cox proportional hazards model was applied on it. Briefly, the model assumed that a patient with covariate values has a cumulative hazard rate related to an unspecified baseline hazard rate seen in the equation:

$$\mathrm{h\,(t, M)} = h_0(\mathrm{t})\exp(\beta_1 Omics_{GE} + \beta_2 Omics_{SCNA} + \beta_3 Omics_{DM} + \beta_4 Omics_{ME}) \quad (2)$$

where h(t, M) was the patient's hazard of death at time t, $h_0(t)$ was the baseline hazard rate, and $B = [\beta_1, \beta_2, \beta_3, \beta_4]$ was a regression coefficient that gives the effect of each M covariate on the hazard rate (Alamartine et al., 1991). Each β could be interpreted as a risk coefficient (Collett, 2015). If the *p*-values of Cox fitting in all three overall tests (likelihood, Wald, and log-rank) were less than 0.05, the model was thought to be significant (Rodriguez-Martin et al., 2020). Therefore, we only kept genes whose all three *p*-values were less than 0.05.

For the retained genes, each gene had a vector including the value of four types of omics data in each sample $V = [v_1, v_2, v_3, v_4]$. The risk score (RS) for the gene in each sample was then calculated:

$$RS = B \cdot V \quad (3)$$

The RS score could be used to predict the patients' risk.

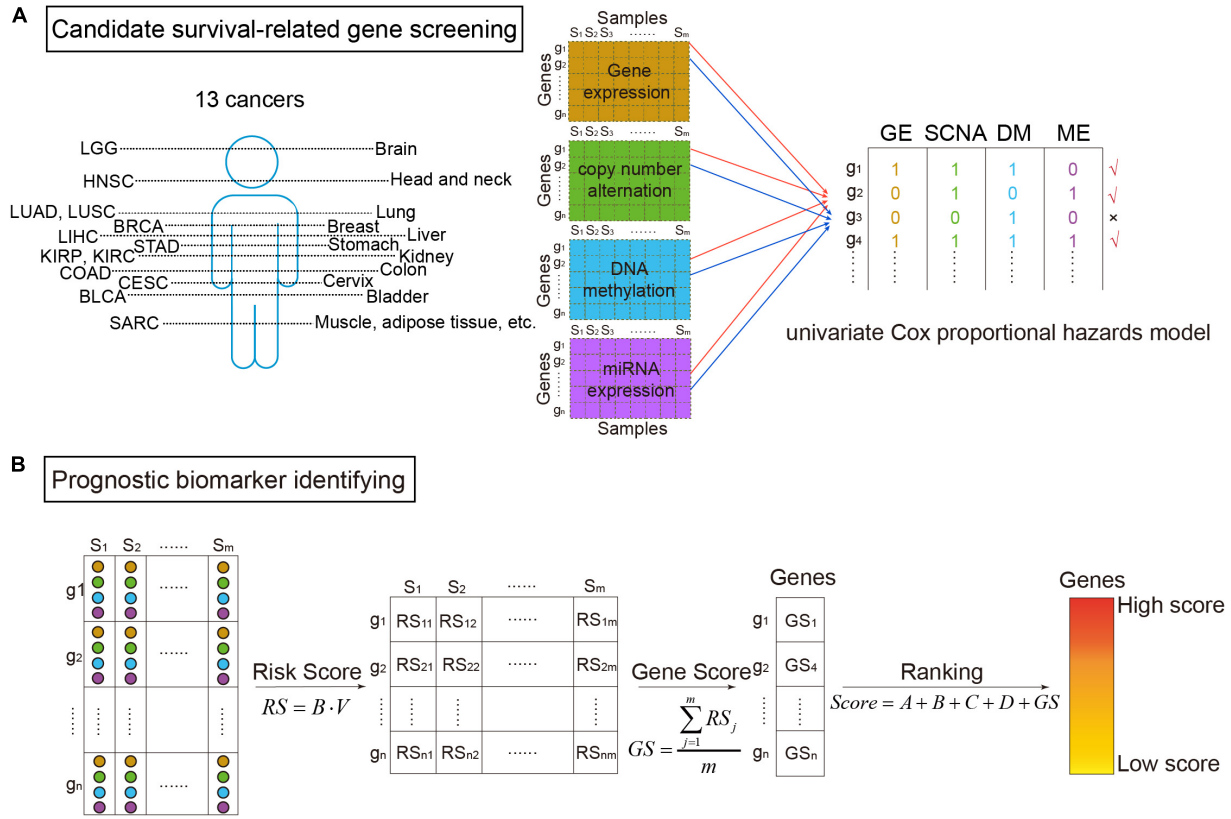


FIGURE 1 | The workflow of survival-related genes identification. **(A)** Candidate survival-related gene screening. DNA methylation, gene expression, somatic copy-number alteration, and microRNA (miRNA) expression profiles of TCGA for the same samples were extracted. miRNA expression data were corresponding to genes according to miRNA–mRNA interactions. Then, we got four types of data in the same samples and the same genes. On each omics data, univariate Cox proportional hazards model was utilized to identify survival-related genes. Only the genes associated with survival in more than two types of data were considered to be candidate genes. **(B)** Prognostic biomarker identifying. For the selected candidate genes, the multivariate Cox proportional hazards model was then applied to get risk scores (RS). Further, scores for ranking genes were obtained by calculating GS scores. In which, A, B, C, and D were binary variables indicating whether the gene was survival-related at the four omics data or not ("1" for related and "0" for not), respectively. The high ranked genes were identified survival-related.

Thereafter, RS scores of the genes were used to calculate each gene's score (GS):

$$GS = \frac{\sum_{j=1}^{m} RS_j}{m} \tag{4}$$

where m was the number of samples. At last, the scores of univariate and multivariate Cox proportional hazards model were combined to calculate the survival-related score of each gene (Score):

$$Score = A + B + C + D + GS \tag{5}$$

where A, B, C, and D represented whether the gene was survival-related at the GE level, copy number level, DM level, and miRNA level, respectively ("1" meant related and "0" meant not). The higher the score, the more relevant between the gene and patients' survival. Therefore, high score genes were identified as prognostic biomarkers.

Functional Analysis

Cumulative hypergeometric inspection was applied to enrichment analysis of Gene Ontology (GO) functions (Gene Ontology, 2015) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto, 2000):

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-1}}{\binom{N}{n}} \tag{6}$$

where N was the whole number of genes. M was the number of genes on a term or a pathway. n was the intersection of interested gene set and N. i was the intersection of M and n. Significant threshold of hypergeometric test was set to $P < 0.05$ (Liu et al., 2018). For the enrichment analysis of GO, we only investigated the biological process (BP) terms.

Different Expression

In order to identify differentially expressed genes, the corresponding normal samples of LUSC and KIRC were downloaded from TCGA. There were 49 LUSC normal samples and 72 KIRC normal samples. After organized into the same gene set, the differentially expressed genes between tumor and

normal samples were identified using the R package "samr" with the threshold value q-value < 0.05 and $|log_2(fold\ change)| > 1$ (Group et al., 2020). The package was based on significance analysis of microarrays (SAM). SAM was developed based on t-test and adjusted the p-value to assess the statistically significant changes for genes (Tusher et al., 2001).

Predictive Model Validation

For each cancer type, in order to evaluate the prognostic power of the biomarker fairly and accurately, the concordance index (C-index) (Harrell et al., 1996) was applied to assess the prognostic power of the classifier. The C-index was a non-parametric measure to quantify the discriminatory power of a predictive model with the value ranging from 0.5 to 1. A C-index of 1 represented perfect prediction accuracy, while C-index of 0.5 indicated a bad prediction like a random guess.

First, we randomly selected 90% of the samples. Second, the Cox regression model was used to calculate the RS score for each sample by multi-omics data of the identified biomarker genes. Based on the RS score, samples were classified into high and low risk groups. Patients in the high risk group were more likely to have poor prognosis while patients in the low risk group were more likely to have good prognosis. Finally, the predicted outcomes for patients were compared with the real status to calculate the C-indexes.

The procedure above was repeated 100 times to generate 100 C-indexes. If the median value of C-index was significantly higher than 0.5, indicating that the model had substantially prognostic power.

Decision Curve Analysis

Decision curve analysis was performed through the multi-omics data and every single omics data, respectively. The method was based on the principle that the relative harms of false positives (e.g., unnecessary biopsy) and false negatives (e.g., missed cancer) could be expressed in terms of a probability threshold (Vickers et al., 2008). Therefore, this threshold probability could be used to determine both whether a patient was defined as test-positive

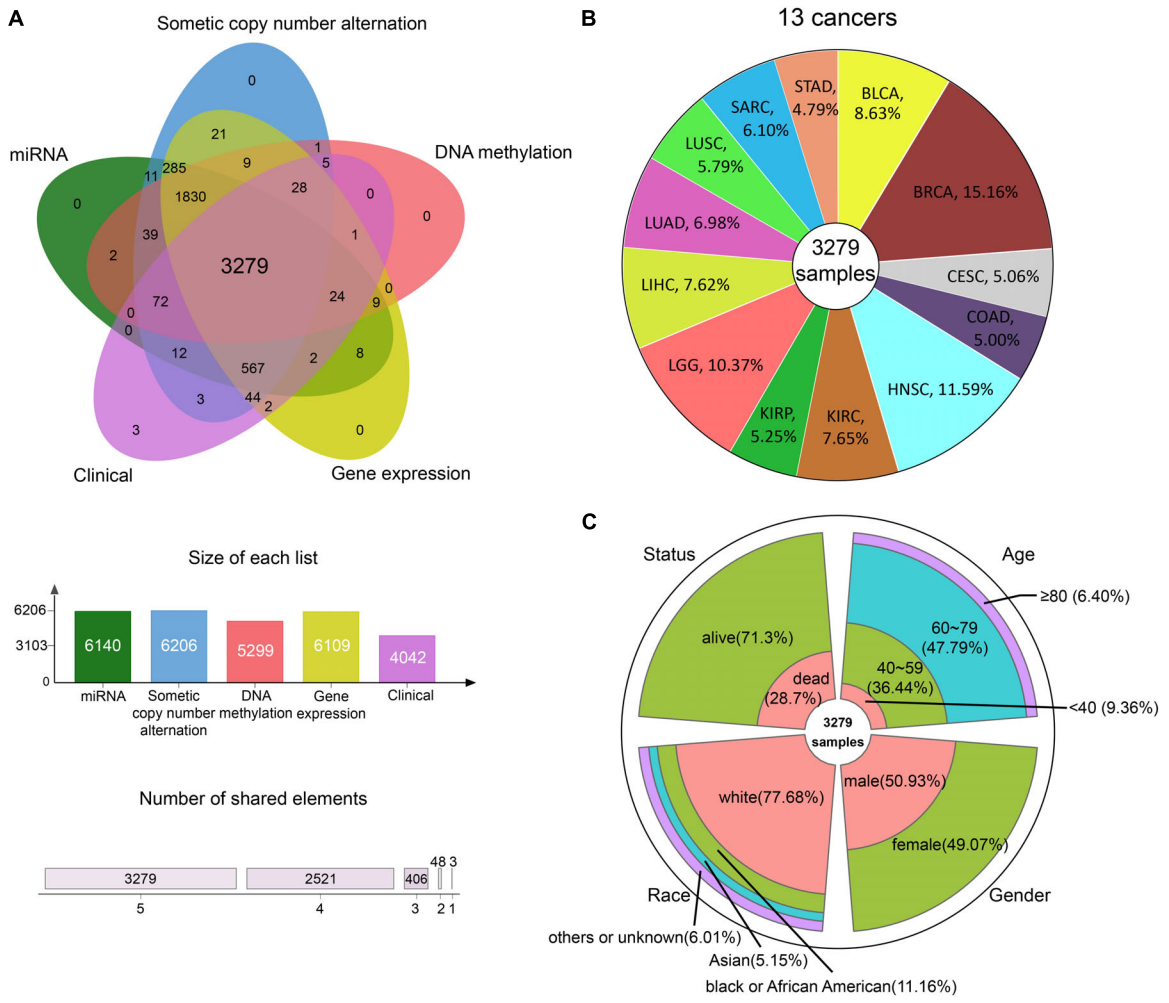


FIGURE 2 | The information of pan-cancer samples. **(A)** The sample set intersections of the multi-omics data. Only the intersecting samples were chosen. We selected 3279 samples in this study. **(B)** The proportion of each cancer. **(C)** The clinical features distribution of the 3279 samples.

TABLE 2 | The prognostic biomarkers of each cancer.

Cancers	Genes
BLCA	*NCBP1*, *RP9*, *UBP1*, *AURKB*, *ARL6IP5*, *LEMD3*, *HSPBP1*, *TMEM214*, *MCMBP*, *FAM107B*
BRCA	*LIMS1*, *NDUFA3*, *NKTR*, *SRP68*, *ARPC3*, *TMEM138*, *DDIT4*, *OCIAD1*, *MAF1*, *DPY19L4*
CESC	*HNRNPA2B1*, *MYO9B*, *EIF3B*, *MTX2*, *MON1B*, *SUN1*, *SSH1*, *SLC35E3*, *MAP7D1*, *PGAM5*
COAD	*SRP72*, *TAF10*, *USP1*, *USP8*, *JKAMP*, *YTHDF1*, *BRIX1*, *ATG101*, *VPS37A*, *TMED4*
HNSC	*DAGLA*, *DNAJA3*, *PTER*, *NUDT5*, *FAM168A*, *GTPBP4*, *PTCD3*, *MINDY3*, *NCALD*, *STEAP2*
KIRC	*BPTF*, *IMPA1*, *RPS6KA4*, *TSFM*, *CEPT1*, *PRKD3*, *PPME1*, *SPIRE1*, *NUDCD1*, *NHLRC2*
KIRP	*DAPK3*, *EPRS*, *IARS*, *SYMPK*, *PCSK7*, *APBB3*, *MSANTD2*, *TBC1D17*, *DOHH*, *CMSS1*
LGG	*FDXR*, *VPS4B*, *WASHC5*, *CITED2*, *BRD8*, *MON2*, *TSPAN13*, *MIOS*, *OGFOD3*, *PIGO*
LIHC	*BCL2L1*, *SELENOW*, *HIST1H2BN*, *MMADHC*, *PNPO*, *ZDHHC11*, *ULBP2*, *CSRNP2*, *SPC24*, *RPL7L1*
LUAD	*FGFR3*, *LTBP3*, *SLC6A4*, *PUM1*, *ARHGAP44*, *SLC39A1*, *NAGPA*, *BTBD2*, *LRRK1*, *ZFC3H1*
LUSC	*HSF2*, *BCLAF1*, *UHRF1BP1L*, *CHORDC1*, *CREBZF*, *FBXO30*, *PCGF6*, *PLCD3*, *HINT3*, *SLC35E2B*
SARC	*BMP1*, *NCAM2*, *PBX1*, *RAD17*, *ARHGEF10*, *PSD3*, *MRPL17*, *FAM160B2*, *CHMP7*, *VPS37A*
STAD	*ANK3*, *GNAI2*, *MARCKS*, *NEDD4*, *PRKAA1*, *UGP2*, *TAF1C*, *INO80D*, *USP37*, *FAM126B*

or negative and to model the clinical consequences of true and false positives using a clinical net benefit function:

$$net\ benefit = \frac{True\ positives}{n} - \frac{False\ positives}{n}\left(\frac{p_t}{1-p_t}\right) \quad (7)$$

where n was the total number of patients in the study and p_t was the threshold probability. Net benefit was weighted by the relative harm of forgoing treatment compared with the negative consequences of an unnecessary treatment. In the decision curve, the thin oblique line represented the assumption that all patients have been treated. The black line represented the assumption that no patients have been treated.

RESULTS

Pan-Cancer Prognostic Biomarker Identification

We integrated GE, SCNA, DM, and miRNA expression data of 13 cancers from TCGA: bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), brain lower grade glioma (LGG), LIHC, lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), sarcoma (SARC), and stomach adenocarcinoma (STAD). After data preprocessing, samples with all the four omics data were kept. Whereupon, we collected the DM, GE, copy number, and miRNA expression of 3279 samples (**Figure 2A**). The percentage of each cancer is shown in **Figure 2B**. We then summarized the clinical characteristics of the 3279 samples. As shown in **Figure 2C**, the majority of these patients were 60–79 years old. And the number of men and women was basically equal. Hence, the sample set could be used to study cancer without gender and age bias. In addition, most of the patients were white people. The

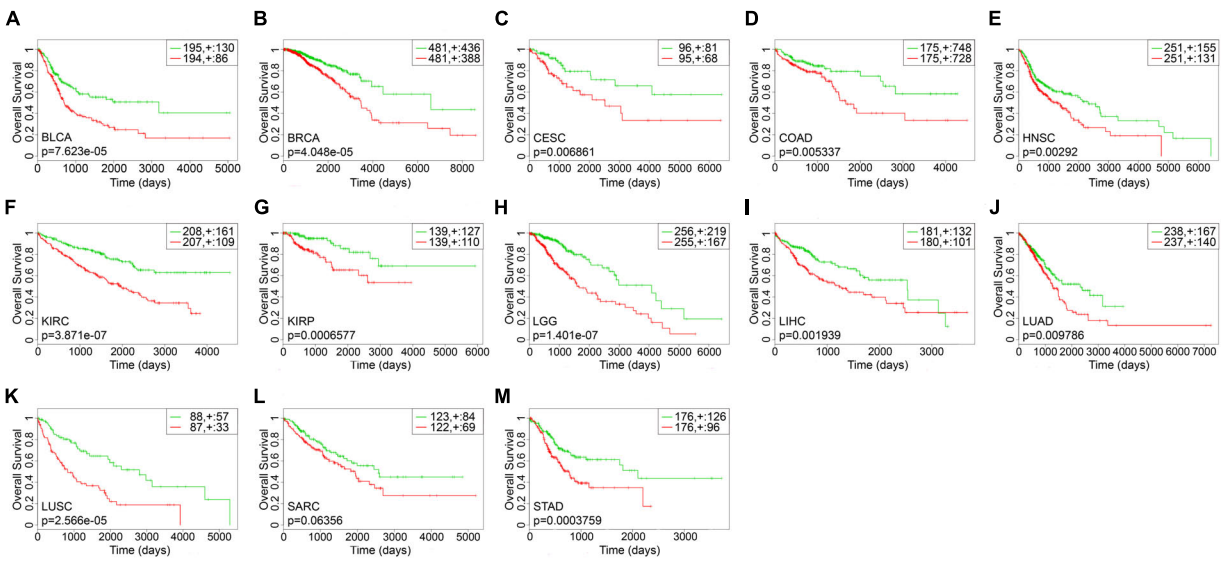


FIGURE 3 | The Kaplan–Meier curves of top-10 survival-related genes for each cancer. The green lines represented the low risk groups and the red lines represented the high risk groups. " + " indicated the censored follow-ups. **(A)** BLCA. **(B)** BRCA. **(C)** CESC. **(D)** COAD. **(E)** HNSC. **(F)** KIRC. **(G)** KIRP. **(H)** LGG. **(I)** LIHC. **(J)** LUAD. **(K)** LUSC. **(L)** SARC. **(M)** STAD.

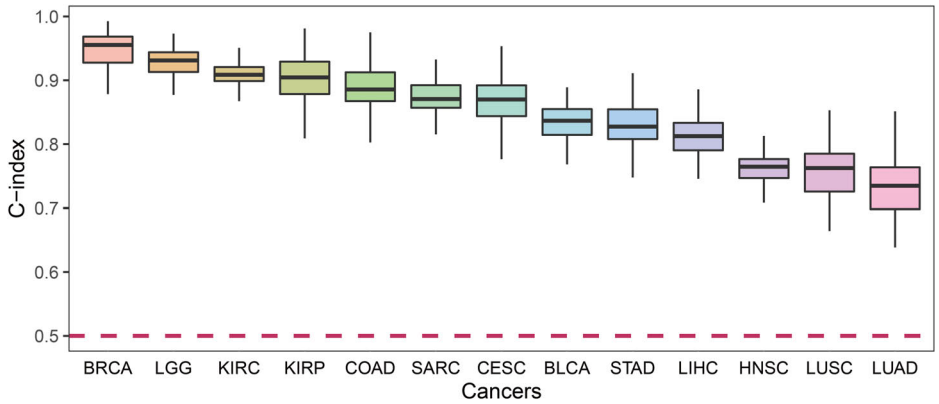


FIGURE 4 | The C-index comparison of the prognostic power of our prognostic biomarkers in 13 cancers.

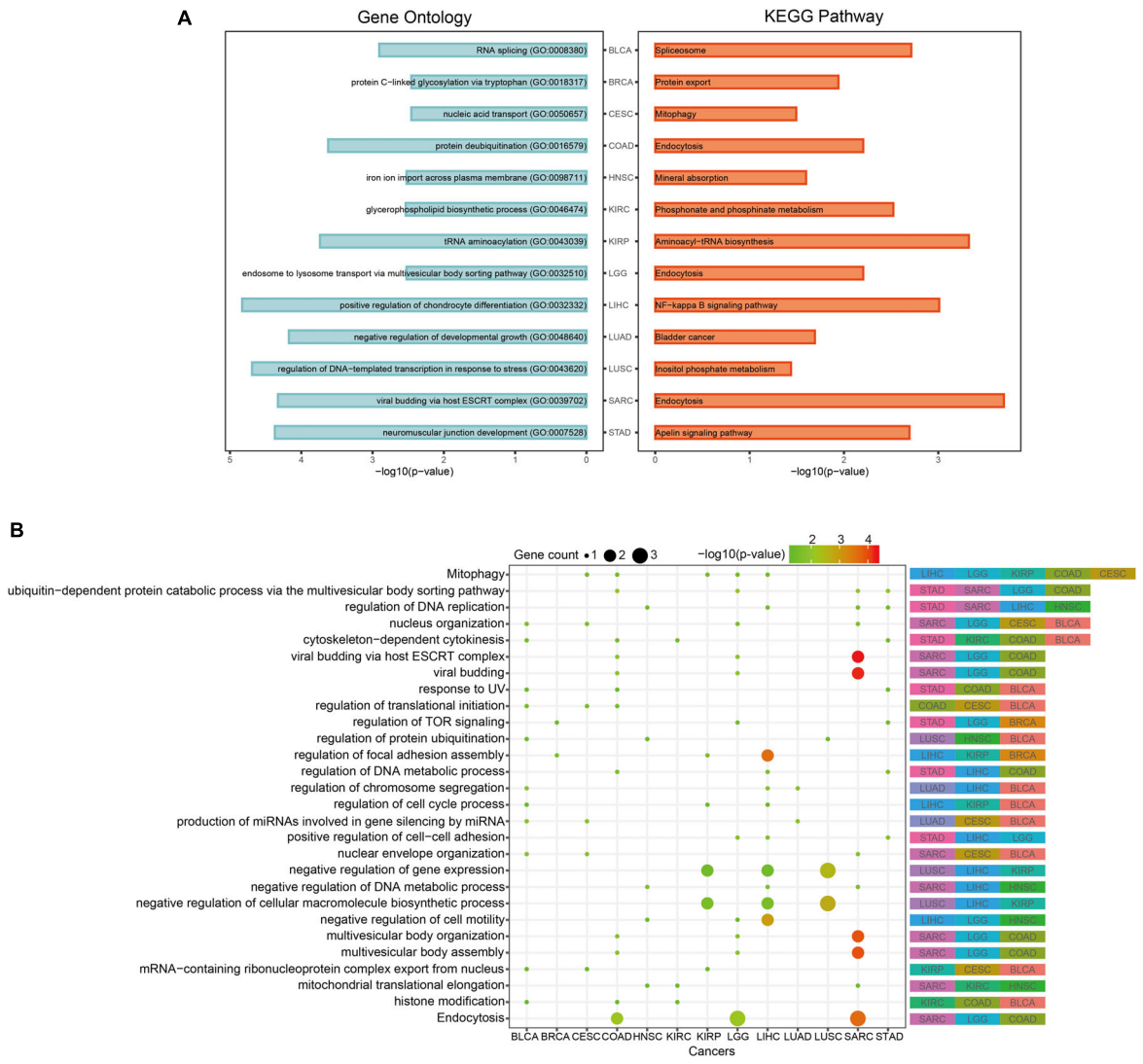
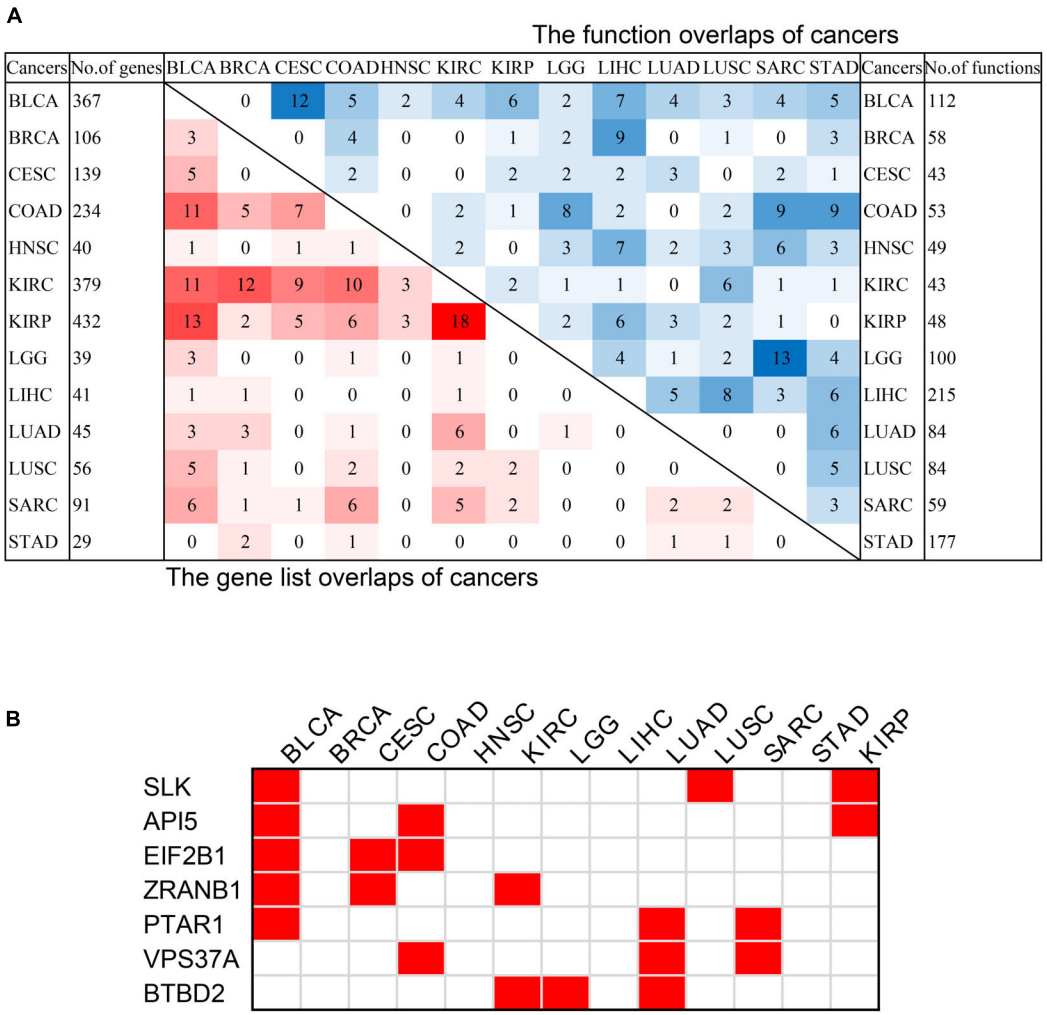


FIGURE 5 | Pan-cancer functional comparison of survival-related genes. **(A)** The representative KEGG pathways and GO functions enriched by the top-10 prognostic genes of each cancer. **(B)** The distribution of cancers enriched to each function. The size of the dots represented the number of enriched genes. The color of the dots represented the *p*-values.



Cancers	No.of genes	BLCA	BRCA	CESC	COAD	HNSC	KIRC	KIRP	LGG	LIHC	LUAD	LUSC	SARC	STAD	Cancers	No.of functions
BLCA	367		0	12	5	2	4	6	2	7	4	3	4	5	BLCA	112
BRCA	106	3		0	4	0	0	1	2	9	0	1	0	3	BRCA	58
CESC	139	5	0		2	0	0	2	2	2	3	0	2	1	CESC	43
COAD	234	11	5	7		0	2	1	8	2	0	2	9	9	COAD	53
HNSC	40	1	0	1	1		2	0	3	7	2	3	6	3	HNSC	49
KIRC	379	11	12	9	10	3		2	1	1	0	6	1	1	KIRC	43
KIRP	432	13	2	5	6	3	18		2	6	3	2	1	0	KIRP	48
LGG	39	3	0	0	1	0	1	0		4	1	2	13	4	LGG	100
LIHC	41	1	1	0	0	0	1	0	0		5	8	3	6	LIHC	215
LUAD	45	3	3	0	1	0	6	0	1	0		0	0	6	LUAD	84
LUSC	56	5	1	0	2	0	2	2	0	0	0		0	5	LUSC	84
SARC	91	6	1	1	6	0	5	2	0	0	2	2		3	SARC	59
STAD	29	0	2	0	1	0	0	0	0	0	1	1	0		STAD	177

FIGURE 6 | The intersection of pan-cancer genes. **(A)** The intersections of the lists of survival-related genes (left bottom) and the intersection of associated functions (top right corner) of cancers. The total numbers of genes associated with survival of each cancer were on the left, and the total associated functions were on the right. The color blocks represented the number of intersecting samples of each two cancers. The darker the color, the greater the intersection was. **(B)** The pan-cancer survival-related genes. The red blocks indicated that the gene was survival-associated with the cancer.

complete clinical information for each sample is provided in **Supplementary Table S1**.

The survival-related gene list of each cancer is shown in **Supplementary Table S2**. We took top-10 genes as a prognostic biomarker of each cancer (**Table 2**) to draw Kaplan–Meier (KM) curves and calculated their log-rank *p*-values. As shown in **Figure 3**, the prognostic markers for each cancer significantly distinguished the high and low risk groups, except for SARC.

For each cancer type, we calculated the C-index which was a non-parametric measure to quantify the discriminatory power of a predictive model. **Figure 4** shows the C-indexes of each cancer. All of the cancers had a C-index significantly higher than 0.5. BRCA had the highest C-index (0.96) while LUSC had the lowest (0.76).

In order to discover the relationship among different cancers based on function, we used the prognostic biomarker genes to perform functional enrichment analysis of GO and KEGG (**Supplementary Table S3**). The most significantly enriched functions and pathways of each cancer are displayed in **Figure 5A**. Among them, COAD, LGG, and SARC were enriched in "endocytosis." BLCA was enriched in "RNA splicing" and CESC was enriched in "mitophagy." The prognostic biomarker genes were enriched in closely cancer-related functions. Then, we calculated the counts of each function enriched by cancers. As shown in **Figure 5B**, "Mitophagy" was enriched by the most cancers. Mitophagy was a tumor suppression mechanism (Bernardini et al., 2017). Besides, we had some interesting findings. First, the most significantly enriched functions of each cancer were their specific functions, while the common functions of cancers were not highly significant generally. For example, "cytoskeleton-dependent cytokinesis" was the common enriched function of STAD, KIRC, COAD, and BLCA, and they had *p*-values about 0.03 which was less significant than their most significantly enriched functions (p-values < 0.003). And

TABLE 3 | The seven pan-cancer survival-related genes.

Genes	Functions	PubMed IDs
SLK	Cell adhesion, regulation of cell cycle	26676752, 27849608, 22057237, 27247392, 22699621
ZRANB1	Protein catabolic process	26676752, 27790711, 25559195, 27006499, 24316982, 27178121, 23685747, 22421440
BTBD2	Protein catabolic process	26676752, 22057237
PTAR1	Cellular protein modification process	26676752
VPS37A	Viral process	27849608, 24316982, 23045694
EIF2B1	Glial cell development	22057237
API5	Regulation of cell death	27790711, 27178121

The most relevant functions of each gene were listed. The references supporting the association of gene with cancer were given in the form of PubMed IDs.

their most significant functions were all their specific functions. Second, even if different cancers were enriched in a same function, the enrichment of function in different cancers was caused by different gene sets. For instance, "Mitophagy" was the common function of LIHC, LGG, KIRP, COAD, and CESC, but the function was hit by different genes (*BCL2L1* of LIHC, *CITED2* of LGG, *TBC1D17* of KIRP, *USP8* of COAD, and *PGAM5* of CESC). Whereafter, we sought the intersection of associated functions for the 13 cancers (**Figure 6A** top right corner). The result showed that the intersection of LGG and SARC was the largest, followed by BLCA and CESC.

In order to discover the relationship among different cancers based on survival-related genes, we first compared the intersection of the survival-related gene lists between each two cancers. We found there was always an overlap between each two gene lists (**Figure 6A** left bottom). The intersection of KIRP and KIRC was the largest. All of the intersections among KIRC, KIRP, and BLCA were large, which might be due to the reason that the three cancers had the largest number of genes. The intersections with other cancers were roughly proportional to the size of the gene list. Second, we compared the gene lists among the 13 types of cancers. We found that seven genes were associated with survival in three kinds of cancers (**Figure 6B**). Subsequently, we downloaded the list of cancer-related genes from the Candidate Cancer Gene Database (CCGD) (Abbott et al., 2015), and retained the human genes that appeared in at least one of the COSMIC and CGC (Sondka et al., 2018). A total of 9265 genes were retained. All of the seven pan-cancer survival-related genes were in the list, and have been verified cancer-related in no less than one literature (**Table 3**). In addition, we investigated the functions of the seven genes (**Supplementary Table S4**) and conducted UCSC Genome Browser (Tyner et al., 2017) analysis on the seven genes. We found that *SLK* had an unconservative exon region, which containing four missense variants (**Figure 7A**).

In order to further verify the close relationship of *SLK* and cancer, we checked the mutation of *SLK* in the COSMIC database (Tate et al., 2019), and found that missense substitution occurred in 36.93% of the samples (**Figure 7B**). Next, enrichment analysis in GO terms (BP) was performed by Enrichr (Kuleshov et al., 2016) and found *SLK* was mainly associated with cell adhesion (**Figure 7C**). Finally, we used STRING database (Szklarczyk et al., 2019) to check the interacted proteins of *SLK*. **Figure 7D** shows there were 10 genes interacting with *SLK*. Half of the interactions have been demonstrated through more than one method, and the genes interacting with *SLK* also had strong relationship between each other.

In order to explore the correlation among prognostic biomarkers of different cancers, we checked the genomic locations of these 130 genes (**Figure 8A**). There were many prognosis-related genes located in chr 6, chr 7, chr 8, chr 11, chr 12, and chr 17, while few genes in chr 13, chr 14, chr 18, chr 20, and chr 21. In addition, we constructed a protein–protein interaction network of these genes based on the STRING database (**Figure 8B**). As shown in the network, the prognosis-related

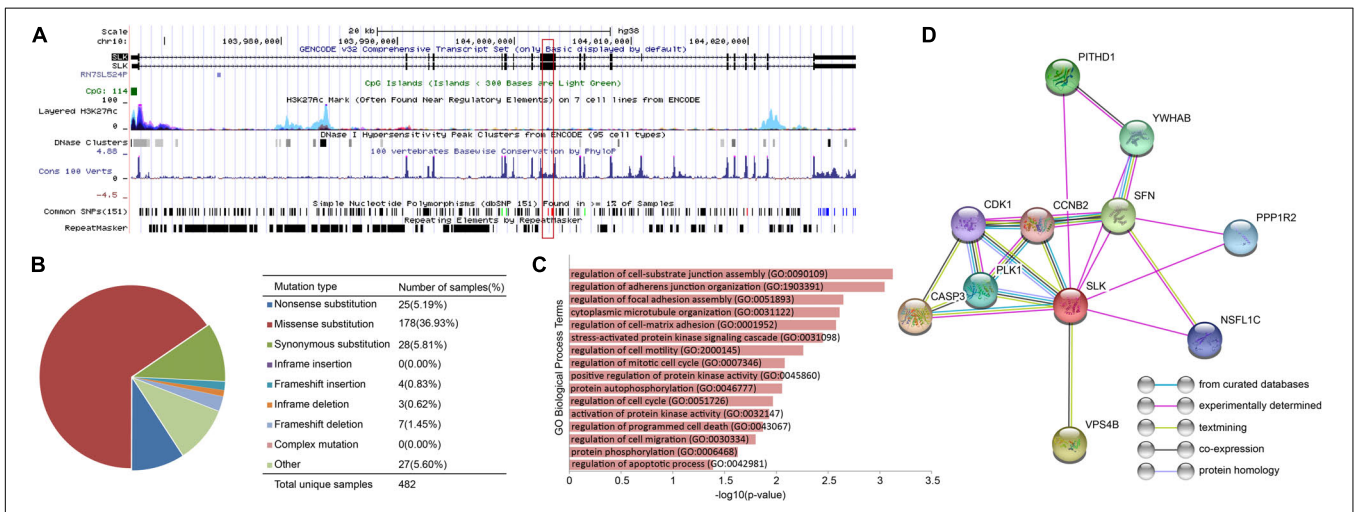


FIGURE 7 | The characteristics of *SLK*. **(A)**The results of UCSC Genome Browser. **(B)** Distribution of mutations on *SLK*. **(C)** The functions of *SLK*. **(D)** The protein–protein interactions of *SLK*.

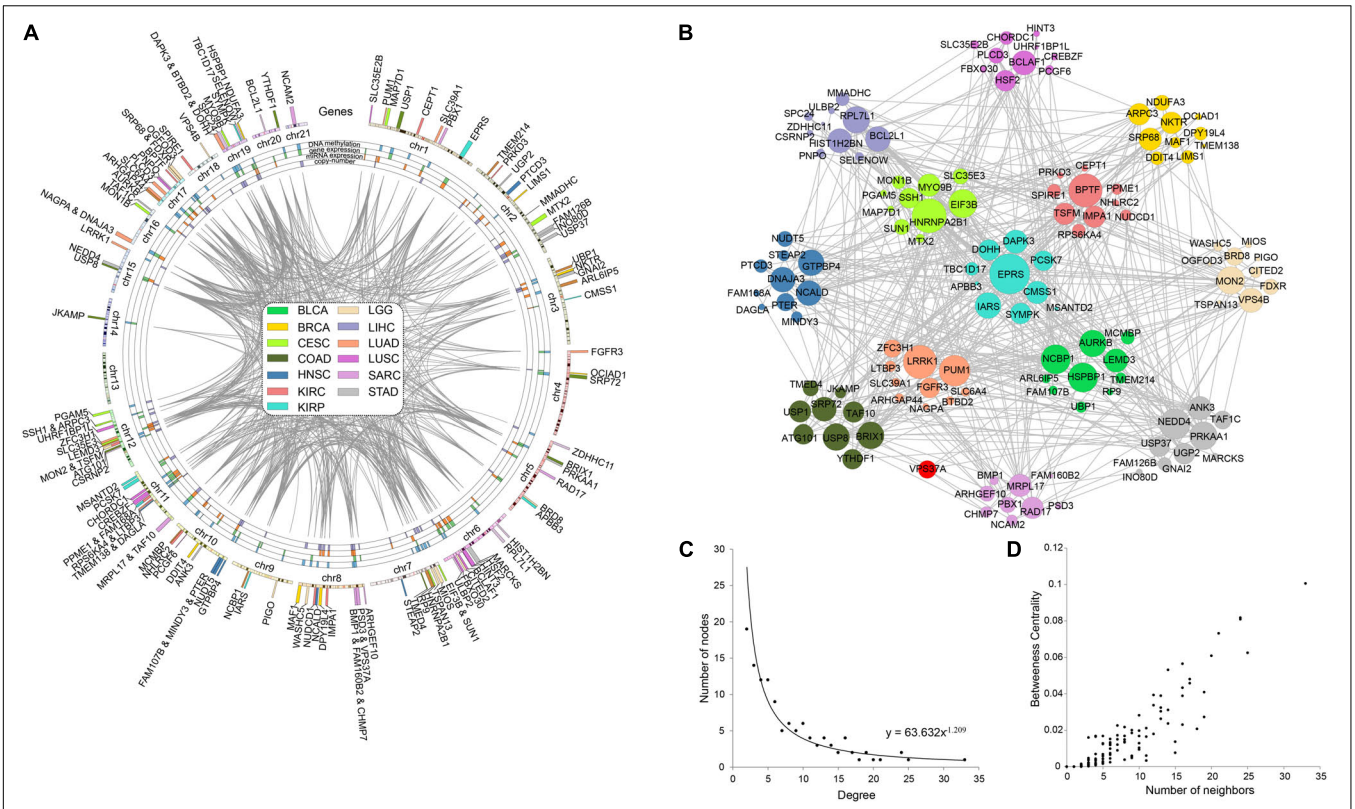


FIGURE 8 | Pan-cancer survival-related gene networks. **(A)** The chromosome distribution of the genes. The blue, green, red, and purple blocks represented the survival correlation of the genes in each omics data, respectively. The links in the middle represented the interaction of the genes. **(B)** The interaction network among the top-10 survival-related genes. Different colors represented prognostic genes of different cancers. Red nodes represented genes prognosis-related in multiple cancers. **(C)** The degree distribution of the prognostic-related gene network. **(D)** The betweeness centrality of the prognostic-related gene network.

genes of different cancers were connected to each other. The degree distribution (**Figure 8C**) and the betweeness centrality (**Figure 8D**) of the network satisfied the condition of scale-free network and were conformed as the general characteristics of biological network. In the network, the degree of *EPRS* had the highest degree of 33. The degrees of *HNRNPA2B1*, *BPTF*, *LRRK1*, and *PUM1* were all greater than 20. These genes mentioned above were widely mutated in many cancers.

Based on the COSMIC database, we found extensive mutations occurred in *EPRS*, and 70% of them were missense mutation. *EPRS* has been shown to be associated with a wide range of cancers by 80 articles. The other four genes also had widespread mutations. In the COSMIC database, 78, 114, 113, and 83 studies confirmed the correlation between *HNRNPA2B1*, *BPTF*, *LRRK1*, and *PUM1* with cancer, respectively.

The Predictive Performance of Our Method

In order to demonstrate the effectiveness of our method, we compared our prognostic biomarkers with previous works. The works using TCGA data were chosen to compare with our work. Chaudhary et al. (2018) used LIHC data of TCGA in their work. Their C-indexs of training and testing set were 0.70(±0.04) and 0.69(±0.08), while our median C-index of LIHC was 0.82. The prognostic power of our method was stronger than theirs. Next, both Yuan et al. (2014) and Zhang et al. (2016) used KIRC and LUSC data of TCGA in their works, so we compared our results of these two cancers with their studies. The comparisons of the C-indexes are shown in **Figures 9A,B**, which showed the higher prognostic power of our 10-gene biomarkers. For KIRC, the median C-index of our work was 0.91. The median C-index of the best performing data (clinical + miRNA) in Yuan's work and the best performing subnetwork (subnetwork K1) of Zhang's study were about 0.76 and 0.74, respectively. For LUSC, the median C-index of our work was 0.76. The median C-index of the best performing data (clinical + protein) in Yuan's work and the best performing subnetwork (subnetwork L1) of Zhang's study were about 0.66 and 0.62, respectively. Therefore, the biomarkers identified by our method could display the better prediction for the patients' survival.

To further confirm reliability of the genes, we downloaded GE data of the corresponding normal samples and used the prognostic biomarkers to cluster the samples. The results showed that the prognostic biomarkers could distinguish the tumor and normal samples (**Figures 9C,D**).

Furthermore, we screened the differentially expressed genes between tumor and normal samples of LUSC and KIRC. After

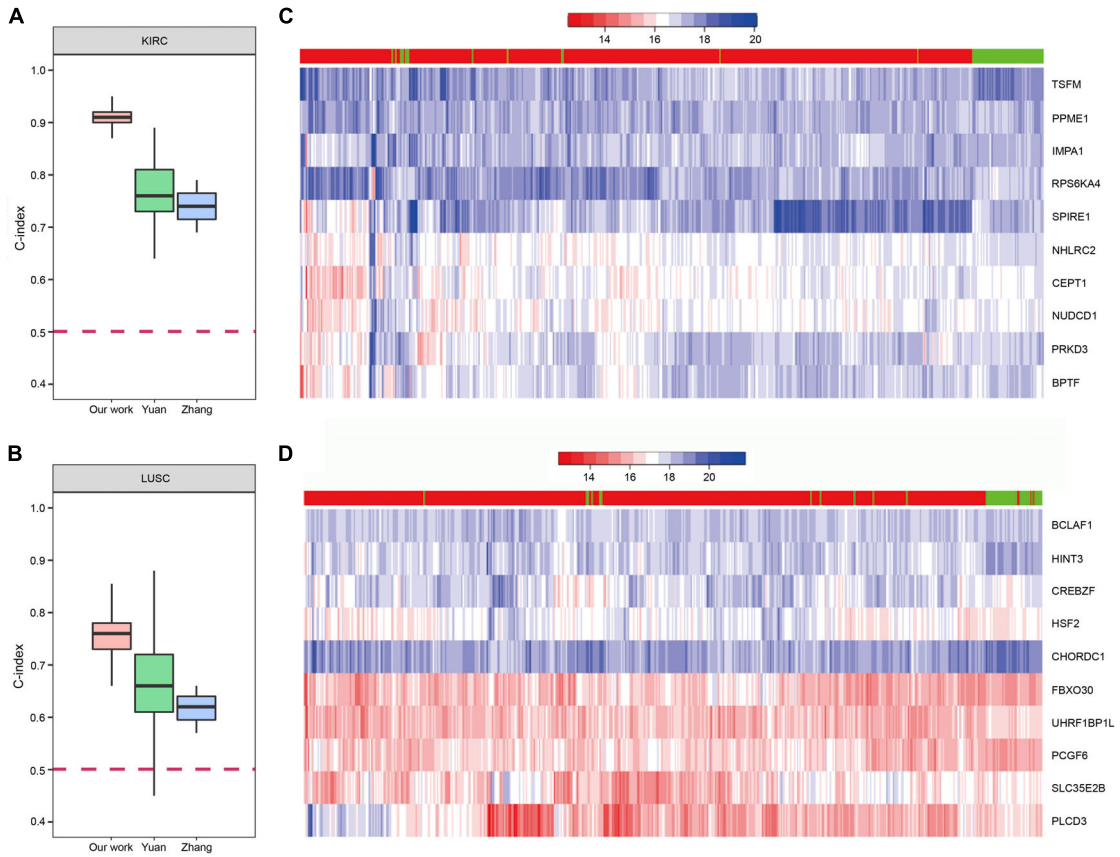


FIGURE 9 | The prognostic biomarkers of KIRC and LUSC. **(A,B)** The C-index comparison of the prognostic power of our 10-gene prognostic biomarkers and other work. **(C,D)** The heatmap of samples hierarchical clustering by the expression of the 10-gene prognostic biomarkers. The bar on the top of the heatmap indicated the group the samples really belong to. Red represented tumor and green represented normal.

comparing them with the list of survival-associated genes, there were 12 differentially expressed genes in LUSC list (**Figure 10A**) and 49 differentially expressed genes in KIRC list (**Figure 10D**). Subsequently, we examined the copy number variation and chromosome location of both the differentially expressed genes and the top-10 biomarker genes (**Figures 10B,C,E,F**). It turned out that among the 22 LUSC genes, six were located in chr 6q, three were in chr 10q, three were in chr 11q, and three were in chr 15q. Of the 59 KIRC genes, 10 were located in chr 8, eight were located in chr 17, and seven were in chr 1. These locations were the peak regions of copy number alternation, suggesting a relationship between these genes and cancer. Moreover, it could be seen that the driver genes of the two cancers were located in different chromosomes, which supported the uniqueness of different cancer-related genes.

Moreover, in the process of univariate Cox regression model, we separately calculated the survival correlation of a gene in four different omics data and then counted them. We also tried the result of considering the same gene in different omics data as different features, and merged the four omics data into one matrix then performed multivariate Cox regression model on it. Only the genes identified as survival-related features more than twice were retained. Finally, the obtained genes were all involved in the gene lists identified through our method and had an incomplete coverage compared with our gene lists. Interestingly, most of these genes were related to survival in GE or SCNA.

In addition, in the process of multivariate Cox regression model, we involved all of the four types of omics data for each candidate gene to perform analysis. Actually, the genes were not survival-related at all of the four omics data in the univariate Cox regression model. To prove the validity of this process, we recalculated the Score of each gene by only using the types of omics data at which the gene was determined to be related to survival in univariate Cox regression model. The results showed that neither the Scores nor the ranks of the genes changed much after recalculation. In consequence, it could suggest the high predictive performance of our multivariable Cox regression model.

The Necessity of Multi-Omics Data Integration

In the process of univariate Cox regression analysis, we found that a gene appeared to be survival-related in one omics dataset, while it might appear to be unrelated to survival on another omics data

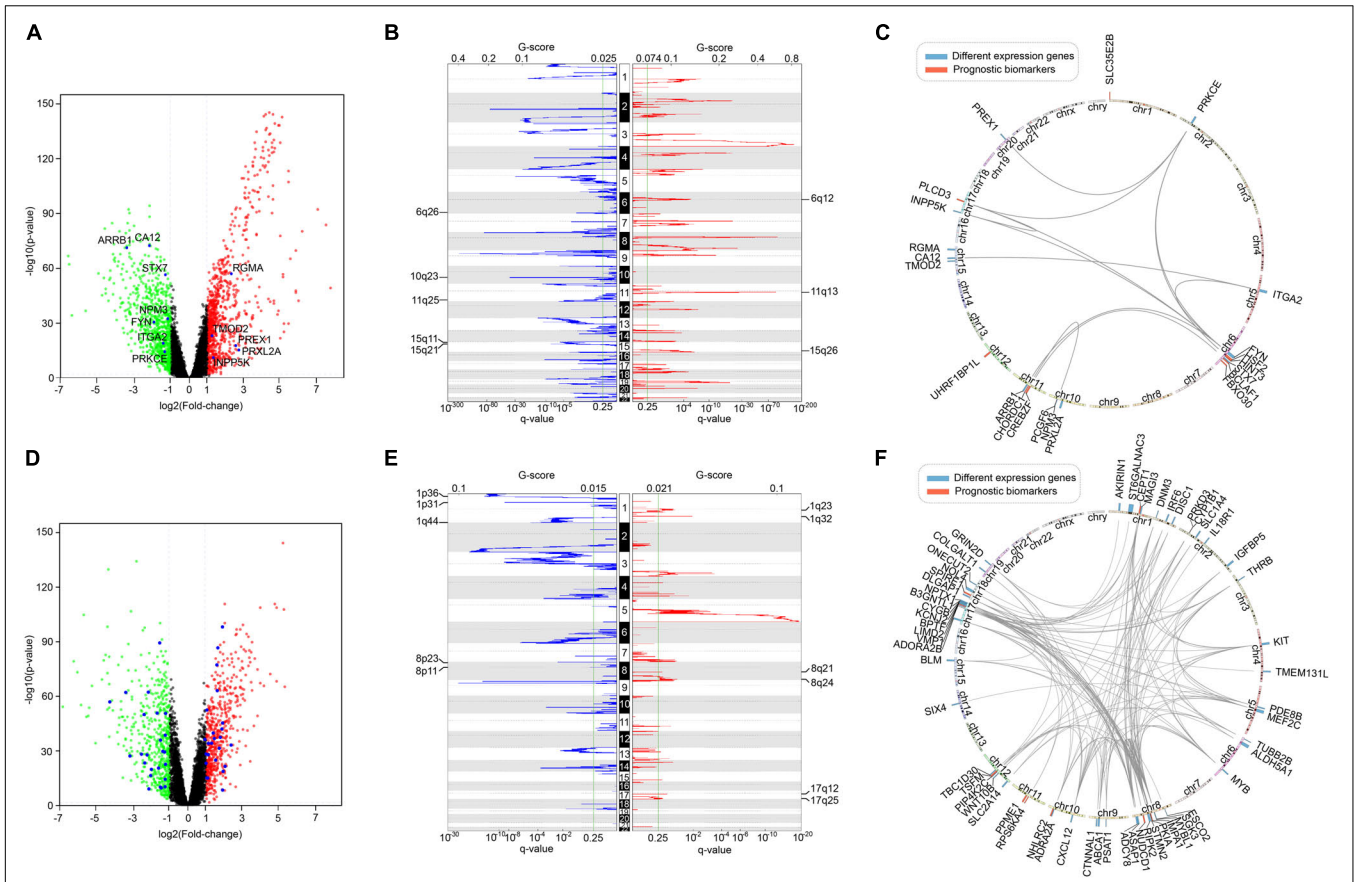


FIGURE 10 | The comparison of survival-related genes and differentially expressed genes. **(A)** The differentially expressed genes of LUSC. Red represented high expression and green represented low expression. Differentially expressed survival-related genes were marked. **(B)** The copy number variation peaks of LUSC. **(C)** Chromosomal positions and interactions of prognostic biomarkers and differentially expressed genes of LUSC. **(D)** The differentially expressed genes of KIRC. Red represented high expression and green represented low expression. Differentially expressed survival-related genes were marked. **(E)** The copy number variation peaks of KIRC. **(F)** Chromosomal positions and interactions of prognostic biomarkers and differentially expressed genes of KIRC.

even under the same model, selection criteria and set of samples. Although this phenomenon might be caused by the error of the data or the imprecision of the experiment, it implied the necessity of multi-omics data integration.

To verify the superiority of integrating multi-omics data, we compared the results of integrating multi-omics data with the results of single omics data in LUSC and KIRC. As shown in **Figure 11**, the results of integrating multi-omics data were significantly higher than those of applying single omics data in decision curve analysis and C-index. The decision curve showed that compared with single omics data, the curve of multi-omics data was further apart from the two extreme curves, which had the greater application value.

DISCUSSION

The recognition of prognostic biomarkers in cancers could predict the prognostic status of each individual patient. This could help to achieve personalized medicine for cancer (Nalejska et al., 2014). Prior work has utilized omics data to predict prognostic status of cancer patients. However, multi-omics data were not used comprehensively.

In this work, we proposed a method to integrate multi-omics data and predict the prognostic status of patients. And gene lists associated with survival were identified in 13 types of cancers. Based on this foundation, the prognostic biomarkers of the cancers were obtained.

Compared with previous studies, this work took a more comprehensive integration of multi-omics data. To verify the reliability and reproducibility of our approach, we confirmed the relationship between our prognostic genes and cancer from multiple perspectives, and the results were stable when changing feature selection strategies. And this method was easy to implement because of its light calculation burden. We obtained candidate survival-related gene lists for 13 types of cancers, and compared the differences and similarities of the lists. The genes associated with survival in multiple cancers were found.

Not only have we successfully verified that genes like *EPRS* were indeed related to various cancers, but also we found that genes such as *SLK* were related to survival of multiple cancers.

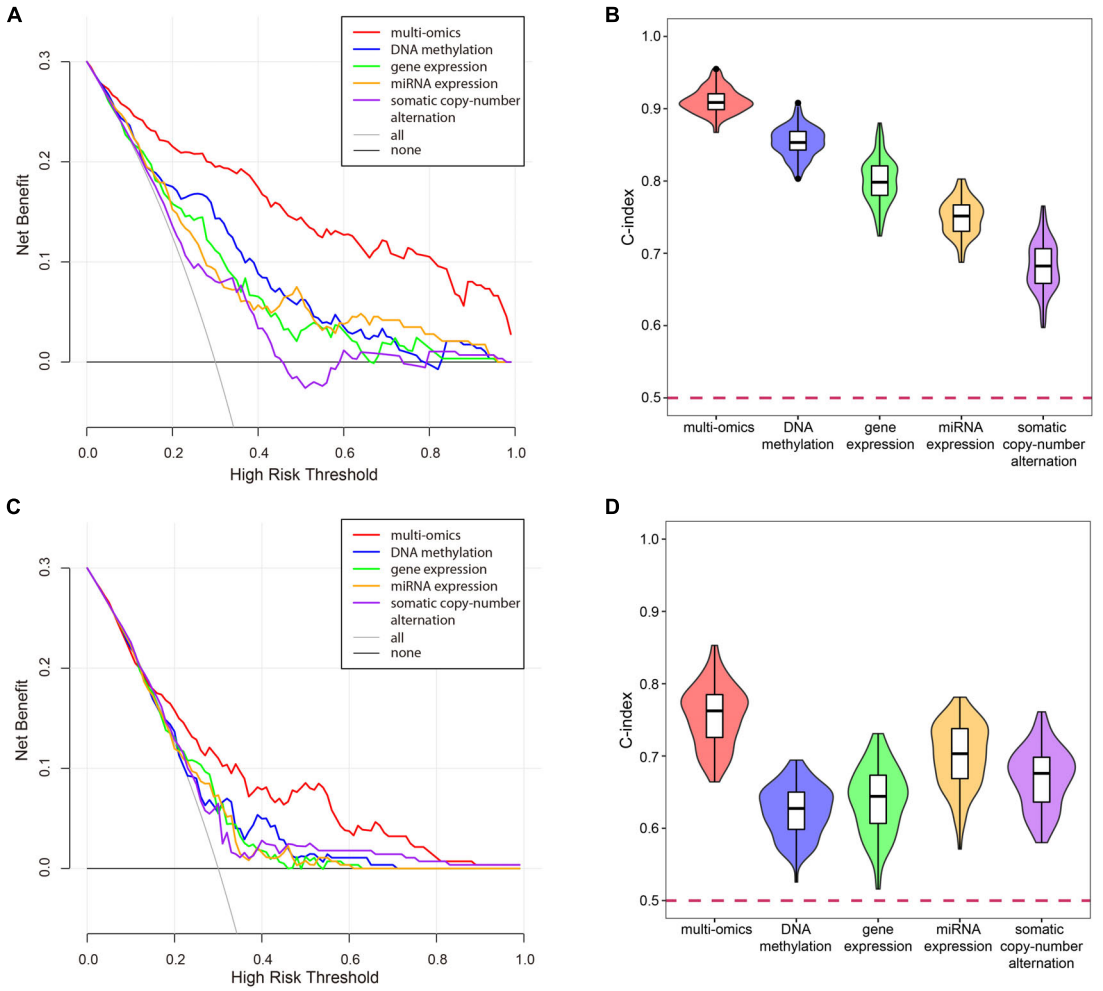


FIGURE 11 | The comparison of the results for multi-omics data and single omics data. **(A)** The decision curve of multi-omics data and each omics data in KIRC. The thin oblique line represented the assumption that all patients have been treated. The black line represented the assumption that no patients have been treated. **(B)** The C-indexes of multi-omics data and each omics data in KIRC. **(C)** The decision curve of multi-omics data and each omics data in LUSC. The thin oblique line represented the assumption that all patients have been treated. The black line represented the assumption that no patients have been treated. **(D)** The C-indexes of multi-omics data and each omics data in LUSC.

SLK has been reported to associate with blood cancer, breast cancer, colorectal cancer, liver cancer, and pancreatic cancer. In our work, we found that it was participated in the BP of patients' survival of bladder cancer, lung cancer, and renal cancer. *SLK* mainly associated with cell adhesion. Cell adhesion plays an important role in the maintenance of tissue structure, whose abnormality results in tumor invasion and metastasis.

However, we also got some confused results in comparing different cancers. As different cancer subtypes of the same tissue, the overlap of gene lists between LUAD and LUSC was small, which was different from the expected outcome. We suspected that this might be due to their different pathogenesis. LUSC commonly occurred in older men and was strongly associated with smoking, but LUAD was more common in women and non-smokers (Kenfield et al., 2008). The differences in the pathogenesis might lead to differences in the genetic mechanisms and the list of related genes. Moreover, the prognostic markers for SARC did not significantly distinguish the high and low risk groups. This might be due to the subtypes of SARC (leiomyosarcoma, liposarcoma, myxofibrosarcoma, synovial SARC, etc.). The subtypes of SARC ought to be considered as different cancer types.

In addition, this might be caused by the bias of data. TCGA patient samples were selected from multiple sources, and were characterized at multiple centers, which might introduce heterogeneity and bias. And the clinical annotations of the patients might not be sufficiently rigorous and comprehensive (Yuan et al., 2014). Even though we only selected the basic information such as age and gender, there were still some missing values.

Till now, only a few molecular prognostic biomarkers based on multi-omics data have been applied to clinic (Yuan et al., 2014). The presence of publication bias and incompletion in the literatures is a major reason why the identified tumor biomarkers

have not been applied in clinic (Mcshane and Hayes, 2012). Further, translational medicine researchers have no access to the results of these studies. Our work clearly provided gene lists related to the survival of various cancers, which could be easily obtained and searched, and help to transform biological data into clinical experiments.

Even so, our work remains inadequate. First of all, overfitting and collinearity of biological data make it technically challenging to effective integration of multi-omics data. Our work did not solve the problem. Although LASSO can well select the most important features to overcome the overfitting problem, it will lose many equally important features at the same time when high pairwise correlations occurred (Zhang et al., 2016). And the intra-tumor heterogeneities of cancer make it almost impossible to find prognostic biomarkers 100% suitable for each patient. Future efforts are still needed to address these problems.

In addition, since the data were downloaded from TCGA which was a program of the National Cancer Institute (NCI) of the United States, most of the patients were white people. The results of this study may be only appropriate for the whites. Although Chaudhary et al. (2018) have validated their model, which was built by TCGA data, on Japanese and Chinese datasets, further validation of other cancer should be done and data of black population should be included in future studies. Furthermore, some studies have suggested a non-linear relationship between miRNA expression and clinical outcomes (Fuchs et al., 2013; Lee et al., 2013). Therefore, some non-parametric algorithms can be applied to the analysis of the prognosis of miRNA in future studies.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the TCGA database: https://www.cancer.gov/tcga. And the core code of this study was merged in the **Supplementary Data Sheet S1**.

AUTHOR CONTRIBUTIONS

NZ collected data, carried out the initial analyses, and drafted the manuscript. MG conceived of the study, participated in its design and coordination, and helped to draft the manuscript. KW and XL coordinated and supervised data collection, and critically commented on the important intellectual content of the manuscript. CZ participated in the design of the study and performed the statistical analysis. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61532014 and 61671189) and the National Key Research and Development Plan Task of China (Grant No. 2016YFC0901902).

ACKNOWLEDGMENTS

We thank the TCGA database for sharing the multi-omics data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00268/full#supplementary-material

TABLE S1 | The complete clinical information for each sample.

TABLE S2 | The survival-related gene list of each cancer.

TABLE S3 | The enriched functions and pathways of each cancer.

TABLE S4 | The functions of the seven pan-cancer survival-related genes.

DATA SHEET S1 | Core code of this study.

REFERENCES

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statist. Med.* 8, 907–925. doi: 10.1002/sim.4780080803

Abbott, K. L., Nyre, E. T., Abrahante, J., Ho, Y. Y., Isaksson Vogel, R., and Starr, T. K. (2015). The candidate cancer gene database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.* 43, D844–D848. doi: 10.1093/nar/gku770

Alamartine, E., Sabatier, J. C., Guerin, C., Berliet, J. M., and Berthoux, F. (1991). Prognostic factors in mesangial IgA glomerulonephritis: an extensive study with univariate and multivariate analyses. *Am. J. Kidney Dis.* 18, 12–19. doi: 10.1016/s0272-6386(12)80284-8

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.

Baylin, S. B. (2005). DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* 2(Suppl. 1), S4–S11.

Berdasco, M., and Esteller, M. (2010). Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Dev. Cell* 19, 698–711. doi: 10.1016/j.devcel.2010.10.005

Bernardini, J. P., Lazarou, M., and Dewson, G. (2017). Parkin and mitophagy in cancer. *Oncogene* 36, 1315–1327. doi: 10.1038/onc.2016.302

Burrell, R. A., Mcgranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi: 10.1038/nature12625

Cagney, D. N., Sul, J., Huang, R. Y., Ligon, K. L., Wen, P. Y., and Alexander, B. M. (2018). The FDA NIH biomarkers, endpoints, and other tools (BEST) resource in neuro-oncology. *Neurol. Oncol.* 20, 1162–1172. doi: 10.1093/neuonc/nox242

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489, 519–525. doi: 10.1038/nature11404

Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43–49. doi: 10.1038/nature12222

Cancer Genome Atlas Research Network, Linehan, W. M., Spellman, P. T., Ricketts, C. J., Creighton, C. J., Fei, S. S., et al. (2016). Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* 374, 135–145. doi: 10.1056/NEJMoa1505917

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* 24, 1248–1259. doi: 10.1158/1078-0432.CCR-17-0853

Collett, D. (2015). *Modelling Survival Data In Medical Research*. Boca Raton, FL: Chapman and Hall/CRC.

Cox, D. (1986). Citation-classic - regression-models and life-tables. *Curr. Contents Agric. Biol. Environ. Sci.* 34:16.

Dalerba, P., Sahoo, D., Paik, S., Guo, X., Yothers, G., Song, N., et al. (2016). CDX2 as a prognostic biomarker in stage II and stage III colon cancer. *N. Engl. J. Med.* 374, 211–222. doi: 10.1056/NEJMoa1506597

Fehrmann, R. S., Karjalainen, J. M., Krajewska, M., Westra, H. J., Maloney, D., Simeonov, A., et al. (2015). Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* 47, 115–125. doi: 10.1038/ng.3173

Fuchs, M., Beissbarth, T., Wingender, E., and Jung, K. (2013). Connecting high-dimensional mRNA and miRNA expression data for binary medical classification problems. *Comput. Methods Program. Biomed.* 111, 592–601. doi: 10.1016/j.cmpb.2013.05.013

Gene Ontology, C. (2015). Gene ontology consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056. doi: 10.1093/nar/gku1179

Group, P. T. C., Calabrese, C., Davidson, N. R., Demircioglu, D., Fonseca, N. A., He, Y., et al. (2020). Genomic basis for RNA alterations in cancer. *Nature* 578, 129–136. doi: 10.1038/s41586-020-1970-0

Harrell, F. E. Jr., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 15, 361–387. doi: 10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8:84. doi: 10.3389/fgene.2017.00084

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

Kenfield, S. A., Wei, E. K., Stampfer, M. J., Rosner, B. A., and Colditz, G. A. (2008). Comparison of aspects of smoking among the four histological types of lung cancer. *Tob. Control.* 17, 198–204. doi: 10.1136/tc.2007.022582

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, I. H., Lee, S. H., Park, T. H., and Zhang, B. T. (2013). Non-linear molecular pattern classification using molecular beacons with multiple targets. *Biosystems* 114, 206–213. doi: 10.1016/j.biosystems.2013.05.008

Lindahl, L. M., Besenbacher, S., Rittig, A. H., Celis, P., Willerslev-Olsen, A., Gjerdrum, L. M. R., et al. (2018). Prognostic miRNA classifier in early-stage mycosis fungoides: development and validation in a Danish nationwide study. *Blood* 131, 759–770. doi: 10.1182/blood-2017-06-788950

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, 400–416e411. doi: 10.1016/j.cell.2018.02.052

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* 12, 323–334. doi: 10.1038/nrc3261

Mcshane, L. M., and Hayes, D. F. (2012). Publication of tumor marker research results: the necessity for complete and transparent reporting. *J. Clin. Oncol.* 30, 4223–4232. doi: 10.1200/JCO.2012.42.6858

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

Mishra, N. K., Southekal, S., and Guda, C. (2019). Survival analysis of multi-omics data identifies potential prognostic markers of pancreatic ductal adenocarcinoma. *Front. Genet.* 10:624. doi: 10.3389/fgene.2019.00624

Morikawa, A., Hayashi, T., Kobayashi, M., Kato, Y., Shirahige, K., Itoh, T., et al. (2018). Somatic copy number alterations have prognostic impact in patients with ovarian clear cell carcinoma. *Oncol. Rep.* 40, 309–318. doi: 10.3892/or.2018.6419

Nalejska, E., Maczynska, E., and Lewandowska, M. A. (2014). Prognostic and predictive biomarkers: tools in personalized oncology. *Mol. Diagn. Ther.* 18, 273–284. doi: 10.1007/s40291-013-0077-9

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* 46, 10546–10562. doi: 10.1093/nar/gky889

Rodriguez-Martin, B., Alvarez, E. G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., et al. (2020). Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* 52, 1–14. doi: 10.1038/s41588-019-0562-0

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

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18, 696–705. doi: 10.1038/s41568-018-0060-1

Swanton, C. (2012). Intratumor heterogeneity: evolution through space and time. *Cancer Res.* 72, 4875–4882. doi: 10.1158/0008-5472.CAN-12-2217

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

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

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

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., et al. (2017). The UCSC genome browser database: 2017 update. *Nucleic Acids Res.* 45, D626–D634. doi: 10.1093/nar/gkw1134

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963. doi: 10.1093/nar/gkx1090

Vickers, A. J., Cronin, A. M., Elkin, E. B., and Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med. Inform. Decis. Mak.* 8:53. doi: 10.1186/1472-6947-8-53

Xu, A., Chen, J., Peng, H., Han, G., and Cai, H. (2019). Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences. *Front. Genet.* 10:236. doi: 10.3389/fgene.2019.00236

Yang, J. H., Li, J. H., Shao, P., Zhou, H., Chen, Y. Q., and Qu, L. H. (2011). starBase: a database for exploring microRNA-mRNA interaction maps from argonaute CLIP-Seq and degradome-seq data. *Nucleic Acids Res.* 39, D202–D209. doi: 10.1093/nar/gkq1056

Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* 32, 644–652. doi: 10.1038/nbt.2940

Zhang, F., Ren, C., Lau, K. K., Zheng, Z., Lu, G., Yi, Z., et al. (2016). A network medicine approach to build a comprehensive atlas for the prognosis of human cancer. *Brief Bioinform.* 17, 1044–1059.

Zhao, S., Geybels, M. S., Leonardson, A., Rubicz, R., Kolb, S., Yan, Q., et al. (2017). Epigenome-wide tumor DNA methylation profiling identifies novel prognostic biomarkers of metastatic-lethal progression in men diagnosed with clinically localized prostate cancer. *Clin. Cancer Res.* 23, 311–319. doi: 10.1158/1078-0432.CCR-16-0549

Zhu, B., Song, N., Shen, R., Arora, A., Machiela, M. J., Song, L., et al. (2017). Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Sci. Rep.* 7:16954. doi: 10.1038/s41598-017-17031-8

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 © 2020 Zhao, Guo, Wang, Zhang 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 03 April 2020
doi: 10.3389/fbioe.2020.00241

Robust Biomarker Screening Using Spares Learning Approach for Liver Cancer Prognosis

Aman Chandra Kaushik[1,2], Aamir Mehmood[2,3], Dong-Qing Wei[2,3] and Xiaofeng Dai[1]**

[1] Wuxi School of Medicine, Jiangnan University, Wuxi, China, [2] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [3] Peng Cheng Laboratory, Shenzhen, China

LncRNAs, miRNAs, mRNAs, methylation, and proteins exert profound biological functions and are widely applied as prognostic features in liver cancer. This study aims to identify prognostic biomarkers' signature for liver cancer. Samples with inadequate tumor purity were filtered out and the expression data from different resources were retrieved. The Spares learning approach was applied to select lncRNAs, miRNAs, mRNAs, methylation, and proteins' features based on their differentially expressed groups. The LASSO boosting technique was employed for the predictive model construction. A total of 200 lncRNAs, 200 miRNAs, 371 mRNAs, 371 methylations, and 184 proteins were observed to be differentially expressed. Five lncRNAs, 11 miRNAs, 30 mRNAs, 4 methylations, and 3 proteins were selected for further evaluation using the feature elimination technique. The highest accuracy of 89.32% is achieved as a result of training and learning by Spares learning methodology. Final outcomes revealed that 5 lncRNA, 11 miRNA, 30 mRNA, 4 methylation, and 3 protein signatures could be potential biomarkers for the prognosis of liver cancer patients.

Keywords: biomarkers, spares learning, liver cancer, prognosis, ncRNA

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Nitish Kumar Mishra,
University of Nebraska Medical Center, United States
Shaoliang Peng,
Hunan University, China

****Correspondence:***
Dong-Qing Wei
dqwei@sjtu.edu.cn
Xiaofeng Dai
1281423490@qq.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

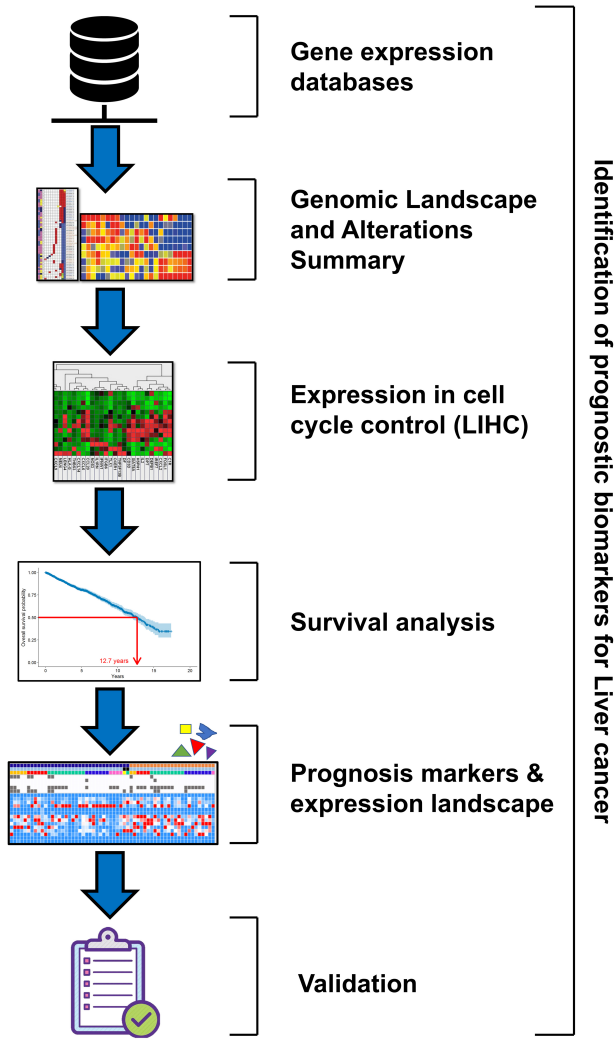
Received: *16 December 2019*
Accepted: *09 March 2020*
Published: *03 April 2020*

Citation:
Kaushik AC, Mehmood A, Wei D-Q and Dai X (2020) Robust Biomarker Screening Using Spares Learning Approach for Liver Cancer Prognosis. Front. Bioeng. Biotechnol. 8:241. doi: 10.3389/fbioe.2020.00241

INTRODUCTION

One of the largest organs in the human body is the liver, which is crucial for metabolism and is helpful in detoxification and maintaining homeostasis. Many ailments are concerned with the liver including hepatitis, fibrosis, genetic and metabolic issues, and liver cancer, which is one of the leading causes of cancer-related expiries (Bray et al., 2018; Dooley et al., 2018; Yang et al., 2019). Hepatocellular carcinoma (HCC) can occur in an ailing liver and encompasses numerous molecular cascades (Kanda et al., 2019). It is reported that more than 90% of liver cancers are HCC, which is an extremely assorted form of cancer verified by high-throughput sequencing and gene expression profiling, at both the molecular and histological levels (Calderaro et al., 2019).

Gene therapy has progressed as an effective source of dealing with disease-causing gene imperfections to attain a typical status. The approaches employed to treat illness by gene therapy consist of gene replacement, gene restoration, gene extension, gene muzzling, vaccination, and, currently, gene-editing technology (Alsaggar and Liu, 2016; Chew et al., 2016; Karimian et al., 2019). Thus, the identification of a gene that can be used as a potential biomarker is an important step in the treatment of liver cancer (West et al., 2019; Zheng et al., 2019; Lu et al., 2020; von Felden and Villanueva, 2020).



GRAPHICAL ABSTRACT | The graphical abstract depicts the pipeline of methodology that represents the flow of work for identification of prognostic biomarkers for liver cancer.

Statistical approaches like artificial neural networks employing BI-RADS (Baker et al., 1995) and logistic regressions have been used in several reports to improve diagnostic performance. It is most beneficial to use statistical approaches as they enhance the identification of breast cancer, with BI-RADS, as well as along the medical and statistical information concerning infected persons' statistical threat aspects (Chhatwal et al., 2009). Regression processes go through overfitting once the prognostic covariates are involved in a large number. Similar situations lead to the well-fitting of a deterioration mockup into the drilling information. However, this doesn't go parallel with the cases of the real world. Variable selection becomes a necessity in an attempt to get exact predictions associated with covariates of a large number, for instance, BI-RADS qualifiers and statistical data. A very famous fact claims the unfavorability of regular step-by-step assortment methods in case of regression models that have numerous covariates (Houssami et al., 2004). On the other hand, sparse penalized methods, like the minimum complete reduction and assortment operative (LASSO), together have gained ample consideration. LASSO is a penalized regression technique that approximates the deterioration constants through enhancing the log-similarity purpose (or adding the squared remainders) having restriction that the addition of the total scores of the deterioration constants, $\Sigma kj = 1||\beta j||$, is </= to a positive constant *s*. LASSO has one of the most fascinating characters that the approximated value belonging to deterioration constants is tenuous in nature, indicating a lot of components that are accurately 0. This proves that unnecessary covariates are automatically deleted by LASSO. It is believed that LASSO has numerous required characteristics that are compulsory for the deterioration mockups with a huge covariate count. Optimization algorithms for the rectilinear deterioration model as well as for general rectilinear mockups are available in large numbers, with good efficiency. As per our information, this work is the first attempt to build a calculated LASSO deterioration mockup that could assist in the diagnosis of breast cancer based on statistical and radiological findings.

This study is aimed to compare the productivity of graphical examination to forecast liver cancer dependent on whether the calculated LASSO deterioration or bit-by-bit calculated (SL) deterioration was employed, along with evaluating the practicality of integrating statistical data into the graphic breakdown for the sake of improving liver cancer diagnosis.

MATERIALS AND METHODS

Gene Expression Databases

The Cancer Genome Atlas (TCGA)[1] catalog can be accessed to gain information regarding alterations in the gene, long non-coding RNAs (lncRNAs), methylations, miRNAs, mRNAs, CAN, mutational expression, and proteins involved in HCC. It is a freely accessible repository at the TCGA (Cerami et al., 2012; Gao et al., 2013). The cancer study "HCC were obtained from TCGA" and information type precedence "Mutation and CNA (DNA copy-number alterations)" were selected before analyzing genomic alterations of cell cycle control in the TCGA data on HCC. This did not require any statement of approval or informed consent for the reason that the information is retrieved from a public repository.

Genomic Alterations Summary

Genomic modifications of cell cycle control via tumor samples were summarized. Genomic modifications inclusive of mutations, CNA (amplifications and homozygous deletions), glyphs, and dye tagging were practiced to summarize the gene expression variations. It was the first step to understand various types of gene signaling in HCC. The shared exclusivity and co-occurrence among cell cycle control were studied as well. Discordant, gene-related happenings linked with a specific cancer are most of the time conflicting in a cluster of tumors,

[1]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

i.e., only single biological happening is expected to occur in every sample of cancer. Another condition is the simultaneous incidence that several genes are changed in each sample (Gao et al., 2013); this was an introductory way to collect information related to various gene signaling in HCC.

Mutations in Cell Cycle Control in HCC

Through the mutations of cell cycle control, the rate and position for all mutations within Pfam protein fields were specified. Colored bars denote an entire extent of cell cycle control proteins and the base of every bar in gray denotes the amino acid count. Protein domains are represented by the boxes colored in red, blue, and green. The lines and points signify the position and rate of genes. The frameshift or nonsense mutations are shown in red, missense mutations are in green, and the black color represents the in-frame deletions (Fang et al., 2015).

Survival Analysis

The survival analysis bears great importance in prognosis to highlight changes in the survival rate. Here, the differences in the overall survival were evaluated via survival analysis among samples having a single or more alteration as that of the inquired genes(s) and also the samples that have no variation.

Statistics

To carry out correlation analysis, a scattered graph of lncRNAs, methylations, miRNAs, mRNAs, CAN, mutational expression, or protein level in every sample was presented. The Kaplan–Meier approach having log-rank tests are carried out for comparing global and healthy survival of HCC that have at least a single modification or lack any adjustment within the inquired gene(s). Samples with up-regulation were recognized by the verge of $Z > 2$ (mean expression over 2 SDs). The standard was fixed at 0.05.

Acquisition of Patient's Data

LncRNAs appeared as potential features in the field of oncology. RNA-seq data are obtained from TCGA while the exploration of lncRNAs in cancer is provided by an open-access web app "TANRIC." The TANRIC (The Atlas of ncRNA In Cancer)[2] allows rapid and intuitive analyses of lncRNAs in the framework of experimental and other molecular information. Through TANRIC, a high amount of lncRNAs were identified with probable biomedical implication, where the majority of them shows

[2]https://ibl.mdanderson.org/tanric/_design/basic/index.html

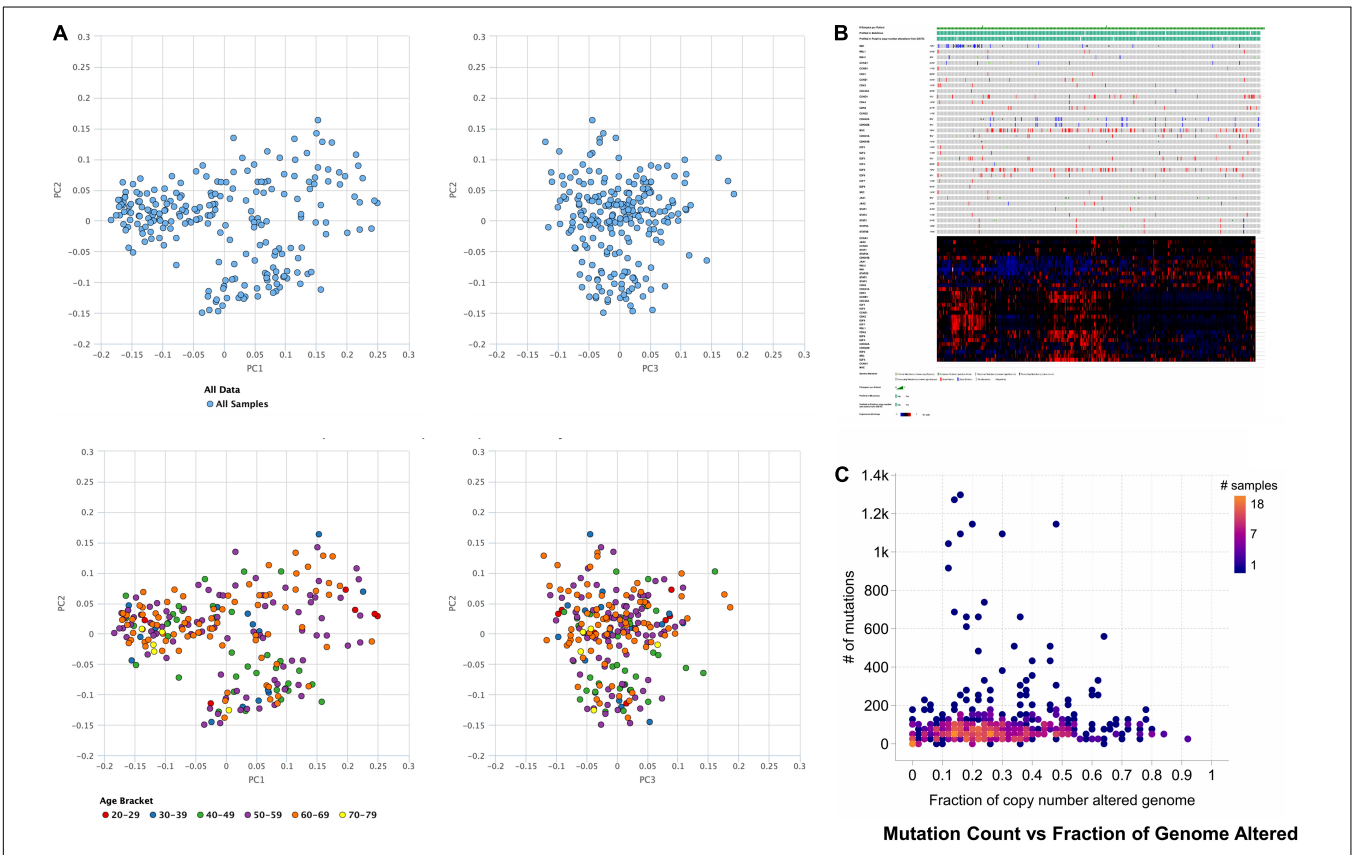


FIGURE 1 | Principal component analysis of hepatocellular carcinoma where the TCGA data and age is **(A)** used for principal component analysis. **(B)** Genetic alterations in the number of samples per patients, illustrating that activation of cell cycle via missense mutations mediates gene signaling in case of the hepatocellular carcinoma, or gene signaling can also be mediated by the inactivation of cell cycle control via truncating mutations. **(C)** The number of mutation counts vs. a fraction of the copy number alterations in the genome.

robust associations with the already formed therapeutic goals and biomarkers across the cell lines or tumor types. We retrieved lncRNA, miRNA, mRNA, methylation, and protein expression data from TANRIC (Li et al., 2015) of all the TCGA liver cancer patients (Ciriello et al., 2015). The corresponding clinical data are retrieved from Genomic Data Commons (GDC)[3].

[3]https://gdc.cancer.gov/

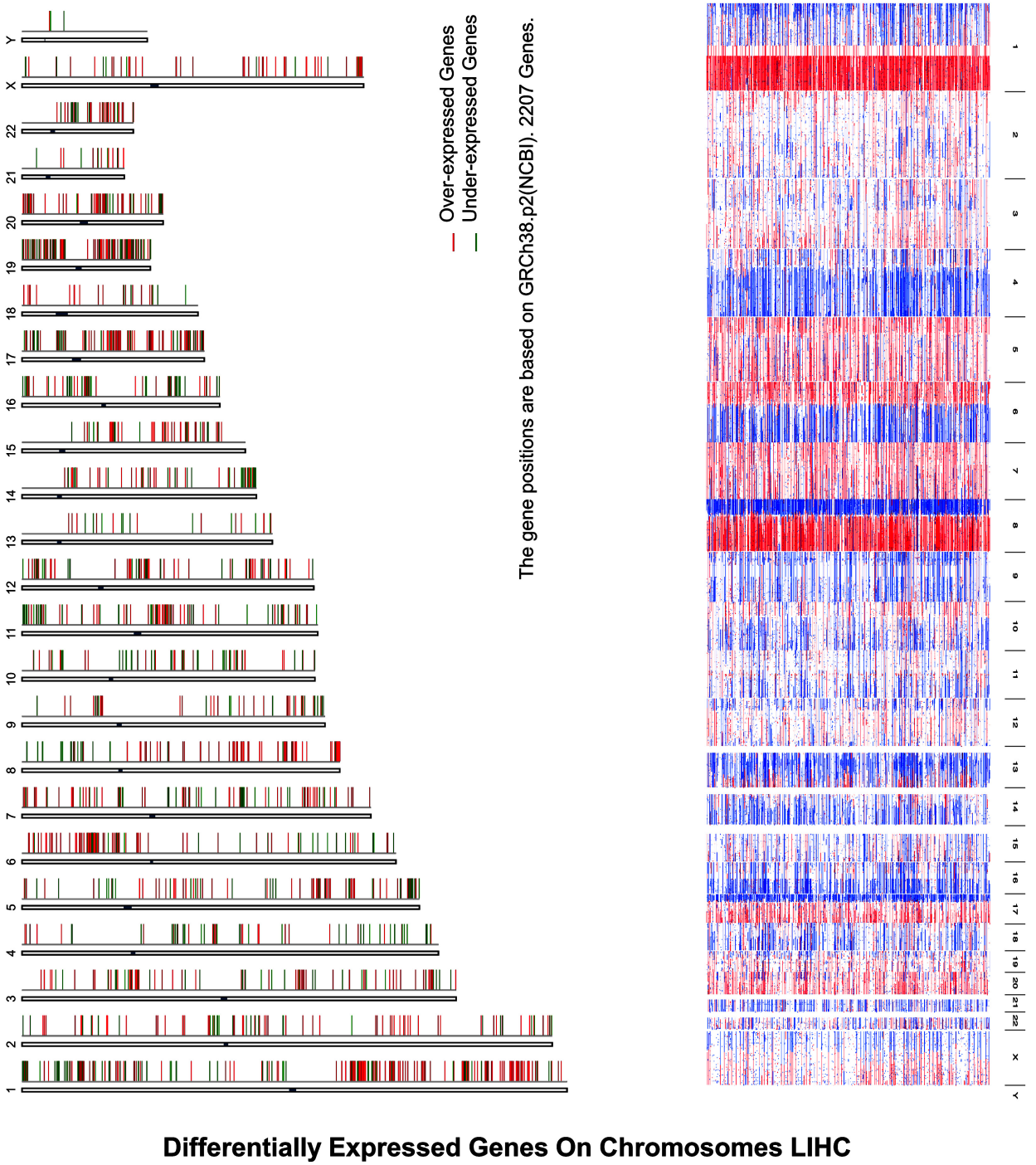


FIGURE 2 | Differentially expressed genes on chromosomes LIHC depict differentially expressed genes on chromosomes in HCC, where expression analysis reveals that there is a significant cell cycle control in under- and overexpressed HCC, explaining that these were hotspots for the activation, where *X*-axis represents the over- and underexpression of genes while *Y*-axis indicates the chromosomes' number.

Purity estimation was performed for the patients using consensus purity estimate and the Clonal Heterogeneity Analysis Tool (Li et al., 2012; Li and Li, 2014). Patients were filtered out with purity estimators below 60%.

Feature Identification for lncRNAs, miRNAs, mRNAs, Methylations, and Proteins

For the identification of promising discriminative lncRNAs, miRNAs, mRNAs, methylations, and proteins of survival groups, the R limma[14] package was used to identify promising discriminative biomarkers by analyzing the differential expression of lncRNAs, miRNAs, mRNAs, methylations, and proteins.

Feature Selection of lncRNAs, miRNAs, mRNAs, Methylations, and Proteins

The differentially expressed lncRNAs, miRNAs, mRNAs, methylations, and proteins were used as input features for predictive modeling. Spares learning was applied to select features. The Spares learning and LASSO method were ranked by features based on specific importance.

Predictive Modeling and Expression Landscape

We used Spares learning and LASSO to construct the predictive model of survival groups. LASSO is a powerful ensemble learning method that has achieved state-of-the-art performance in many biomedical tasks.

$$g_i = G_{-I} b_i + \varepsilon_i, \text{ subject to } |b_i| \leq C_i, \quad (1)$$

where b_i is the coefficient of expressions other than RNA's i, $|\cdot|$ is an L-1 norm, and the residual is denoted as ε_i. The j_{th} coefficient element in b_i indicates a regulatory relationship from RNAs j to RNAs i (with a direction) in the linear model, where zero shows non-relationship between them. In contrast with correlation-based RNA regulatory networks, linear regression-based RNA regulatory networks can capture the main effects of multiple

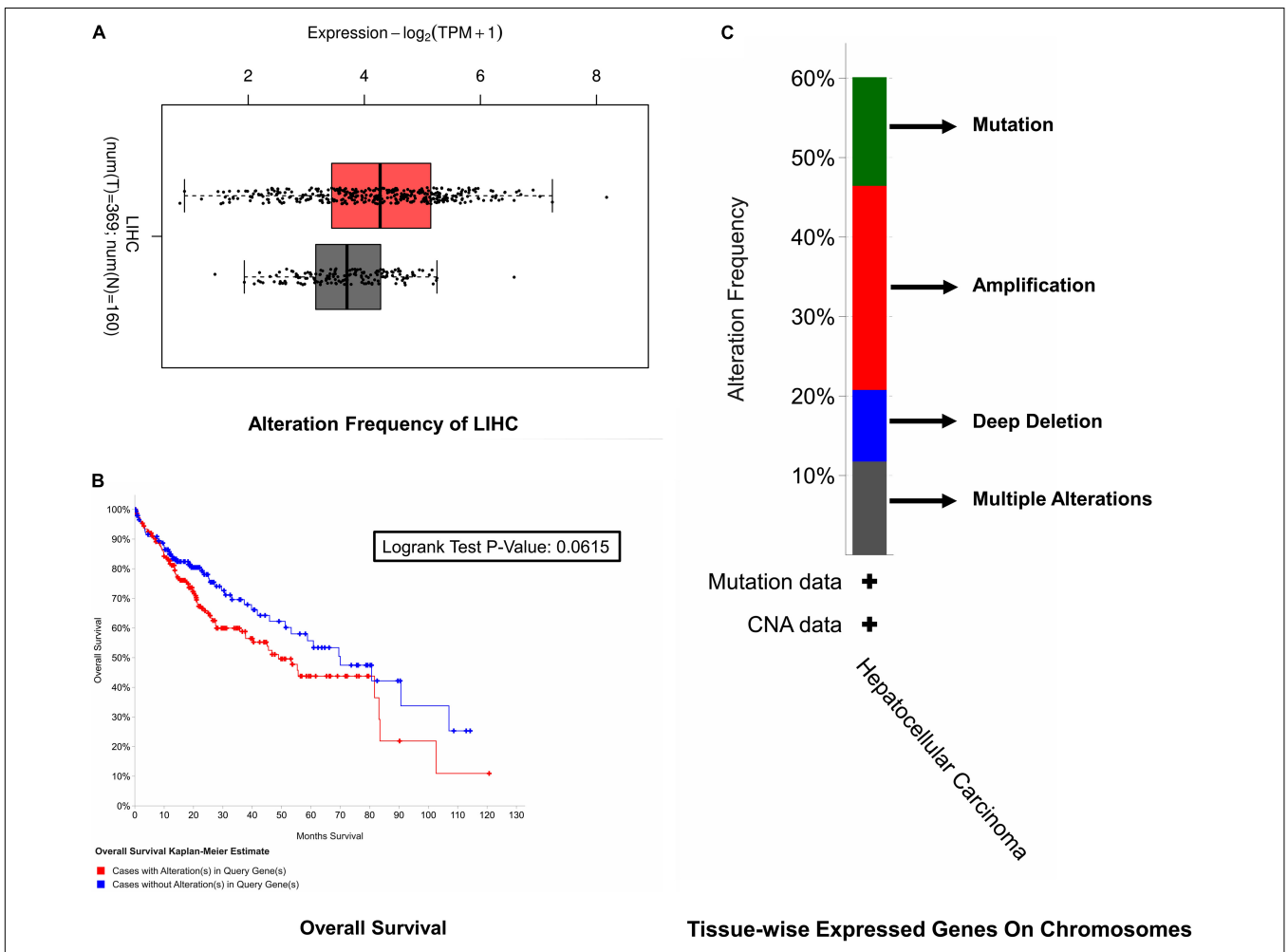


FIGURE 3 | Survival rate inspection. **(A)** Tissue-wise expressed gene on chromosomes. **(B)** Overall survival analysis of the cell cycle control neighborhood in HCC. **(C)** CAN and alteration frequency of cell cycle control in HCC.

RNAs. Correlation-based RNA regulatory networks may fail to infer RNA regulation if the correlation is not significantly high and if multiple RNAs regulate simultaneously. The coefficient vector b_i for RNAs i is used for constructing the adjacency matrix B of the group-specific RNAS regulatory network, i.e., $B = \{b_{1,\ldots,} b_p\}^T$. Then, Eq. (1) can be optimized by:

$$\|g_i - G_{-I} b_i\|^2 + \lambda |b_i|, \quad (2)$$

where λ is a hyper-parameter for sparsity regularization, and $\|\cdot\|^2$ is an L-2 norm of a vector.

Algorithm

1: $I \in \{1, \ldots, p\}$ ***do***
2: $\boldsymbol{b_i^T}$ = Random LASSO in equation (2)
3: $\boldsymbol{b_i^C}$ = Random LASSO in equation (2)
4: ***end for***
5: $\boldsymbol{B^T} = \{\boldsymbol{b_I^T}, \ldots, \boldsymbol{b_p^T}\}$
6: $\boldsymbol{B^C} = \left\{\boldsymbol{b_I^C}, \ldots, \boldsymbol{b_p^C}\right\}$

RESULTS AND DISCUSSION

In this study, data from the TCGA Cancer Genomics have been used to explore, visualize, and analyze the genetic and medical features of alterations in cell cycle control found in cases of HCC from databases of TCGA. As per our knowledge, this study is the opening data mining approach that tends to discover the existing connection among modifications occurring in control of cell cycle and patients' prognosis. A lot of conclusions in this study are coherent with the previously reported data. Remarkably, we detected in our study that alterations in the cell cycle control mostly exist in HCC. Variations in these genetic factors are on autonomous cascades to HCC and are in an uncommon fashion of increasing gene changes. Although no cell cycle control was linked with any of the survival events (disease-free and global survival) in this work, it provides us with a fresh perspective to concurrently investigate biological modifications and medical features through information exploration.

Genomic Landscape and Alterations Summary

Based on obtained outcomes, it was observed that the majority of the cases undergo alteration in the cell cycle control, and nearly all of them were missense mutations. Others incorporated deep deletions and few amplifications. However, the rest of the cases remained had modifications in the cell cycle control that comprises most of the truncating and missense mutations. The shared exclusivity analysis implies that events that occurred in cell cycle control were liable to occur again in HCC as shown in **Figure 1** through principal component analysis.

Expression in Cell Cycle Control in HCC

Inspection of the expression analysis reveals that there is a significant cell cycle control in under- and overexpressed

TABLE 1 | Biomarkers list of lncRNAs, miRNAs, mRNAs, methylations, and proteins in liver cancer with their estimated coefficient score.

ncRNAs	Biomarkers	Prediction Scores
lncRNAs	ENSG00000234577 1	−0.009981348
	ENSG00000236753 1	0.002278461
	ENSG00000257883 1	−0.000943499
	ENSG00000260920 1	0.423349833
	ENSG00000273059 1	−1.47647639
miRNAs	has-mir-139	−0.000141116
	has-mir-148a	−2.26E−07
	has-mir-23c	−0.00822489
	has-mir-3156-1	−0.076147329
	has-mir-3680	0.062526769
	has-mir-3682	0.001235262
	has-mir-3910-2	0.008941494
	has-mir-3924	−0.035235217
	has-mir-548v	0.006351925
	has-mir-592	0.000212203
	has-mir-632	0.107689301
mRNAs	ACSM4	0.212644291
	C15orf33	−0.004560181
	CAD	4.21E−05
	CDCA8	0.000189659
	CHAC2	0.000861851
	DYNC1LI1G6PD	5.19E−05
	G6PD	6.03E−06
	GOT2	−2.16E−06
	GRPEL2	0.000536562
	KLRB1	−8.62E−05
	LDHA	7.57E−07
	MYCN	0.000171789
	MYOD1	0.00620387
	NBPF22P	0.004514996
	PDE6A	0.014750229
	PSD4	−2.48E−05
	PSMD1	352E-05
	RPL21P44	0.00651285
	SEPHS1	0.000165664
	SLC1A7	2.08E−05
	SNAR-C4	0.101384012
	SNORD94	−0.009345186
	TAF3	0.00158546
	TMEM69	3.91E−05
	TTC26	0.000160117
	TXNRD1	1.10E−05
	UCK2	0.000240815
	UTP11L	0.00051883
	YARS	9.13E−06
	ZNF643	1.88E−05
Methylations	Cg02603140	0.0886803
	Cg04104820	−0.612239
	Cg15744128	−0.119538
	Cg24153171	−1.46113
Proteins	Acetyla-Tubulin-Lys40	0.041212138
	B-Raf	0.102643389
	PAI-1	0.114385869

HCC, explaining that these were hotspots for the activation as illustrated in **Figure 2**.

Survival Analysis

For the sake of survival rate inspection, Kaplan–Meier plots were used in an order to complete survival analysis in cases of HCC with as well as without cell cycle control overexpression. For the overall survival analysis, mutations in the cell cycle control were found to be concurrent and not linked to a decreased overall survival ($p = 0.0615$). Likewise, none of the cell cycle control was linked with any of the survival events (**Figure 3**).

Liver Cancer Prognosis Markers and Expression Landscape

It was observed that lncRNAs, miRNAs, mRNAs, methylations, and proteins may exert a more profound biological impact than a single gene by virtue of its intrinsic regulatory nature. Therefore, predictive modeling is also performed for the sake of liver cancer

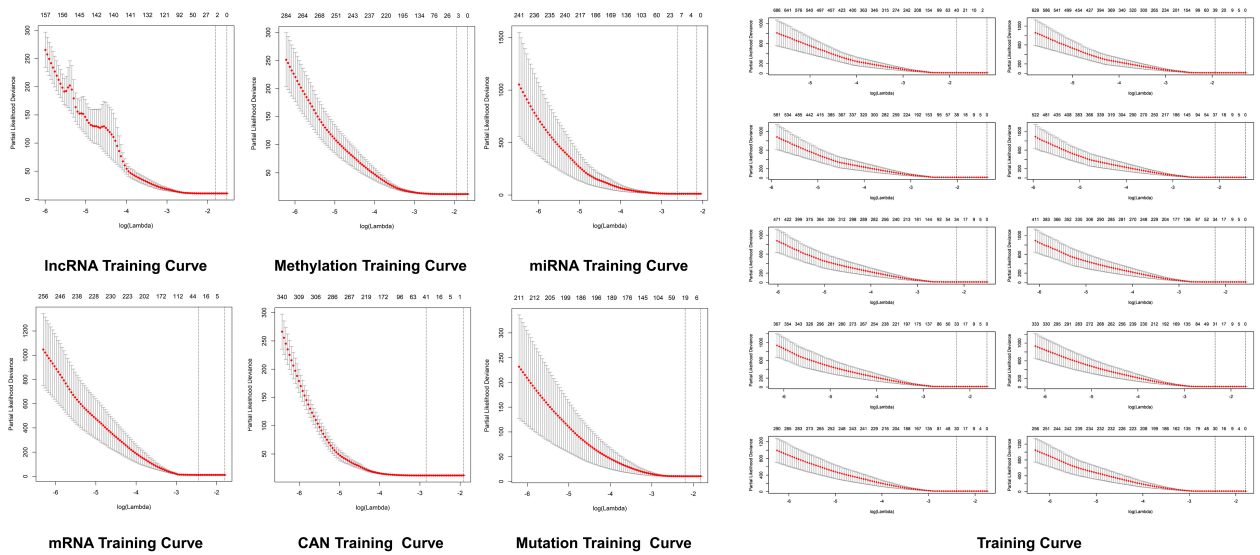


FIGURE 4 | Predictive modeling through LASSO. LASSO with the features weighted training curve in HCC, where optimization equation was resolved through LASSO for RNA implication based on features reputation generation, drilling LASSO with the features biased by reputation, which depend on bootstrapping approach by drilling only a small set of variables instead of drilling the whole variables directly as depicted in figure.

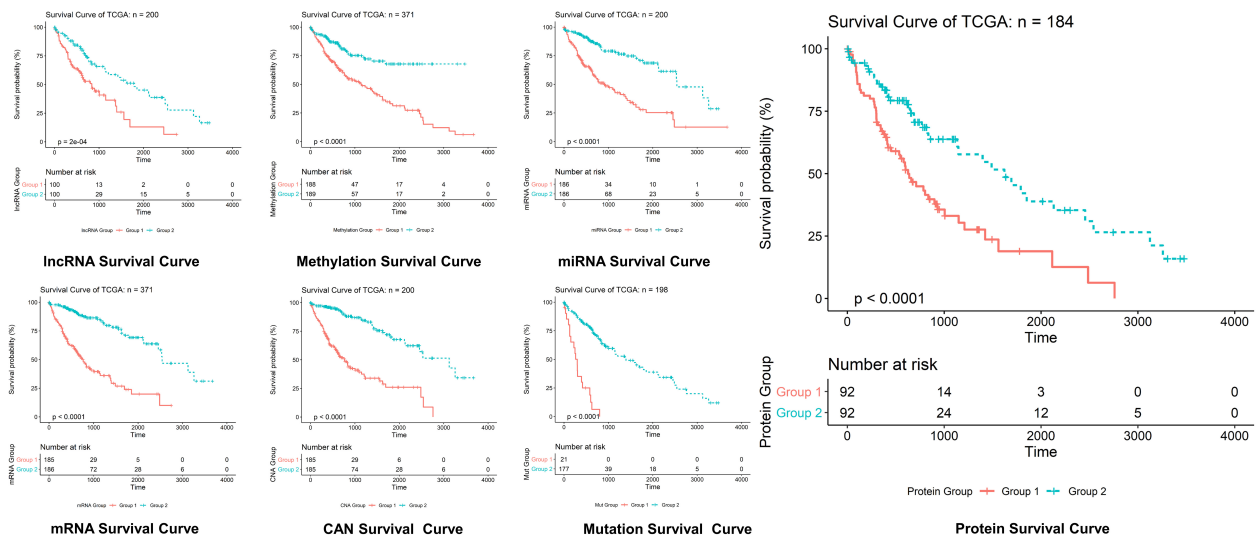


FIGURE 5 | Survival curves prediction through LASSO. Survival curves of lncRNAs, methylations, miRNAs, mRNAs, protein, CAN, and mutation in hepatocellular carcinoma, where optimization equation was resolved through LASSO for RNAs implication based on features reputation generation, drilling LASSO with the features biased by reputation, which depend on bootstrapping approach by drilling only a small set of variables instead of drilling the whole variables directly, and constant approximation is much dependable on every individual training through high-dimensional information as shown here.

prognosis (**Table 1**) and the expression landscape of survival patients as shown in **Figures 4**, **5**.

Predictive Modeling

The optimization of an equation can be resolved by a LASSO explanation. For reliable RNA implication, we applied the Random LASSO (Lee et al., 2014). This technique is divided into two main steps: (1) features' reputation generation for RNAs and (2) drilling LASSO with the features biased by reputation. This method uses a bootstrapping approach by drilling only a small set of variables instead of drilling the whole variables directly. The constant approximation is much dependable on every individual training through high-dimensional information (**Figures 4**, **5**).

Validation of Biomarkers

Based on literature reports, an initial major component assessment leads to separation. So, the first thing that we asked was whether the whole organization is practical by using a standardized and shared dataset. Hence, to evaluate our data for equality and applicability, a fivefold leave group cross-validation was employed by the use of LASSO and spare learning. All the datasets were examined distinctly, so each dataset was utilized as a drilling set along with group tags that were equitable to the general receptor position, via reference against liver cancer. Consequently, the receptor status of the entire untrained datasets was predicted by the use of attained Spares learning and LASSO model. Nevertheless, the accuracy of the categorization of the patient's data builds by the utilization of Spares learning and LASSO models is high. Reasonable Matthews correlation coefficients and low error rates were 0.1 for the grouping of references against liver cancer. The bootstrapping sampling was modified at the time of model formation for addressing inequity in the class in various datasets for drawing an equal number of samples out of each group.

CONCLUSION

Detection of cancer at an early curable phase and eradicating the tissues can be capable of preventing the expansion of lethal intrusive cancers, which would save countless lives. Presently, it is extensively stated that lncRNAs, mRNAs, and miRNAs could be probable biomarkers for various cancers. Identification of lncRNAs, mRNAs, and miRNAs related to disease adds to the enhancement of understanding of diagnosis and pathogenesis. Therefore, for the investigation of disease association of lncRNAs, mRNAs, and miRNAs, development of numerous potential computational models has been done. Nevertheless, only some studies centered on the identification of lncRNA, mRNA, and miRNA signs for the diagnosis of liver cancer in the early stage. Consequently, in the current study, we put forward a new classification technique based on lncRNAs, mRNAs, and miRNAs for the categorization of early and advanced phases of liver cancer. The increasing trend in the implementations of machine learning methods and the latest developments of personalized medicine enhanced the forecasting of cancer. For the sake of identifying main aspects that could influence cancer development, recurrence, and survival, different machine learning techniques and algorithms employed for feature selection are globally cast. In general, cancer prediction studies based on machine learning employed expression profiles of mRNA/miRNA, clinical factors, and histological variables as an input for the procedure of cancer prediction. Success in the development of computational models for prediction of cancer rests on comprehending the biological information and shortcomings of the drilling dataset, for instance, minor collection of high-dimensional samples known as "curse of dimensionality." Nevertheless, the over-drilling problem may be overcome through appropriate feature selection and cross-validation approaches. Our findings provide a new vision for exploring biological functions of lncRNAs, miRNAs, mRNAs, methylations, and proteins in liver cancer, and screening novel potential biomarker (lncRNAs, miRNAs, mRNAs, methylations, and proteins) signature could be a biomarker for the prognosis of liver cancer patients. Better performance toward liver cancer was shown by logistic LASSO regression descriptor where significant improvement was seen in predicting liver cancer.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

AK, XD, and D-QW designed the experiments. XD, D-QW, and AK performed the entire computational experiments and assisted in writing the manuscript. D-QW and AK analyzed the data and wrote the manuscript. AK, D-QW, XD, and AM read the manuscript and advised on method development. All authors approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 81972789), the National Science and Technology Major Project (Grant No. 2018ZX10302205-004-002), the Six Talent Peaks Project in Jiangsu Province (Grant No. SWYY-128), the Fundamental Research Funds for the Central Universities (Grant No. JUSRP22011), and Technology Development Funding of Wuxi (Grant No. WX18IVJN017).

REFERENCES

Alsaggar, M., and Liu, D. (2016). "Liver-targeted gene and cell therapies: an overview," in *Gene Therapy and Cell Therapy Through the Liver*, eds S. Terai and T. Suda (Tokyo: Springer), 1–11. doi: 10.1007/978-4-431-55666-4_1

Baker, J. A., Kornguth, P. J., Lo, J. Y., Williford, M. E., and Floyd, C. E. Jr. (1995). Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology* 196, 817–822. doi: 10.1148/radiology.196.3.7644649

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

Calderaro, J., Ziol, M., Paradis, V., and Zucman-Rossi, J. (2019). Molecular and histological correlations in liver cancer. *J. Hepatol.* 71, 616–630. doi: 10.1016/j.jhep.2019.06.001

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*. Philadelphia: AACR.

Chew, W. L., Tabebordbar, M., Cheng, J. K., Mali, P., Wu, E. Y., Ng, A. H., et al. (2016). A multifunctional AAV–CRISPR–Cas9 and its host response. *Nat. Methods* 13:868. doi: 10.1038/nmeth.3993

Chhatwal, J., Alagoz, O., Lindstrom, M. J., Kahn, C. E. Jr., Shaffer, K. A., and Burnside, E. S. (2009). A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *Am. J. Roentgenol.* 192, 1117–1127. doi: 10.2214/AJR.07.3345

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* 163, 506–519. doi: 10.1016/j.cell.2015.09.033

Dooley, J. S., Lok, A. S., Garcia-Tsao, G., and Pinzani, M. (2018). *Sherlock's Diseases of the Liver and Biliary System*. Hoboken, NJ: John Wiley & Sons.

Fang, B., Mehran, R. J., Heymach, J. V., and Swisher, S. G. (2015). Predictive biomarkers in precision medicine and drug development against lung cancer. *Chin. J. Cancer* 34, 295–309. doi: 10.1186/s40880-015-0028-4

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:pl1. doi: 10.1126/scisignal.2004088

Houssami, N., Irwig, L., Simpson, J. M., McKessar, M., Blome, S., and Noakes, J. (2004). The influence of clinical information on the accuracy of diagnostic mammography. *Breast Cancer Res. Treat.* 85, 223–228. doi: 10.1023/b:brea.0000025416.66632.84

Kanda, T., Goto, T., Hirotsu, Y., Moriyama, M., and Omata, M. (2019). Molecular mechanisms driving progression of liver cirrhosis towards hepatocellular carcinoma in chronic Hepatitis B and C infections: a review. *Int. J. Mol. Sci.* 20:1358. doi: 10.3390/ijms20061358

Karimian, A., Azizian, K., Parsian, H., Rafieian, S., Shafiei-Irannejad, V., Kheyrollah, M., et al. (2019). CRISPR/Cas9 technology as a potent molecular tool for gene therapy. *J. Cell. Physiol.* 234, 12267–12277. doi: 10.1002/jcp.27972

Lee, T. F., Chao, P. J., Ting, H. M., Chang, L., Huang, Y. J., Wu, J. M., et al. (2014). Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of xerostomia after intensity-modulated radiotherapy for head and neck cancer. *PLoS One* 9:e89700. doi: 10.1371/journal.pone.0089700

Li, B., and Li, J. Z. (2014). A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.* 15:473. doi: 10.1186/s13059-014-0473-4

Li, B., Senbabaoglu, Y., Peng, W., Yang, M.-L., Xu, J., and Li, J. Z. (2012). Genomic estimates of aneuploid content in glioblastoma multiforme and improved classification. *Clin. Cancer Res.* 18, 5595–5605. doi: 10.1158/1078-0432.CCR-12-1427

Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., et al. (2015). TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.* 75, 3728–3737. doi: 10.1158/0008-5472.CAN-15-0273

Lu, Y., Min, Z., Qin, A., Wu, J., Jiang, X., and Qiao, Z. (2020). Role of miR-18a and miR-25 disruption and its mechanistic pattern in progression of liver cancer. *3 Biotech* 10:74.

von Felden, J., and Villanueva, A. (2020). Role of molecular biomarkers in liver transplantation for hepatocellular carcinoma. *Liver Transpl.* doi: 10.1002/lt.25731

West, C. A., Black, A. P., and Mehta, A. S. (2019). "Analysis of hepatocellular carcinoma tissue for biomarker discovery," in *Hepatocellular Carcinoma*, ed. Y. Hoshida (Cham: Humana Press), 93–107. doi: 10.1007/978-3-030-21540-8_5

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

Zheng, Y., Shi, Y., Yu, S., Han, Y., Kang, K., Xu, H., et al. (2019). GTSE1, CDC20, PCNA, and MCM6 synergistically affect regulations in cell cycle and indicate poor prognosis in liver cancer. *Anal. Cell. Pathol.* 2019:1038069.

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 © 2020 Kaushik, Mehmood, Wei and Dai. 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 Genetics

ORIGINAL RESEARCH
published: 07 March 2020
doi: 10.3389/fgene.2020.00230

Effect for Human Genomic Variation During the BMP4-Induced Conversion From Pluripotent Stem Cells to Trophoblast

Hai-tao Li[1†], Yajun Liu[1,2,3†], Hongde Liu[1] and Xiao Sun[1*]*

[1] State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China, [2] The Second Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [3] Academy of Medical Sciences of Zhengzhou University Translational Medicine Platform, Zhengzhou University, Zhengzhou, China

The role of genomic variation in differentiation is currently not well understood. Here, the genomic variations were determined with the whole-genome sequencing for three pairs of pluripotent stem cell lines and their corresponding BMP4-induced trophoblast cell lines. We identified ∼3,500 single nucleotide variations and ∼4,500 indels by comparing the genome sequenced data between the stem cell lines and the matched BMP4-induced trophoblast cell lines and annotated them by integrating the epigenomic and transcriptomic datasets. Relatively, introns enrich more variations. We found ∼45% (42 genes) of the differentially expressed genes in trophoblasts that associate genomic variations. Six variations, located at transcription factor binding sites where H3K4me3 and H3K27ac are enriched in both H1 and H1_BMP4, were identified. The epigenetic status around the genomic variations in H1 was similar to that in H1_BMP4. This means that the variation-associated gene's expression change can not be attributed to epigenetic alteration. The genes associated with the six variations were upregulated in differentiation. We inferred that during the differentiation, an increased in the expression level of the MEF2C gene is due to a genomic variation in chromosomes 5: 88179358 A > G, which is at a binding site of TFs KLF16, NR2C2, and ZNF740 to MEF2C. Allele G shows a higher affinity to the TFs in the induced cells. The increased expression of MEF2C leads to an increased expression of TF MEF2C's target genes, subsequently affecting the differentiation. Although genomic variation should not be a dominant factor in differentiation, we believe that genomic variation could indeed play a role in the differentiation from stem cells into trophoblast.

Keywords: genomic variations, pluripotent stem cells, trophoblast, whole genome sequencing, epigenomic and transcriptomic data

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Wei Jiang,
Nanjing University of Aeronautics and Astronautics, China
Ruiping Wang,
The University of Texas MD Anderson Cancer Center, United States

****Correspondence:***
Hongde Liu
liuhongde@seu.edu.cn
Xiao Sun
xsun@seu.edu.cn

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

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *07 December 2019*
Accepted: *26 February 2020*
Published: *07 March 2020*

Citation:
Li H, Liu Y, Liu H and Sun X (2020) Effect for Human Genomic Variation During the BMP4-Induced Conversion From Pluripotent Stem Cells to Trophoblast. Front. Genet. 11:230. doi: 10.3389/fgene.2020.00230

INTRODUCTION

Stem cell differentiation involves a complex but poorly understood biological process. Genetic and epigenetic factors have in the past been intensively studied for this process. Some key transcription factors (TFs) affecting differentiation have been identified, such as POU5F1, SOX2, and KLF4 (Nichols et al., 1998; Takahashi and Yamanaka, 2006; Goolam et al., 2016). Previously, our lab

identified TFs for DNase I hypersensitive sites based on public available chromatin accessibility data of human embryonic stem cells (ESCs) and BMP4-induced trophoblast cell lines (Liu et al., 2017). The chromatin structure, including nucleosome positions, DNA methylation, and histone modifications also changed during the differentiation (Su et al., 2012; Xie et al., 2013; de Boni et al., 2018). However, genomic variation, another important factor, was rarely considered in the differentiation-related study.

The cells forming the outer layer of a blastocyst in early development are referred to as trophoblasts. These cells significantly contribute to the placenta and membrane, and also provides nutrients to the embryo. This layer of trophoblasts is collectively named "the trophoblast." The trophoblast is the first cell type differentiated from the zygote during the first stage of pregnancy. Human ESCs provide a very useful model for studying the early development of human embryos and trophoblasts. However, the genomic variation that occurs in the trophoblast differentiation is not well studied *in vivo*. Since Xu et al. proved that bone morphogenetic protein 4 (BMP4) could induce human ESCs, efficiently differentiating to trophoblast lineage, multiple research institutes have adopted this system as a model to study trophoblast lineage specification *in vitro* (Xu et al., 2002). Roberts's et al. considered this model reliability by analyzing gene expression profiling through RNA sequencing (RNA-Seq) technology (Yabe et al., 2016; Liu et al., 2017).

Recently, studies have demonstrated how genetic variation affects gene expression (Banovich et al., 2018; Delahaye et al., 2018). Kilpinen et al. (2013) suggest that genetic variation appears to be the primary driver of gene expression variation. Furthermore, DeBoever et al. (2017) show that the vast majority of genetic variations associated with gene expression levels are located in the regulatory regions of human induced pluripotent stem cells (iPSCs).

An integrative analysis using combinations of genomic, epigenomic, and transcriptomic data will provide a basis for biomarker discovery and help to provide insight of disease etiology. Genomic sequence variation, epigenetic factor, and gene expression are interdependent and jointly contribute to the normal functioning or dysfunction of tissue (Delahaye et al., 2018). For example, the sequencing variations can alter the TF binding strength to regulate gene expression directly (Karczewski et al., 2011; Madsen et al., 2018; Johnston et al., 2019). We are interested in the role of the genome variations in differentiation from human ESCs to trophoblasts.

In this study, we analyzed the genomic DNA sequences for three paired human ESCs and iPSC (H1, H9, and MRucR) (Sheridan et al., 2019). We demonstrated that some of the genomic variations can affect the differentiation by altering TFs' binding affinity.

MATERIALS AND METHODS

Whole Genome Sequence of Cell Lines in This Study

Genomic DNA sequences were determined for three stem cells and their corresponding BMP4-induced cell lines (**Supplementary Table S1**), in which two types of stem cells were human ESC lines (H1 and H9), and third an iPSC cell line, MRucR. The cell lines that differentiated into the trophoblast were named H1_BMP4, H9_BMP4, and MRucR_BMP4, respectively. The three cell lines and the matched BMP4 induced cell lines was obtained from the Roberts's laboratory, University of Missouri. For more details of these three pairs of cell lines see Sheridan et al. (2019). MRucR iPSC was established from umbilical cords of babies born to mothers who experienced an early-onset form of preeclampsia during their pregnancies. The BMP4-inducing experiment was performed as previously described (Sheridan et al., 2019). Briefly, the trophoblast stem cells were exposed to BMP4 in combination with signaling inhibitors of ACTIVIN-A (A83-01) and FGF2 (PD173074) (BAP treatment) (Sheridan et al., 2019).

Library Preparation and Sequencing

The quality of isolated genomic DNA was verified using these three methods: (1) DNA purity and concentration were identified by NanoPhotometer® spectrophotometer (IMPLEN, CA, United States) (OD260/OD280). The Optical Density (OD) value of the qualified sample ranged between 1.8 – 2.0. (2) DNA degradation, and suspected RNA/Protein contamination were verified by electrophoresis on 1% agarose gels. (3) The concentration and purity of DNA samples were further quantified precisely by the Qubit DNA Assay Kit in Qubit®2.0 Flurometer (Life Technologies, CA, United States). A total amount of 1 μg DNA per sample was required for library generation.

Paired-end DNA libraries were prepared according to the manufacturer's instructions (Illumina Truseq Library Construction). First, 1.0 μg Genomic DNA was sheared into an average size of 350 base pair (bp) fragments by Covaris S220 sonicator. Second, the ends of the gDNA fragments were repaired; 3′ ends were adenylated. Both ends of the gDNA fragments were ligated at the 3′ ends with paired-end adaptors (Illumina) with a single 'T' base overhang, and purified using AMPure SPRI beads from Agencourt. The size distribution and concentration of the libraries were then determined using Agilent 2100 Bioanalyzer and were qualified by real-time PCR (2 nM), respectively. Lastly, DNA libraries were sequenced on Illumina Hiseq X according to the manufacturer's instructions for paired-end 150 bp reads.

Whole Genome Sequencing Data Analysis

The raw image files obtained from the Hi-Seq platform were processed with the Illumina pipeline for base calling and stored as FASTQ format (Raw data). Quality control (QC) was as follows: (1) to filter reads with adapter contamination (>10 bp aligned to the adapter allowing ≤ 10% mismatches). (2) to discard the reads containing more than 10% uncertain nucleotides. (3) to discard the paired reads when a single read has more than 50% low quality nucleotides (Phred quality < 5).

After quality control, the sequenced reads were mapped to the GRCh37 assembly of the human genome by Burrows-Wheeler Aligner (BWA) software using default settings (Li

and Durbin, 2009). Subsequently, we used Samtools (Li et al., 2009) and Picard[1] with default settings to sort reads, remove duplicated reads, and to generate the final bam file. If one or more pair read(s) had multiple mapping positions, the best one was selected. If there were multiple best mapping positions, we randomly picked one.

Variation Calling and Functional Annotation

The analysis flowchart is shown in **Figure 1A**. The genomic variations were determined with the whole-genome sequencing for three pairs of pluripotent stem cell lines and their corresponding BMP4-induced trophoblast cell lines and annotated them by integrating the epigenomic and transcriptomic datasets. We only focused on the direct effect on the differentiation of genomic variation which is located in those genomic regions where epigenetic markers remain comparable between H1 and H1_BMP4.

Genomic variations were identified by comparing genomic DNA sequences between stem cell lines and matched BMP4-induced trophoblast cell lines using the following steps. Single nucleotide variation (SNV) and small somatic insertions and deletions (indels) were identified using the Strelka2 (Kim et al., 2018) tool. All 'PASS' calls identified by Strelka2 were retained for downstream analyses. BreakDancer (Chen et al., 2009) was applied to detect structural variation (SV). The SVs that received minimal confidence scores (90 for insertions, inversions, deletions, and translocations) were selected for downstream analyses.

The ANNOVAR tool was used to produce statistical analyses of the SNVs/indels (Wang et al., 2010). The variation position, variation type, conservative prediction, and other information were obtained at this step through a variety of databases, including DbSNP, 1000 Genome, and the reference sequence (Sherry et al., 2001; Pruitt et al., 2007; The 1000 Genomes Project Consortium, 2010). The variations were then assigned to genes by associating them with the nearest transcription start sites (TSSs) using the BEDOPS toolkit (Neph et al., 2012) and Gencode v21 human annotation (Harrow et al., 2012).

The regions include exon, intron, intergenic region, UTR5, UTR3, and the upstream (variant overlaps 1-kb region upstream of transcription start site)/downstream (variant overlaps 1-kb region downstream of transcription end site) regions.

The annotation detail of these three cell lines is listed in **Supplementary Excel File 1**. The version of the bioinformatics tools used is listed in **Supplementary Table S2**.

Public Datasets Analyzed in This Study

The goal of the work is to assess the effect of the genomic variations during stem cell differentiation. We therefore need to exclude the effect of the epigenomic variation (**Figure 1A**). We thus combined DNA methylation, histone modifications, chromatin accessibility, and transcriptomic datasets for the H1 and H1_BMP4 (differentiated with BAP treatment) cell lines. All the ChIP-Seq and DNase-Seq data in H1 and H1_BMP4 were generated by the ENCODE Consortium[2] and retrieved from the GEO database according to their accession number (**Supplementary Table S3**) (Lister et al., 2011; Roadmap Epigenomics Consortium et al., 2015; Davis et al., 2018). The sequencing data were mapped to the human reference genome GRCH37/hg19 by BWA. For ChIP-Seq data, we performed peak calling by using the MACS2 tool with default settings (Zhang et al., 2008). Differential peaks between the two cell lines were also identified with MACS2. We are interested in the genomic variations where there is no significant change between the two matched cell lines in the epigenomic data, especially for histone marks H3K27ac and H3K4me3 since the two represent the active state for enhancers and promoters.

The methylation state at CpG sites in whole genome bisulfite sequencing (WGBS) data were mapped to GRCh38/hg38 (**Supplementary Table S3**) (Roadmap Epigenomics Consortium et al., 2015). We used the liftOver tool to transform the genome coordinates from hg38 to hg19 (Dreszer et al., 2012).

The gene expression data (Xie et al., 2013) were retrieved from the ENCODE project (**Supplementary Table S3**) with the following identifiers: ENCFF245NXP, ENCFF809MAC, ENCFF051SBM, and ENCFF787HWI. RNA abundance was represented as the logarithm of the transcripts per million (TPM) provided by the RSEM program (Li and Dewey, 2011). The data was used in two aspects. One to identify the differentially expressed genes, so as to check which genes are influenced by the genomic variations. The other is to find the transcription factors that truly have a function in cells. This was done by checking the expression level of the gene that encodes the transcription factor (**Figure 1A**). Differentially expressed mRNAs were identified with limma (Ritchie et al., 2015). Genes with Fold Change > 2 or $<1/2$ and false discovery rate (FDR) p-value < 0.05 were identified as significantly differentially expressed genes between ESC H1 and differentiated to trophoblast H1_BMP4.

Prediction for Transcription Factor Binding Sites (TFBSs) Around the Genomic Variations

To estimate the change of the TF binding affinity due to the genomic variation, the 150-bp DNA sequences were extracted around the genomic variations and inputted into a bioinformatics tool, HOMER (Heinz et al., 2010), to calculate the affinity (**Figure 1A**). Each of the sequences includes one kind of genomic variation. The affinity of a TF to a specific DNA sequence can be estimated by comparing the DNA sequence and the motif of the TF. The motif is the most favorable binding DNA pattern of the TF and can be represented with Position Weight Matrix (PWM). The comparison will be a score on the motif (motif score). A high score means a high affinity between the DNA sequence and the TF. We listed the TFs with a motif score ≥ 10 in the stem cell and its BMP4-induced cells and found the difference of the TFs between the cells.

[1] http://broadinstitute.github.io/picard

[2] https://www.encodeproject.org/

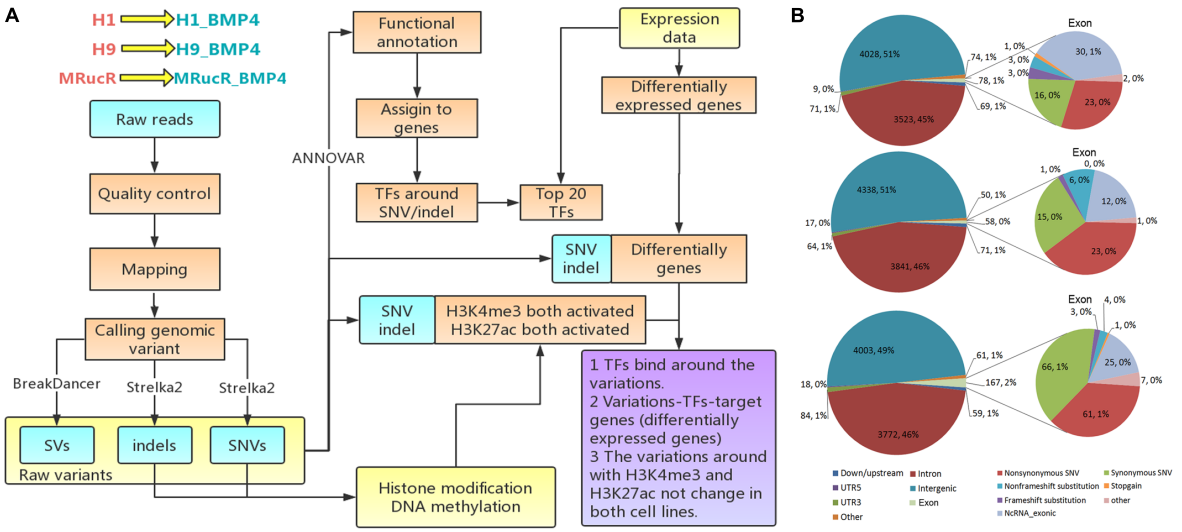


FIGURE 1 | Genomic variations during the differentiation. **(A)** Flow chart of the analysis. **(B)** The variations occurred in different genomic regions. In this figure, down/upstream means the variant overlaps the 1-kb region downstream of transcription end site or upstream of TSS.

RESULTS

The Distribution of Genomic Variations During the Differentiation

We performed the high-throughput DNA sequencing for three pluripotent stem cell lines (H1, H9, and MRucR) and differentiated trophoblasts (H1_BMP4, H9_BMP4, and MRucR_BMP4). After quality control (**Supplementary Figure S1**), genomic variations were identified by comparing the DNA sequence between the stem cell lines and matched BMP4-induced trophoblast cell lines. Reads mapping rate was more than 98%, and the sequencing depth was beyond 31X (**Supplementary Tables S4, S5**). We used Strelka2 to identify SNV and indels. SV was identified with BreakDancer (Chen et al., 2009; Kim et al., 2018). The Flow chart of our analysis is shown in **Figure 1A**. We identified approximately 3,500 SNVs and 4,500 indels in the ESC and iPSC lines, respectively (**Table 1A**). The count of SV was relatively rare (∼230) (**Table 1A**).

The frequency of SNVs was counted. According to our statistics, the single-base substitution patterns were similar in the three pluripotent stem cells (**Supplementary Figure S2a**). The highest frequency of SNV was the conversion of cytosine to thymine (C > T), and the lowest frequency of SNV was the conversion of guanine to cytosine (G > C). Among the SVs (**Table 1B**), the proportion of translocation, including inter- and intra-, was high (about 77%∼87%). The percentage of insertion inversion and deletion was 12% ∼ 22%. We did not detect the structure deletion events here. SV had similar distributions in these three pairs of cell lines (**Figure 1B**).

We then annotated SNV and indels with ANNOVAR (Wang et al., 2010). The SNV and indels were counted in different genomic regions (**Supplementary Excel File 1**). To calculate fold enrichments, ratios of the variation counts in a certain category of genomic regions were divided by the proportions of this category in the whole genome length. The variations were significantly enriched in the introns (**Figure 2**).

Distribution profiles of SNV/indels in the 2500-bp genomic regions around the TSSs seemly occurred in a periodic pattern in the cell lines (**Supplementary Figure S2b**). SNP frequency spectra show striking periodicities across nucleosomal regions, and SNPs have a preference for nucleosomes (Langley et al., 2014). We therefore speculated that there is an association between the patterns and nucleosome distribution in this region.

Transcription Factors That Bind Around Genomic Variations Sites

A genomic variation may directly alter the affinity of TF's binding as this variation occurs exactly at TFBSs, thus influencing the

TABLE 1 | Number of genomic variants **(A)** and structural variants **(B)** in which stem cells differentiate into trophoblast cells.

(A)

Matched hESC-trophoblast cell lines	Indels	SNV	SV
H1	4071	3781	233
H9	4703	3736	235
MRucR	4729	3435	225

(B)

Matched hESC-trophoblast cell lines	INS	INV	ITX	CTX	DEL
H1	31	2	92	108	0
H9	19	10	96	110	0
MRucR	44	6	84	91	0

Indels, small somatic insertions and deletions; SNV, single nucleotide variant; SV, structural variation; INS, insertion; INV, inversion; ITX, intra-chromosomal translocation; CTX, inter-chromosomal translocation; DEL, deletions.

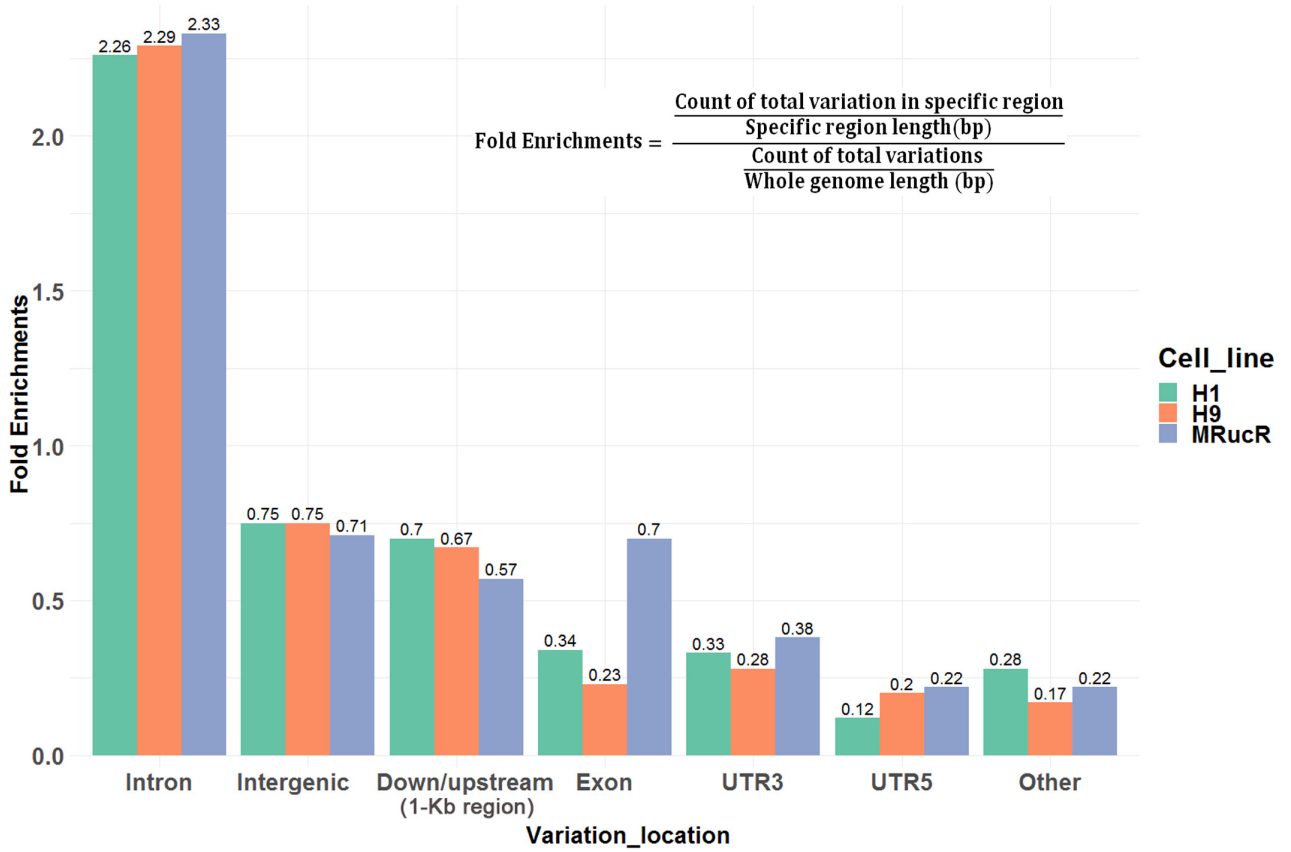


FIGURE 2 | Ratios of the number of variations occurred in certain categories of the genomic regions. Fold enrichments, ratios of the variation counts in a certain category of genomic regions were divided by the proportions of this category in the whole genome length. Down/upstream means the variant overlaps the 1-kb region downstream of the transcription end site or upstream of TSS.

transcription levels of the downstream target genes of the TFs. We assessed the effect of the variations on the alteration of TF binding with HOMER (Heinz et al., 2010). The 150-bp DNA sequences around the variation sites were inputted into HOMER to identify motifs of the TFs and compared the affinity of the TFs variation between the stem cell and matched BMP4-induced trophoblast cell lines. In **Figures 3A,B**, although counts of TF motifs around SNV/indels positions decreased, the number was not zero. There were some motifs that were far away from the SNV/indels. We intend to find out what TFs bind at the SNV/indels-harbored sites.

A TF with its motif in one genomic region does not mean the TF indeed binds at the region in a cell because the TF maybe is not expressed in the cell. We therefore need to check if the TF is abundant enough in the cell, namely, to check if the expression level of the gene encoding the TF is high enough. To do this, a necessary step is to exclude those TFs with a low abundance. The expression of TF-encoding genes was shown for H1 and H1_BMP4 (**Supplementary Figure S3**). Here, the cutoff of TPM $\geq$ 3 was used to select the TFs that indeed exist in the cell. We identified 154 and 193 TFs in H1 and H1_BMP4 cells, respectively, with an overlap of 135 TFs (**Figure 3C**). The top 20 key TFs were chosen by setting TPM $\geq$ 3 and motif score $\geq$ 10 (**Figures 3D,E** and **Supplementary Tables S6, S7**). In the top 20 TFs of H1 and H1_BMP4, some widely known to be involved in placental development were found, such as GATA3 (Kubaczka et al., 2015), POU5F1 (Wang et al., 2012), and KLF4 (Abad et al., 2013). A target gene of SREBP1A is the transcriptional repressor BHLHB2, which also promotes the differentiation of stem cells to trophoblast giant cells (Lecomte et al., 2010). ZEB2 has recently been identified to play critical roles in the regulation of the epithelial-mesenchymal transition and trophoblast differentiation (DaSilva-Arnold et al., 2019).

We identified the TFs that differ in affinity around the genomic variation sites between H1 and H1_BMP4. The TFs were chosen under conditions of TPM $\geq$ 3, motif score $\geq$ 10, and motif score count difference (abs) $\geq$ 10. We found three TFs (Zfp281, OCTs, and KLF3), whose affinities near the genomic variations differ between H1 and H1_BMP4 cell lines. The three kinds of TFs have a higher motif score count (count of motif score $\geq$ 10) in H1 than in H1_BMP4. Interestingly, these TFs are widely known to be involved in trophoblast differentiation. For example, Zfp281 (Krüppel-like zinc finger transcription factor), a zinc finger transcription factor, which shapes the transcriptome of trophoblast cells and regulates early placental development, has also been investigated in a recent article (Ishiuchi et al., 2019).

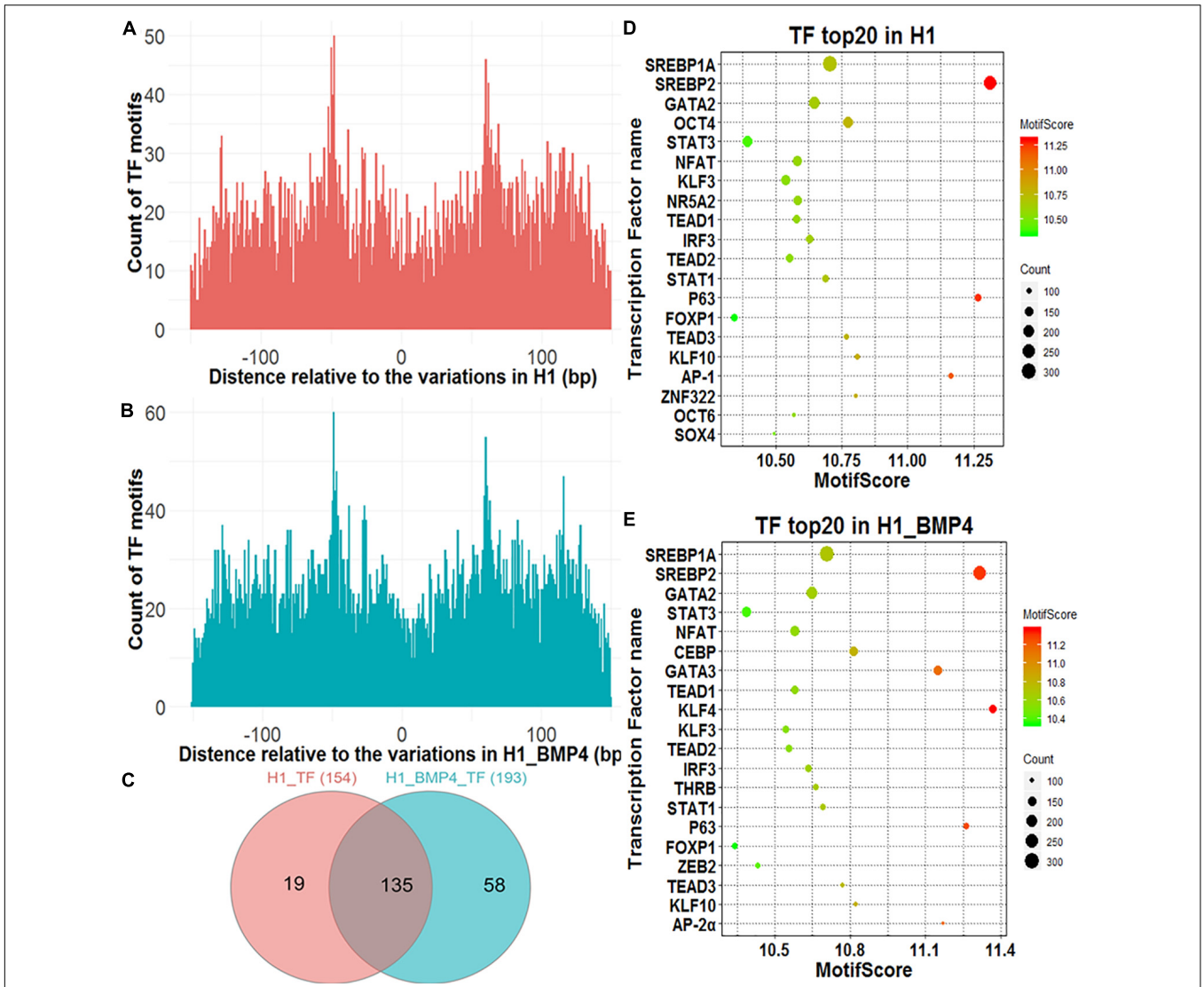


FIGURE 3 | Transcription factors identified around genomic variations sites. **(A)** The Venn diagram describing the overlap between the TFs in H1 and H1_BMP4. **(B,C)** The count of the TF motif around SNV/indels in H1 and H1_BMP4. **(D,E)** The bubble chart of the top 20 key TFs in H1 and H1_BMP4.

Moreover, the other TFs were widely known to be important in placental development (**Supplementary Table S8**).

In short, we found that there were indeed some genomic variations at TFBSs.

Correlation Between Genome Variations and Epigenomics

Genome variation was reported to be correlated to epigenetic alteration. Li et al. demonstrated that about 2/3 of eQTLs were due to variations that altered chromatin accessibility or histone marking (Li et al., 2016). We studied the correlation between genetic variations and chromatin alteration in H1 and H1_BMP4 cell lines. The sequencing data of H3K4me3, H3K27ac, H3K27me3, and DNA methylation data were retrieved from ENCODE (**Supplementary Table S5**).

In this analysis, the genomic variations were divided into four categories according to whether the variation site was at the peaks of the histone modifications and DNA methylation in H1 and H1_BMP4 (Lister et al., 2011; Roadmap Epigenomics Consortium et al., 2015). We counted the peaks of the histone modifications and DNA methylation within the 4K-bp genomic region around SNV/indel sites (**Figure 4** and **Supplementary Figure S4**).

The enriching regions (peaks) of DNA methylation and H3K27me3 were considered as "inhibited" regions. Modified regions (peaks) by H3K4me3 and H3K27ac indicate "promoter activated" and "enhancer inactivated" regions, respectively. By comparing the epigenomic states of the regions around the genomic variations between the stem cells and matched BMP4-induced trophoblast cell lines, 66 variations that do not associate an epigenomic alteration were identified (**Table 2**, and

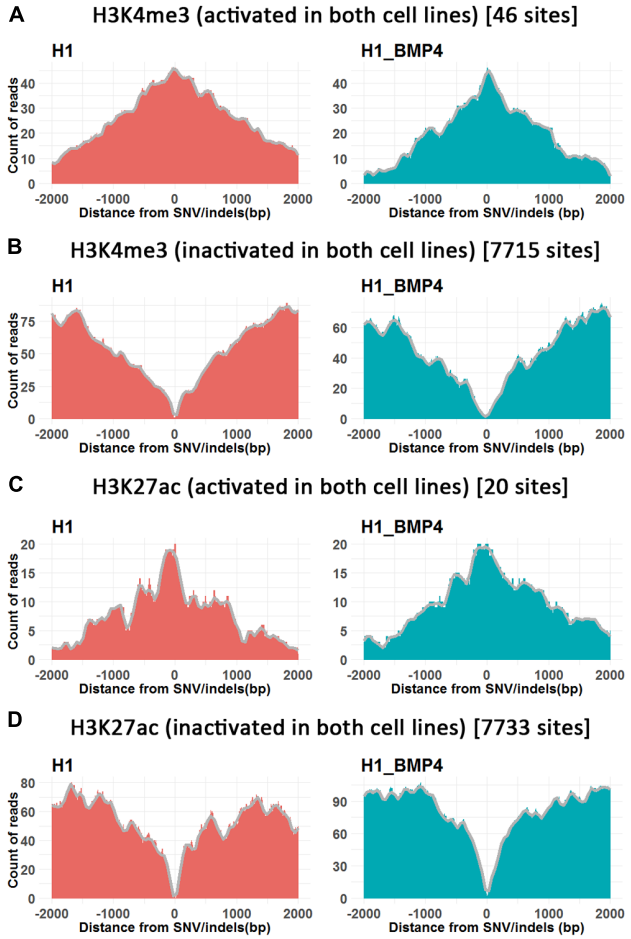


FIGURE 4 | The histone modifications surrounding SNV/indels. *X*-axis is the upstream and downstream 2 Kbp relative to SNV/indels sites. **(A–B)** The number of peak counts on the genomic variations which are located in both promoter activated **(A)** or inactivated **(B)** around H3K4me3 between H1 and H1_BMP4. **(C–D)** The number of peak counts on the genomic variations which are located in both enhancer activated **(C)** or inactivated **(D)** around H3K27ac between H1 and H1_BMP4. The number in parentheses indicates the number of SNV/indels. Complete profiles are shown in **Supplementary Figure S4**.

Supplementary Excel File 2). Next, we evaluated the effects of the genomic variation around which the epigenetic modifications do not alter between H1 and H1_BMP4.

The Direct Effect of Genomic Variation on Differentiation

We are interested in the gene expression change that is only caused by the genomic variation instead of the epigenetic variation. We therefore only focused on the genomic variations which are located in those genomic regions where epigenetic markers remain comparable between H1 and H1_BMP4.

We identified 5,688 differentially expressed genes between H1 and H1_BMP4 (FDR p-value ≤ 0.05 and fold change ≥ 2 or $\leq 1/2$) (**Figure 5A**). In the genes, there are 3,763 up-regulated and 1,925 down-regulated genes. Since each genomic variation was assigned to a gene, we could identify the genomic variations that associate differentially expressed genes.

TABLE 2 | Number of the genomic variations that occur in regions associating different switches of the epigenomic modifications between H1 and H1_BMP4.

Regions	Modification types	State switch from H1 to H1_BMP4	The number of genomic variations
–	H3K27me3 and DNA methylation	In - > T	0
–	H3K27me3 and DNA methylation	T - > In	0
Enhancer	H3K27ac	I - > A	18
Enhancer	H3K27ac	A - > I	12
Promoter	H3K4me3	I - > A	30
Promoter	H3K4me3	A - > I	30
Promoter	H3K4me3	A - > A	46
Enhancer	H3K27ac	A - > A	20

In, inhibited; T, transcribed; I, inactivated; A, activated.

One-hundred-and-twenty-nine differential expression genes were determined using stricter criteria (p-value $\leq$ 1E–5 and fold change $\geq$ 4 or $\leq$1/4). We found a total of 26 genes with variations in the differentiation process out of the 129 differential expression genes (**Figure 5B**). The functional annotation tool of KEGG was applied to the 26 genes (**Figure 5C**). As expected, two pathways implicated in regulating human ESC differentiation was found with an adjusted p-value < 0.05, namely the "signaling pathways regulating pluripotency of stem cells" and the "TGF-beta signaling pathway," which represent the differentiation pathways. For instance, inhibition of the TGF-beta signaling pathway could be sufficient for the derivation and long term expansion of human trophoblast cells and the placenta (Xu et al., 2017; Knöfler et al., 2019).

In the induction by BMP4, BMP ligands are bound to the BMP4 complex, then transphosphorylate the intracellular signaling proteins, Smad1/5/8. The phospho-Smad1/5/8 interacts with Smad4, and the complex translocates into the nucleus where it interacts with transcription cofactors and regulates expression of target genes in a cell type-specific manner (Shimasaki et al., 2004). Based on these biological processes, 235 target genes of Smad 1/4/5/6 were retrieved from the TF targets in the Harmonizome database (Rouillard et al., 2016). Among them, 94 are differentially expressed between H1 and H1_BMP4 (FDR p-value ≤ 0.05 and fold change ≥ 2 or $\leq$1/2). There are 72 up-regulated (30.64%) and 22 down-regulated genes (9.36%). In the 94 BMP4-induced genes, 42 (44.68%) contain the genomic variations (**Figure 5D**), suggesting a tight association between the variation and BMP4 induced differentiation.

In particular, by excluding the variations associating the epigenetic modification alteration, six genomic variations, which associate differentially expressed genes, were identified (**Table 3**, and **Supplementary Excel File 3**). Five of the six variations were located in the promoter regions of genes MEF2C, TNFAIP8,

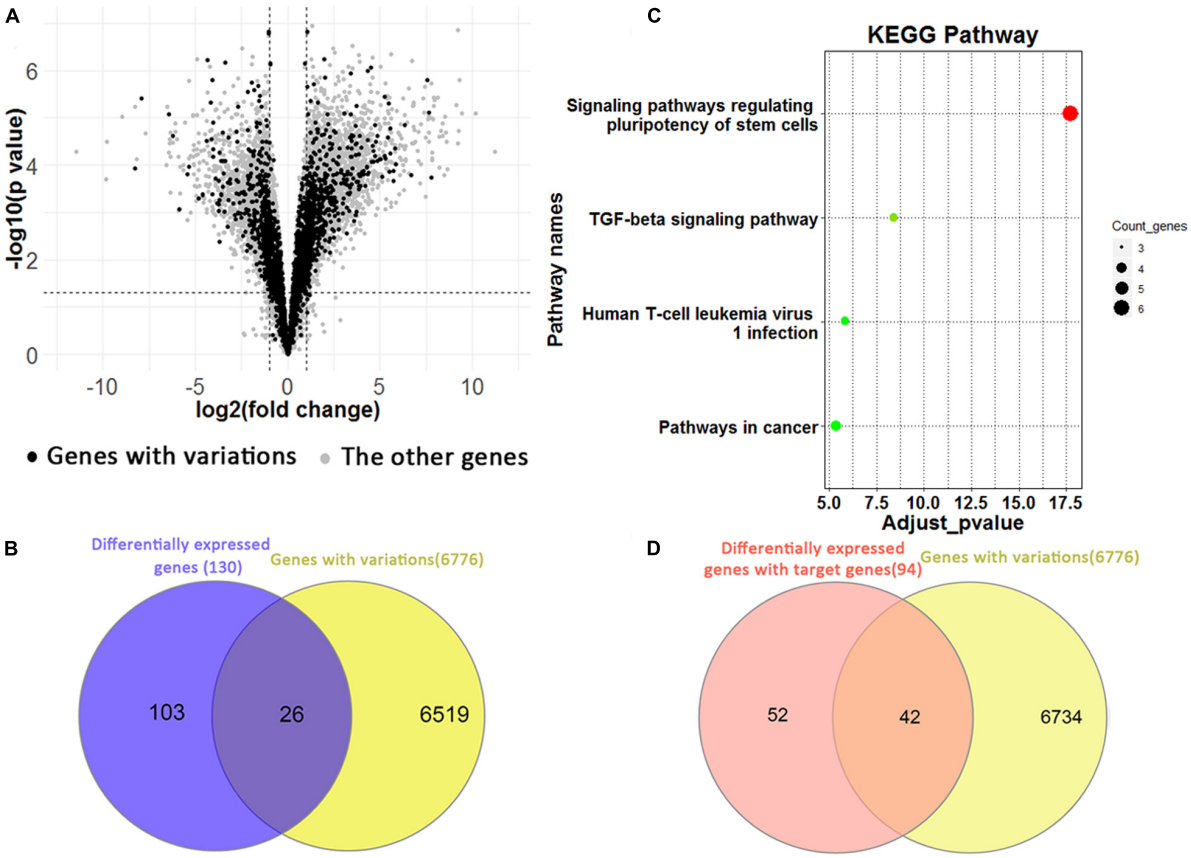


FIGURE 5 | (A) Volcano plotting shows the differentially expressed genes between human ESC and BMP4-induced trophoblast cell lines. The black points are for the genes with a genomic variation, while the gray points are for the other genes. **(B)** The Venn plotting shows the genes both with a genomic variation and those exhibiting a differential expression from H1 to H1_BMP4. **(C)** The significant KEGG pathway for 26 genes. **(D)** The Venn plotting shows the genes both containing variations, and 94 differentially expressed target genes, which were regulated by downstream signaling proteins of BMP4.

TABLE 3 | Information of six genomic variations.

Chr	Start	Ref	Alt	Location	Gene.refGene	Modification types
chr5	118605129	ATG	A	Intronic	TNFAIP8	H3K4me3
chr5	88179358	A	G	ncRNA_intronic	MEF2C	H3K4me3
chr17	62339443	G	T	Intronic	TEX2	H3K4me3
chr7	41740828	G	A	ncRNA_intronic	INHBA-AS1	H3K4me3
chr1	20811135	T	C	Intronic	CAMK2N1	H3K4me3
chr7	94023911	A	G	UTR5	COL1A2	H3K27ac

Column Location lists the genomic region of the variation. ncRNA_intronic means non-coding RNA genes. Gene.refGene is the gene associating the variation. Modification types indicate the H3K4me3 or H3K27ac around the variation in both H1 and H1_BMP4.

TABLE 4 | The TF motif scores of MEF2C.

TF_Name	Start	End	Strand	H1 Expression(TPM)	H1_BMP4 Expression (TPM)	H1_BMP4_motif score	H1_motif score	motif score_diff
KLF16	11	21	–	14	15.52	14.81	10.90	3.90
ZNF740	7	16	–	5.4	7.29	12.63	7.09	5.54
ZNF740	6	15	–	5.4	7.29	10.66	5.53	5.13
NR2C2	10	24	+	9.27	12.99	10.09	3.87	6.22

The TFs in this table were chosen as follows: (1) the motif scores of the TFs larger than 10 in H1 or H1_BMP4; (2) TPM value of the TF larger than 3 in both H1 and H1_BMP4; (3) the fold change of the expressed TF genes between H1 and H1_BMP4 less than 2.

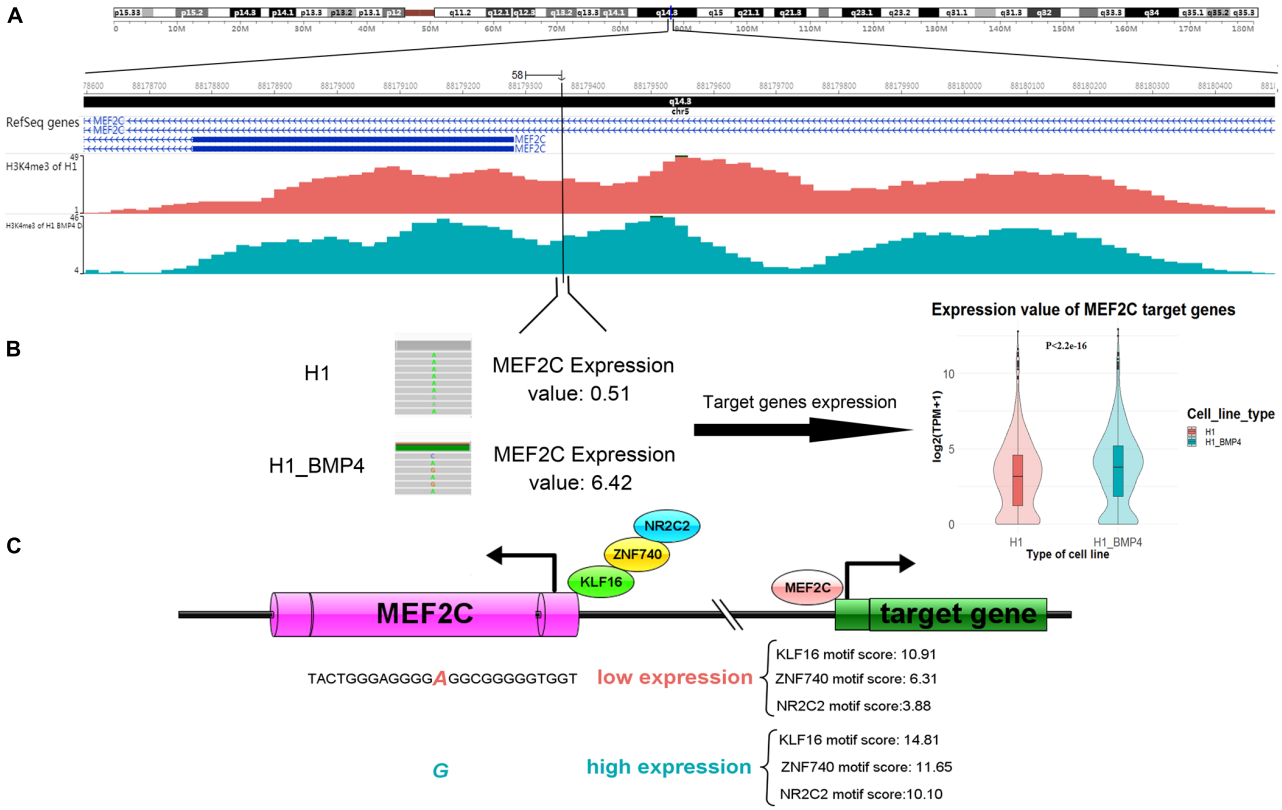


FIGURE 6 | A genomic variation (chromosomes 5: 88179358 A > G) inserts its function in differentiation by affecting MEF2C' expression, by altering the affinity of TFs binding. **(A)** The H3K4me3 track of H1 and H1_BMP4 around the variation site. **(B)** The gene expression of TF MEF2C was significantly increased (left panel). Violin plotting (right panel) shows that the gene expression of target genes of TF 'MEF2C' was significantly up-regulated between H1 and H1_BMP4 ($p < 2.2E-16$). **(C)** During differentiation, increased expression levels of the MEF2C gene is due to a genomic alteration, chromosomes 5: 88179358 A > G, which is at a binding site of TFs KLF16, NR2C2, and ZNF740. Allele G shows a higher affinity to TFs in the induced cell. A high expression of MEF2C leads to an increased expression of TF MEF2C's target genes, subsequently affecting the differentiation.

TEX2, INHBA-AS1, and CAMK2N1. The remaining one was at the enhancer of gene COL1A2.

We highlighted a core function of the six genes (MEF2C, TNFAIP8, TEX2, INHBA-AS1, CAMK2N1, and COL1A2) in differentiation and placenta development. TNFAIP8 plays a role in immune homeostasis, inflammatory responses, tumor genesis, and development. TNFAIP8 is also highly expressed in most normal human tissues especially for immunity-related tissues like the placenta (Zhang et al., 2018). INHBA encodes a member of the TGF-β (transforming growth factor-beta) superfamily of proteins which has been proven to promote the differentiation of human embryonic stem cells into trophoblasts (Pucéat, 2007). COL1A2 encodes one of the chains for type I collagen, the fibrillar collagen found in most connective tissues, and it is an early stage marker of osteoblast differentiation (Parisuthiman et al., 2005). MEF2C plays a pivotal role in myogenesis, neural crest, and craniofacial development, and may have an influence on maintaining the differentiated state of muscle cells (Zweier et al., 2010). In a recent study, dysregulation of MEF2 expression or signaling in early pregnancy may be associated with placenta-related pregnancy disorders, including preeclampsia (Li et al., 2017). In total, the six genes both associate a genomic variation in their regulatory regions and show differential expression from H1 to H1_BMP4, suggesting that the genomic variations associated with differentiation.

Since the six genomic variations are in non-coding regions, we assessed the effect of the variation in altering the binding affinity of TFs. We applied JASPAR2018[3] to calculate the motif score of 12-bp DNA sequences around the six SNVs/indels to the TFs that bind at those variation-harbored regions (Khan et al., 2018). We listed the motif scores in **Supplementary Excel File 3**. We found that the allele G of a variation (chr5: 88179358 A > G), which is at the enhancer of gene MEF2C, can increase the motif scores ≥ 10 in the induced cells (**Table 4**). it should be considered, however, that the MEF2C expression level is ∼4 fold higher in H1_BMP4 than in H1. The results mean that the genomic variation accounts for the increase of MEF2C expression by increasing TFs' affinity at MEF2C's promoter. Moreover, the H2K27ac does not show a significant change between the two cell lines. We

[3]http://jaspar.genereg.net/

confirmed that the MEF2C expression increase is caused by the genomic variation.

Importantly, the MEF2C gene itself encodes a TF protein. More recently, MEF2 was proven to regulate human trophoblast differentiation and invasion (Li et al., 2017). We retrieved the list of 954 MEF2C target genes reported in ENCODE TF targets (Rouillard et al., 2016). We then compared the TF MEF2C target genes' expression between H1 and H1_BMP4. The result showed that they are significantly up-regulated in the transition from human ESC and the trophoblast (paired sample *t*-test, $p < 2.2E\text{-}16$) (**Figure 6**).

Functional annotations by Enrichr (Kuleshov et al., 2016) on KEGG (Kyoto Encyclopedia of Genes and Genomes) and GO (Gene ontology) suggest the target genes of TF MEF2C were enriched in the "NF-kappa B signaling pathway" (Pollheimer and Knöfler, 2005), "Primary immunodeficiency," "Systemic lupus erythematosus," "positive regulation of NF-kappaB transcription factor activity," "positive regulation of immune response," and so on (adjust *p*-value ≤ 0.05, **Supplementary Figure S5**). Literature suggests that the placenta is an important immune-related tissue (Erkers et al., 2017). We therefore considered that the target genes of TF MEF2C have a significant impact on the cell differentiation process.

Briefly, the gene expression of MEF2C was significantly increased due to the high-affinity binding sites of TFs like KLF16, NR2C2, and ZNF740, a variation site located in MEF2C (chromosomes 5: 88179358 A > G). Further, the expression levels of these three TFs were not significantly different between H1 and H1_BMP4, suggesting that MEF2C's expression increase is not caused by a rise in the abundance of the TFs of MEF2C, but due to the genomic variation (**Figure 6**). Interestingly, literature indicates that simultaneous depletion of KLF2, KLF4, and KLF5 leads to differentiation of the embryonic stem cell, and it has been postulated that other members of the KLF family, such as KLF16, may play similar roles in embryonic stem cells (Jiang et al., 2008; Andreoli et al., 2010). It has been established that NR2C2 (TR4) plays a critical role in embryonic development and differentiation.

NR2C2 is expressed in blastocysts and embryonic stem cells and can act as transcriptional activators in hESC (Shyr et al., 2009; O'Geen et al., 2010). Altogether, this indicates that due to the up-regulation of MEF2C, its target genes were up-regulated. Importantly, the upregulation of MEF2C is caused by the genomic variation (chr5: 88179358 A > G), which alters the affinity of MEF2C's TFs.

DISCUSSION

Understanding the genetic underpinnings of complex traits remains a major challenge in human genetics. In this study, we obtained paired genomic DNA sequences of human ESC and the trophoblast from three cell lines (H1, H9, and MRucR) through whole genome sequencing, and integrated the epigenomic (DNA methylation, histone modifications and chromatin accessibility) and transcriptomic datasets to investigate the impact of the genome variations in human ESC differentiation to the trophoblast. We found that the SNVs and indels generally tend to be located in the intron regions rather than in the other regions.

We focused on the gene expression variation caused by genomic variation rather than the epigenetic variation. Six genomic variations were identified. One of them, located in MEF2C (chromosomes 5: 88179358 A > G), is a TF. This SNV increased the TF binding strength to regulate gene expression directly. Thereby, leading to an increase in the expression of downstream target genes affecting the differentiation of human ESC into the trophoblast. It suggested that the variations in the non-coding region played an important role in the differentiation process.

The inducer, BMP4, is the most significant factor to differentiate to the trophoblast. BMP4 is able to inhibit the Activin/Nodal signaling pathway and activate the BMP signaling pathway, which is required for human ESCs to differentiate into trophoblasts (Xu et al., 2002). We showed that ∼45% (42 genes) of the differentially expressed BMP4-induced genes associated with genomic variations. Although genomic variations are not possible to be the only dominant factor in differentiation, some genomic variations indeed have an effect on differentiation. There are two limitations in this work. The first is that the conclusion is confined to only three pairs of cell lines. A similar analysis should be carried out in more cell lines of iPSC. The second limitation is that comprehensive biochemical experiments are still needed to validate the conclusion. CRISPR technology in cultured cells could be employed to prove the role of genomic variation in the differentiation process (Zhou et al., 2014, p. 2).

DATA AVAILABILITY STATEMENT

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Wang et al., 2017) at the BIG Data Center (Zhang et al., 2019), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number CRA002190, which is publicly accessible at https://bigd.big.ac.cn/gsa.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design, analyzed and discussed the results. XS conceived the project and completed the core program. HaL, HoL, and YL performed the computational analysis. HaL and HoL wrote the manuscript.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (Nos. 81830053, 61972084)

and the Key Research & Development Program of Jiangsu Province (BE2016002-3).

ACKNOWLEDGMENTS

We thank Professor Toshihiko Ezashi (University of Missouri) and Professor R. Michael Roberts (University of Missouri) for providing genomic materials obtained from three human pluripotent stem cells (H1, H9, and MRucR). We would also like to thank Professor Toshihiko Ezashi and Yue Hou for critical reading and helpful discussions for the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00230/full#supplementary-material

REFERENCES

Abad, M., Mosteiro, L., Pantoja, C., Cañamero, M., Rayon, T., Ors, I., et al. (2013). Reprogramming *in vivo* produces teratomas and iPS cells with totipotency features. *Nature* 502, 340–345. doi: 10.1038/nature12586

Andreoli, V., Gehrau, R. C., and Bocco, J. L. (2010). Biology of Krüppel-like factor 6 transcriptional regulator in cell life and death. *IUBMB Life* 62, 896–905. doi: 10.1002/iub.396

Banovich, N. E., Li, Y. I., Raj, A., Ward, M. C., Greenside, P., Calderon, D., et al. (2018). Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* 28, 122–131. doi: 10.1101/gr.224436.117

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* 6, 677–681. doi: 10.1038/nmeth.1363

DaSilva-Arnold, S. C., Kuo, C. Y., Davra, V., Remache, Y., Kim, P. C. W., Fisher, J. P., et al. (2019). ZEB2, a master regulator of the epithelial-mesenchymal transition, mediates trophoblast differentiation. *Mol. Hum. Reprod.* 25, 61–75. doi: 10.1093/molehr/gay053

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

de Boni, L., Gasparoni, G., Haubenreich, C., Tierling, S., Schmitt, I., Peitz, M., et al. (2018). DNA methylation alterations in iPSC- and hESC-derived neurons: potential implications for neurological disease modeling. *Clin. Epigenet.* 10:13. doi: 10.1186/s13148-018-0440-440

DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K. M., et al. (2017). Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells. *Cell Stem Cell* 20, 533.e7–546.e7. doi: 10.1016/j.stem.2017.03.009

Delahaye, F., Do, C., Kong, Y., Ashkar, R., Salas, M., Tycko, B., et al. (2018). Genetic variants influence on the placenta regulatory landscape. *PLoS Genet.* 14:e1007785. doi: 10.1371/journal.pgen.1007785

Dreszer, T. R., Karolchik, D., Zweig, A. S., Hinrichs, A. S., Raney, B. J., Kuhn, R. M., et al. (2012). The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* 40, D918–D923. doi: 10.1093/nar/gkr1055

Erkers, T., Stikvoort, A., and Uhlin, M. (2017). Lymphocytes in placental tissues: immune regulation and translational possibilities for immunotherapy. *Stem Cells Int.* 2017:5738371. doi: 10.1155/2017/5738371

Goolam, M., Scialdone, A., Graham, S. J. L., Macaulay, I. C., Jedrusik, A., Hupalowska, A., et al. (2016). Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* 165, 61–74. doi: 10.1016/j.cell.2016.01.047

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors Prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Ishiuchi, T., Ohishi, H., Sato, T., Kamimura, S., Yorino, M., Abe, S., et al. (2019). Zfp281 shapes the transcriptome of trophoblast stem cells and is essential for placental development. *Cell Rep.* 27, 1742–1754. doi: 10.1016/j.celrep.2019.04.028

Jiang, J., Chan, Y.-S., Loh, Y.-H., Cai, J., Tong, G.-Q., Lim, C.-A., et al. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat. Cell Biol.* 10, 353–360. doi: 10.1038/ncb1698

Johnston, A. D., Simões-Pires, C. A., Thompson, T. V., Suzuki, M., and Greally, J. M. (2019). Functional genetic variants can mediate their regulatory effects through alteration of transcription factor binding. *Nat. Commun.* 10, 1–16. doi: 10.1038/s41467-019-11412-11415

Karczewski, K. J., Tatonetti, N. P., Landt, S. G., Yang, X., Slifer, T., Altman, R. B., et al. (2011). Cooperative transcription factor associations discovered using regulatory variation. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13353–13358. doi: 10.1073/pnas.1103105108

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 46, D260–D266. doi: 10.1093/nar/gkx1126

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., et al. (2013). Coordinated effects of sequence variation on DNA binding. Chromatin structure, and transcription. *Science* 342, 744–747. doi: 10.1126/science.1242463

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* 15, 591–594. doi: 10.1038/s41592-018-0051-x

Knöfler, M., Haider, S., Saleh, L., Pollheimer, J., Gamage, T. K. J. B., and James, J. (2019). Human placenta and trophoblast development: key molecular mechanisms and model systems. *Cell. Mol. Life Sci.* 76, 3479–3496. doi: 10.1007/s00018-019-03104-3106

Kubaczka, C., Senner, C. E., Cierlitza, M., Araúzo-Bravo, M. J., Kuckenberg, P., Peitz, M., et al. (2015). Direct induction of trophoblast stem cells from murine fibroblasts. *Cell Stem Cell* 17, 557–568. doi: 10.1016/j.stem.2015.08.005

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

Langley, S. A., Karpen, G. H., and Langley, C. H. (2014). Nucleosomes shape DNA polymorphism and divergence. *PLoS Genet.* 10:e1004457. doi: 10.1371/journal.pgen.1004457

Lecomte, V., Meugnier, E., Euthine, V., Durand, C., Freyssenet, D., Nemoz, G., et al. (2010). A new role for sterol regulatory element binding protein 1 transcription factors in the regulation of muscle mass and muscle cell differentiation. *Mol. Cell. Biol.* 30, 1182–1198. doi: 10.1128/MCB.00690-699

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

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

Li, L., Gong, X., and Rubin, L. P. (2017). Expression of MEF2 transcription factors in human placenta and involvement in trophoblast invasion and differentiation. *FASEB J.* 31:692.8. doi: 10.1096/fasebj.31.1_supplement.692.8

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., et al. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600–604. doi: 10.1126/science.aad9417

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., et al. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* 471, 68–73. doi: 10.1038/nature09798

Liu, Y., Ding, D., Liu, H., and Sun, X. (2017). The accessible chromatin landscape during conversion of human embryonic stem cells to trophoblast by bone morphogenetic protein 4. *Biol. Reprod.* 96, 1267–1278. doi: 10.1093/biolre/iox028

Madsen, J. G. S., Rauch, A., Van Hauwaert, E. L., Schmidt, S. F., Winnefeld, M., and Mandrup, S. (2018). Integrated analysis of motif activity and gene expression changes of transcription factors. *Genome Res.* 28, 243–255. doi: 10.1101/gr.227231.117

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., et al. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics* 28, 1919–1920. doi: 10.1093/bioinformatics/bts277

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., et al. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* 95, 379–391. doi: 10.1016/S0092-8674(00)81769-81769

O'Geen, H., Lin, Y.-H., Xu, X., Echipare, L., Komashko, V. M., He, D., et al. (2010). Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. *BMC Genomics* 11:689. doi: 10.1186/1471-2164-11-689

Parisuthiman, D., Mochida, Y., Duarte, W. R., and Yamauchi, M. (2005). Biglycan modulates osteoblast differentiation and matrix mineralization. *J. Bone Miner. Res.* 20, 1878–1886. doi: 10.1359/JBMR.050612

Pollheimer, J., and Knöfler, M. (2005). Signalling pathways regulating the invasive differentiation of human trophoblasts: a review. *Placenta* 26, S21–S30. doi: 10.1016/j.placenta.2004.11.013

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65. doi: 10.1093/nar/gkl842

Pucéat, M. (2007). TGFβ in the differentiation of embryonic stem cells. *Cardiovasc. Res.* 74, 256–261. doi: 10.1016/j.cardiores.2006.12.012

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

Roadmap Epigenomics Consortium Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., et al. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016:baw100. doi: 10.1093/database/baw100

Sheridan, M. A., Yang, Y., Jain, A., Lyons, A. S., Yang, P., Brahmasani, S. R., et al. (2019). Early onset preeclampsia in a model for human placental trophoblast. *Proc. Natl. Acad. Sci. U.S.A.* 116, 4336–4345. doi: 10.1073/pnas.1816150116

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shimasaki, S., Moore, R. K., Otsuka, F., and Erickson, G. F. (2004). The bone morphogenetic protein system in mammalian reproduction. *Endocr. Rev.* 25, 72–101. doi: 10.1210/er.2003-0007

Shyr, C.-R., Kang, H.-Y., Tsai, M.-Y., Liu, N.-C., Ku, P.-Y., Huang, K.-E., et al. (2009). Roles of testicular orphan nuclear receptors 2 and 4 in early embryonic development and embryonic stem cells. *Endocrinology* 150, 2454–2462. doi: 10.1210/en.2008-1165

Su, J., Shao, X., Liu, H., Liu, S., Wu, Q., and Zhang, Y. (2012). Genome-wide dynamic changes of DNA methylation of repetitive elements in human embryonic stem cells and fetal fibroblasts. *Genomics* 99, 10–17. doi: 10.1016/j.ygeno.2011.10.004

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 663–676. doi: 10.1016/j.cell.2006.07.024

The 1000 Genomes Project Consortium, (2010). A map of human genome variation from population scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., et al. (2017). GSA: genome sequence archive*. *Genomics Proteom. Bioinform.* 15, 14–18. doi: 10.1016/j.gpb.2017.01.001

Wang, Z., Oron, E., Nelson, B., Razis, S., and Ivanova, N. (2012). Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell Stem Cell* 10, 440–454. doi: 10.1016/j.stem.2012.02.016

Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., et al. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153, 1134–1148. doi: 10.1016/j.cell.2013.04.022

Xu, J. G., Zhu, S. Y., Heng, B. C., Dissanayaka, W. L., and Zhang, C. F. (2017). TGF-β1-induced differentiation of SHED into functional smooth muscle cells. *Stem Cell Res. Ther.* 8:10. doi: 10.1186/s13287-016-0459-450

Xu, R.-H., Chen, X., Li, D. S., Li, R., Addicks, G. C., Glennon, C., et al. (2002). BMP4 initiates human embryonic stem cell differentiation to trophoblast. *Nat. Biotechnol.* 20:1261. doi: 10.1038/nbt761

Yabe, S., Alexenko, A. P., Amita, M., Yang, Y., Schust, D. J., Sadovsky, Y., et al. (2016). Comparison of syncytiotrophoblast generated from human embryonic stem cells and from term placentas. *PNAS* 113, E2598–E2607. doi: 10.1073/pnas.1601630113

Zhang, L., Liu, R., Luan, Y., and Yao, Y. (2018). Tumor necrosis factor-α induced protein 8: pathophysiology, clinical significance, and regulatory mechanism. *Int. J. Biol. Sci.* 14, 398–405. doi: 10.7150/ijbs.23268

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Zhang, Z., Zhao, W., Xiao, J., Bao, Y., Wang, F., Hao, L., et al. (2019). Database resources of the BIG data center in 2019. *Nucleic Acids Res.* 47, D8–D14. doi: 10.1093/nar/gky993

Zhou, H. Y., Katsman, Y., Dhaliwal, N. K., Davidson, S., Macpherson, N. N., Sakthidevi, M., et al. (2014). A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* 28, 2699–2711. doi: 10.1101/gad.248526.114

Zweier, M., Gregor, A., Zweier, C., Engels, H., Sticht, H., Wohlleber, E., et al. (2010). Mutations in MEF2C from the 5q14.3*q*15 microdeletion syndrome region are a frequent cause of severe mental retardation and diminish MECP2 and CDKL5 expression. *Hum. Mutat.* 31, 722–733. doi: 10.1002/humu.21253

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 © 2020 Li, Liu, Liu and Sun. 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 Genetics

CORRECTION
published: 28 May 2020
doi: 10.3389/fgene.2020.00538

Corrigendum: Effect for Human Genomic Variation During the BMP4-Induced Conversion From Pluripotent Stem Cells to Trophoblast

Hai-tao Li[1], *Yajun Liu*[1,2,3], *Hongde Liu*[1]* *and Xiao Sun*[1]*

[1] State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China, [2] The Second Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [3] Academy of Medical Sciences of Zhengzhou University Translational Medicine Platform, Zhengzhou University, Zhengzhou, China

Keywords: genomic variations, pluripotent stem cells, trophoblast, whole genome sequencing, epigenomic and transcriptomic data

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

***Correspondence:**
Hongde Liu
liuhongde@seu.edu.cn
Xiao Sun
xsun@seu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *03 May 2020*
Accepted: *04 May 2020*
Published: *28 May 2020*

Citation:
Li H, Liu Y, Liu H and Sun X (2020) Corrigendum: Effect for Human Genomic Variation During the BMP4-Induced Conversion From Pluripotent Stem Cells to Trophoblast. Front. Genet. 11:538. doi: 10.3389/fgene.2020.00538

A Corrigendum on

Effect for Human Genomic Variation During the BMP4-Induced Conversion From Pluripotent Stem Cells to Trophoblast

by Li, H., Liu, Y., Liu, H., and Sun, X. (2020). Front. Genet. 11:230. doi: 10.3389/fgene.2020.00230

In the original article, there was an error in the Acknowledgments. It was written that Professor Toshihiko Ezashi and Professor R. Michael Roberts provided "three human pluripotent stem cells"; however, this is not accurate. They provided "genomic materials obtained from three human pluripotent stem cells."

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

Copyright © 2020 Li, Liu, Liu and Sun. 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 April 2020
doi: 10.3389/fcell.2020.00242

Detection of Allosteric Effects of lncRNA Secondary Structures Altered by SNPs in Human Diseases

Xiaoyan Lu[1,2†], Yu Ding[1,2†], Yu Bai[1,2†], Jing Li[1,2], Guosi Zhang[1], Siyu Wang[1], Wenyan Gao[1,2], Liangde Xu[1,2] and Hong Wang[1,2*]*

[1] School of Ophthalmology and Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou, China, [2] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Recent studies have shown that structuralized long non-coding RNAs (lncRNAs) play important roles in genetic and epigenetic processes. The spatial structures of most lncRNAs can be altered by distinct *in vivo* and *in vitro* cellular environments, as well as by DNA structural variations, such as single-nucleotide polymorphisms (SNPs) and variants (SNVs). In the present study, we extended candidate SNPs that had linkage disequilibria with those significantly associated with lung diseases in genome-wide association studies in order to investigate potential disease mechanisms originating from SNP structural changes of host lncRNAs. Following accurate alignments, we recognized 115 ternary-relationship pairs among 41 SNPs, 10 lncRNA transcripts, and 1 type of lung disease (adenocarcinoma of the lung). Then, we evaluated the structural heterogeneity induced by SNP alleles by developing a local-RNA-structure alignment algorithm and employing randomized strategies to determine the significance of structural variation. We identified four ternary-relationship pairs that were significantly associated with SNP-induced lncRNA allosteric effects. Moreover, these conformational changes disrupted the interactive regions and binding affinities of lncRNA-HCG23 and TF-E2F6, suggesting that these may represent regulatory mechanisms in lung diseases. Taken together, our findings support that SNP-induced changes in lncRNA conformations regulate many biological processes, providing novel insight into the role of the lncRNA "structurome" in human diseases.

Keywords: lncRNA secondary structure, linkage-disequilibrium SNPs, structural heterogeneity, transcription factors, human diseases

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Dapeng Hao,
Baylor College of Medicine, United States
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

****Correspondence:***
Liangde Xu
xuld@eye.ac.cn
Hong Wang
wanghong84@ems.hrbmu.edu.cn

† These authors have contributed equally to this work

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

Received: *06 February 2020*
Accepted: *23 March 2020*
Published: *08 April 2020*

Citation:
Lu X, Ding Y, Bai Y, Li J, Zhang G, Wang S, Gao W, Xu L and Wang H (2020) Detection of Allosteric Effects of lncRNA Secondary Structures Altered by SNPs in Human Diseases. Front. Cell Dev. Biol. 8:242. doi: 10.3389/fcell.2020.00242

INTRODUCTION

With the development of whole-genome sequencing technology, long non-coding RNAs (lncRNAs) have been studied and discovered to play a key role in complex diseases. LncRNAs regulate gene expression at epigenetic, transcriptional, and post-transcriptional levels (Chen et al., 2019). In lung cancers, HOX antisense intergenic RNA (HOTAIR), a well-studied lncRNA, has been shown to correlate with metastasis and poor prognosis (Loewen et al., 2014; Wang et al., 2018). In addition, aside from regulating expression levels of genes, lncRNA structures govern a complex post-transcriptional regulatory program in diseases (Fujimoto et al., 2016). LncRNAs have been shown to form structural domains that function as landing pads for transcription factors (TFs)

to participate in transcriptional regulation (Wang et al., 2017). Since lncRNAs are known to play important roles in various diseases, considerable research has focused on elucidating potential relationships between disease phenotypes and lncRNA structural conformations.

Single-nucleotide polymorphisms (SNPs) are the most common type of variants in the human genome. Functional SNPs not only affect gene expression, but they also influence the structures and stabilities of RNAs (Ramírez-Bello and Jiménez-Morales, 2017). By affecting binding affinities, SNPs regulate gene expression in various diseases at the post-transcriptional level and can thus decrease invasion ability of genes (Halvorsen et al., 2010; Pirooz et al., 2018). Moreover, disease-associated linkage-disequilibrium (LD) SNPs have been predicted to alter the ensemble of RNA structures and to further affect RNA-protein binding sites (Martin et al., 2012). Therefore, investigating haplotypes that include specific pairs of SNPs in high LD may contribute to better understanding pathogenic mechanisms in various diseases.

Recently, lncRNAs have been implicated in several diseases. In addition, many disease-associated SNPs modify the secondary structures of lncRNAs, which affect their expressions and functions, thus leading to the development of diseases (Castellanos-Rubio and Ghosh, 2019). Furthermore, risk variants and their LD SNPs decrease binding affinities of TFs and lncRNAs (Hua et al., 2018). Taking together, known disease-associated or their LD SNPs may cause structural rearrangements of molecules and contribute to disease progression.

In the present study, we investigated LD SNPs of lung-disease-associated SNPs and mapped them onto lncRNA transcripts across the whole human genome. Connections among single LD SNPs, lncRNAs, and lung diseases were then determined using this methodology. Additionally, the structural heterogeneity of lncRNAs generated by single LD SNPs and their haplotypes were quantified via a computational algorithm. We identified single LD SNPs that significantly altered second structures of lncRNAs. Furthermore, we predicted changes in binding affinities between lncRNAs and TFs. Our comprehensive pipeline was divided into three parts (**Figure 1**). Collectively, our findings provide further insight into potential molecular mechanisms of lung diseases by demonstrating that lung-disease-associated LD SNPs affect RNA structural rearrangements and concomitantly modulate many biological processes.

MATERIALS AND METHODS

Obtaining and Preprocessing Data

Human disease-associated SNPs were obtained from the Database of Genotypes and Phenotypes (dbGaP), which provided large genetic and phenotypic datasets (Wong et al., 2017). A total of 32 samples of disease phenotypes were downloaded. We identified 42 SNPs associated with lung diseases by searching the following keywords: "lung," "lung cancer," and "lung carcinoma." These SNPs were associated with five types of lung-related diseases, namely, adenocarcinoma of the lung, non-small-cell carcinoma of the lung, small-cell lung carcinoma, lung neoplasms, and squamous-cell carcinoma of the lung. All of these lung-associated genotypes and phenotypes were used for follow-up analyses.

All of the lncRNA-sequence datasets from the whole human genome were downloaded from The GENCODE consortium version 29 (GENCODE V29), which involved comprehensive genomic annotations of lncRNAs that were recruited from GRCh38 (Harrow et al., 2012). Ultimately, 16,042 mature lncRNA genes and 29,566 alternative isoforms were selected for further study.

Identifying Linkage-Disequilibrium Blocks

LD SNPs can induce substantial changes in the structural ensemble of RNAs (Martin et al., 2012). We identified LD blocks around disease-associated SNPs (LD SNPs), from which we estimated the structural influences of SNPs around lung-disease-associated SNPs. Datasets of SNPs from the 1000 Genomes Project—including chromosome files with genotypes for all of the samples and detailed descriptions of each individual sample—were used as raw LD datasets (Genomes Project et al., 2015; Sudmant et al., 2015). We chose the GRCh38 reference genome to ensure consistency of data sources.

LD blocks associated with lung diseases were extracted as follows. First, samples and SNPs derived from East Asian individuals were selected. Second, only SNPs with two alleles were selected. Third, only SNPs with minor allele frequencies (MAFs) exceeding 5% (common variants) and missing value proportions under 25% were selected; additionally, we required that the SNP genotype of each included sample reach up to 75%. Only samples with P values less than 0.01 were selected as significant SNPs. Based on these inclusion criteria and the PLINK toolset, we obtained 42 LD blocks associated with 42 disease-associated SNPs (Purcell et al., 2007).

Repositioning SNPs in lncRNA Transcripts

Variation analysis of lncRNA transcripts was completed by repositioning SNPs. Bowtie 2, an ultrafast and memory-efficient tool, was applied to map SNPs onto lncRNA transcripts (Langmead and Salzberg, 2012). First, we chose mature lncRNA transcripts as reference sequences. According to the input, Bowtie 2 built a library of long reference sequences. The dbSNP database records sequence information around SNPs (Sherry et al., 2001). The 25-bp upstream and downstream flanking regions of each identified LD SNP were collected from the dbSNP database. Then, at the center of each SNP site, the 25-bp upstream and downstream regions (as short reads) were aligned with lncRNAs. Based on this short-read alignment strategy, we set strict parameters (e.g., end-to-end, –score-min) to ensure precise locations of SNPs. Finally, the output-SAM file contained the symbols of lncRNA transcripts and SNPs, the positions of nucleic acids where matching reads appeared, and the components of the corresponding short reads. We screened start positions both in left and right side of identical lncRNA transcripts. Next, the distance of both

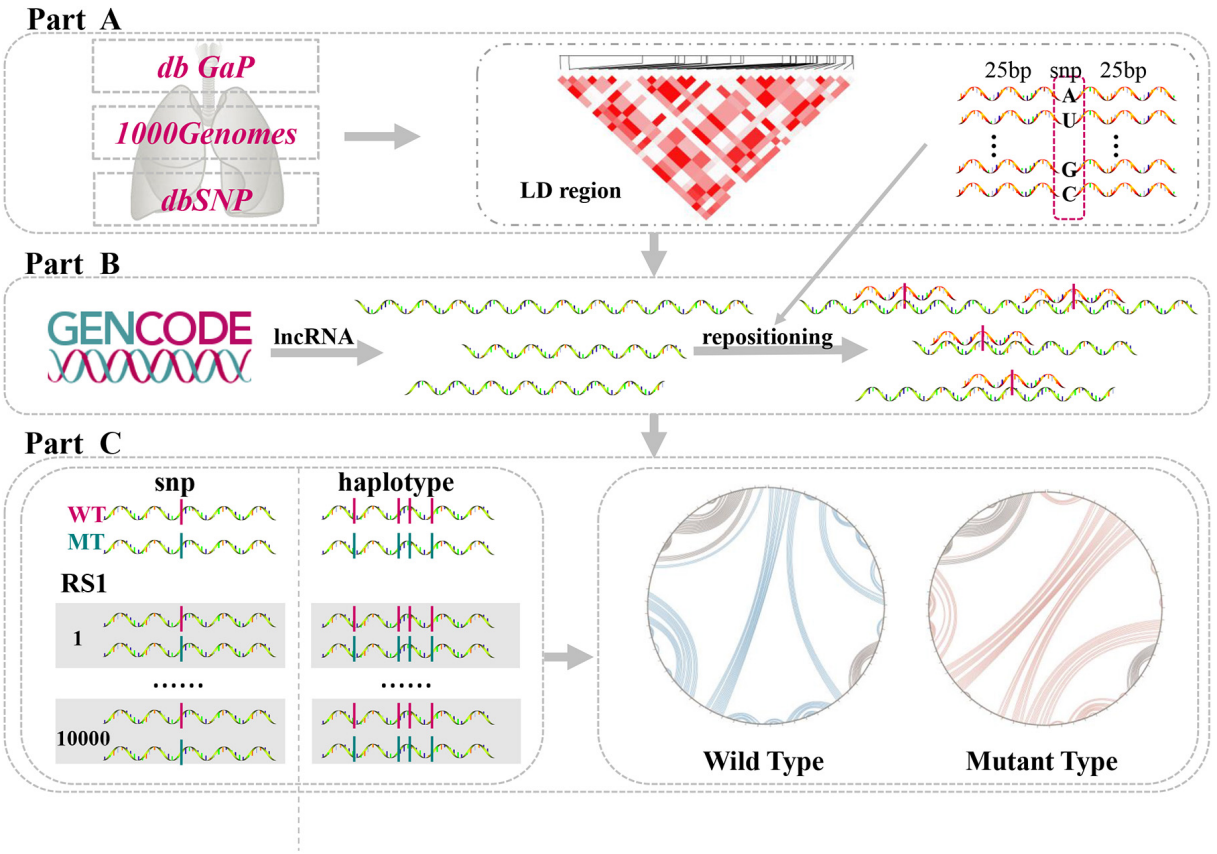


FIGURE 1 | Comprehensive analysis of lncRNA structural heterogeneity generated by linkages-disequilibrium SNPs (LD SNPs). **(A)** Sources of data of lung-disease-associated SNPs and their LD SNPs. **(B)** Information on lncRNA transcripts and positions of LD SNPs within lncRNAs. **(C)** Perturbations to evaluate the differences between WT and MT lncRNAs.

ends was used to decide whether SNPs mapped on lncRNA transcripts. The direction of positive and negative in short-read alignment should be taken into account. If the absolute value of distance was 26, it generally indicated SNPs located on lncRNA transcripts.

Quantifying Structural Heterogeneity of lncRNAs

The exact locations of lung-associated SNPs are a foundation for assessing lncRNA structural disturbances. First, mature lncRNA transcripts downloaded from GENCODE were defined as wild-type (WT) sequences. Meanwhile, lncRNA transcripts with one or more mapped SNPs were assigned as mutant (MT) sequences. Furthermore, we used Linux-based RNA-structure software packages to identify the secondary structures of WT and MT sequences (Reuter and Mathews, 2010). Subsequently, the structural heterogeneity of lncRNAs was quantified via the RNAsmc score designed by our research group, which is the output of an algorithm that computes the difference between two lncRNAs. The stem loop (S), bulge loop (B), interior loop (I), hairpin (H), and multi-branched loop (M) were considered to represent the most essential elements for RNA secondary structures. The locations and amounts of these structural elements were used to calculate the value of the RNAsmc score. The principle of RNAsmc score is as follows:

$$SS = \sum_{u\in\{S,H,I,B,M\}} \frac{u_{p1} \bigcap u_{p2}}{u_{p1} \bigcup u_{p2}} + \sum_{u\in\{S,H,I,B,M\}} \frac{\min(u_{n1}, u_{n2})}{\max(u_{n1}, u_{n2})}$$

Here, SS is equal to the RNAsmc score which represent the similarity between lncRNA structures; *S*, *H*, *I*, *B*, *M* represents five sub-units as mentioned above; u_{p1},$u_{p2}$are the location set of two lncRNA's base for each kind of sub-units; u_{n1} and u_{n2} are the number of each sub-units in each lncRNA structures. We can infer from the scored rules that if there is no difference between two structures, the score is 10; however, if two structures have no overlapping, the score becomes 0. The RNAsmc score was limited to a range of 0 to 10, in which values close to 0 represent a large difference between the two analyzed lncRNA structures, whereas a value of 10 represents structural homogeneity. In addition, in order to show the RNAsmc score was well designed to robustly evaluate the structural heterogeneity, we chose four different score and illustrated their second structure in **Supplementary Figure S1**. As we expected, the lower score suggested the greater difference between wild-type and mutant

lncRNA second structure. This result illustrated that the RNAsmc score was robust.

Assessing Haplotype-Induced Structural Disturbances of lncRNAs

After assessing the structural heterogeneity of lncRNAs from single SNPs, we next investigated structural transformations induced by haplotype blocks (a series of SNPs within an lncRNA transcript). As we expected, the haplotype was consisted of multiple SNPs in random way. However, the combination among SNPs had not only in reference to linkage disequilibrium, but also closely associated with populations. In population, haplotypes followed special rules to regulate individual biological procedure. Therefore, a comprehensive quality control was essential to acquire haplotypes. First, the annotations of SNPs within lncRNA transcripts from the 1000 Genomes Project were integrated, including the sample, sex, alleles, and genotypes of each SNP. Then, we used PLINK, an open-source toolset for analyzing whole-genome associations, to predict possible combinations of SNPs in the population. In addition, the RNAsmc score was calculated to evaluate structural disturbances by comparing the architectures of WT and MT lncRNA transcripts, which carried haplotype blocks.

Evaluating Significance of SNP-Modulated Structural Heterogeneity

We further assessed the significance of SNP-modulated lncRNA structural heterogeneity in two ways. First, while keeping the WT and MT SNP sites within lncRNA transcripts unchanged, we performed 10,000 permutations of the flanking sequences of these sites. Additionally, the background distributions of RNAsmc scores between random WT and MT transcripts were calculated and ranked. The *P* value, defined as the Random Score 1 (RS1), was determined by the order of real RNAsmc scores among random scores.

As a second strategy, for a lncRNA sequence with N-bp, we mutated each base into three other bases and obtained all of the possible 3N mutations. The background distributions of scores were computed between the WT sequence and all of the mutated sequences. Subsequently, the *P* value was computed as described above. The mean estimated significance was defined as the Random Score 2 (RS2). In our study, a $P < 0.05$ was used to assign SNPs that significantly altered the conformation of lncRNA transcripts.

Predicting Variation in Molecular Binding Ability

We evaluated the association between molecular function and modifications in lncRNA conformation. LncRNAs involved in transcriptional regulation of molecular interactions were annotated via manual searching from published papers and LncMAP databases (Li et al., 2018). The LncMAP database has integrated genome-wide transcriptional regulation with paired lncRNAs and gene expressions in pan-cancer. In this database, the regulatory states of lncRNAs and TFs in adenocarcinoma of the lung were detected via transcriptional regulatory network perturbation.

Although the relationships between lncRNAs and TFs are well known, their specific structural interactions are less understood. Here, we used CatRAPID software to predict the interactive region induced by structural units between WT and MT lncRNA transcripts and TFs (Agostini et al., 2013). The intuitive lncRNA secondary structures were visualized by VARNA (Darty et al., 2009). The PDB format of lncRNA transcripts and TFs were obtained by RNAComposer and I-TASSER, respectively (Yang and Zhang, 2015; Biesiada et al., 2016). Additionally, these datasets were then predicted via HDOCK, a web server for protein-RNA docking based on a hybrid strategy (Yan et al., 2017).

RESULTS

Mapping SNPs Onto lncRNA Transcripts

First, 42 SNPs (from an East Asian population) associated with 5 types of lung diseases were downloaded from dbGap (**Figure 2A**). These SNPs were filtered based on the Hardy-Weinberg Law. Then, we identified LD blocks around disease-associated SNPs (LD SNPs) using PLINK. According to short-read alignments, the LD SNPs were mapped onto lncRNA transcripts in GENCODE V29. We obtained 115 items consisting of 41 LD SNPs (expanded by rs3817963 and rs7216064; red label in **Figure 2A**), 4 lncRNA symbols (HCG23, AC134407.1, AC134407.2, AC134407.3) with 10 different transcripts, and 1 disease association (adenocarcinoma of the lung; **Supplementary Table S1**). Three SNPs mapped onto three transcripts, namely, AC134407.1, AC134407.2, and AC134407.3 (**Figure 2B**). Meanwhile, the lncRNA HCG23, suspected to be correlated with prostate cancer (Eeles et al., 2013), was matched with seven transcripts and 97.39% of all obtained items (**Figure 2C**). This result suggests that the above four lncRNAs contribute to the onset and development of pan-cancer, or act as necessary regulatory molecules in processes related to adenocarcinoma of the lung. In addition, we found that several SNPs were located in different regions within the same lncRNA transcript, for instance, rs17208657, rs57652561, rs12525722, rs117384660, rs17202309, rs9268475, rs3117099, rs117130854, rs115303880,and rs3117098, all SNPs located in ENST00000646550.1, which may have been due to the distance between each of these linkage SNPs being close to one another. Furthermore, in some cases, one SNP matched with several diverse lncRNA transcripts (**Figure 2C**). This representation may result from SNPs matched within overlapped fragments of lncRNA transcripts. For example, as shown in **Figure 2C** and **Supplementary Table S1**, rs17208657 was mapped onto six lncRNA transcripts (ENST00000642577.1, ENST00000644884.1, ENST 00000645134.1, ENST00000646550.1, ENST00000646628.1, and ENST00000647036.1). These one-to-one correspondences allowed us to explore the effects of LD SNPs on lncRNA transcripts. Additionally, these correspondences suggested that one lncRNA transcript may be influenced by several LD SNPs,

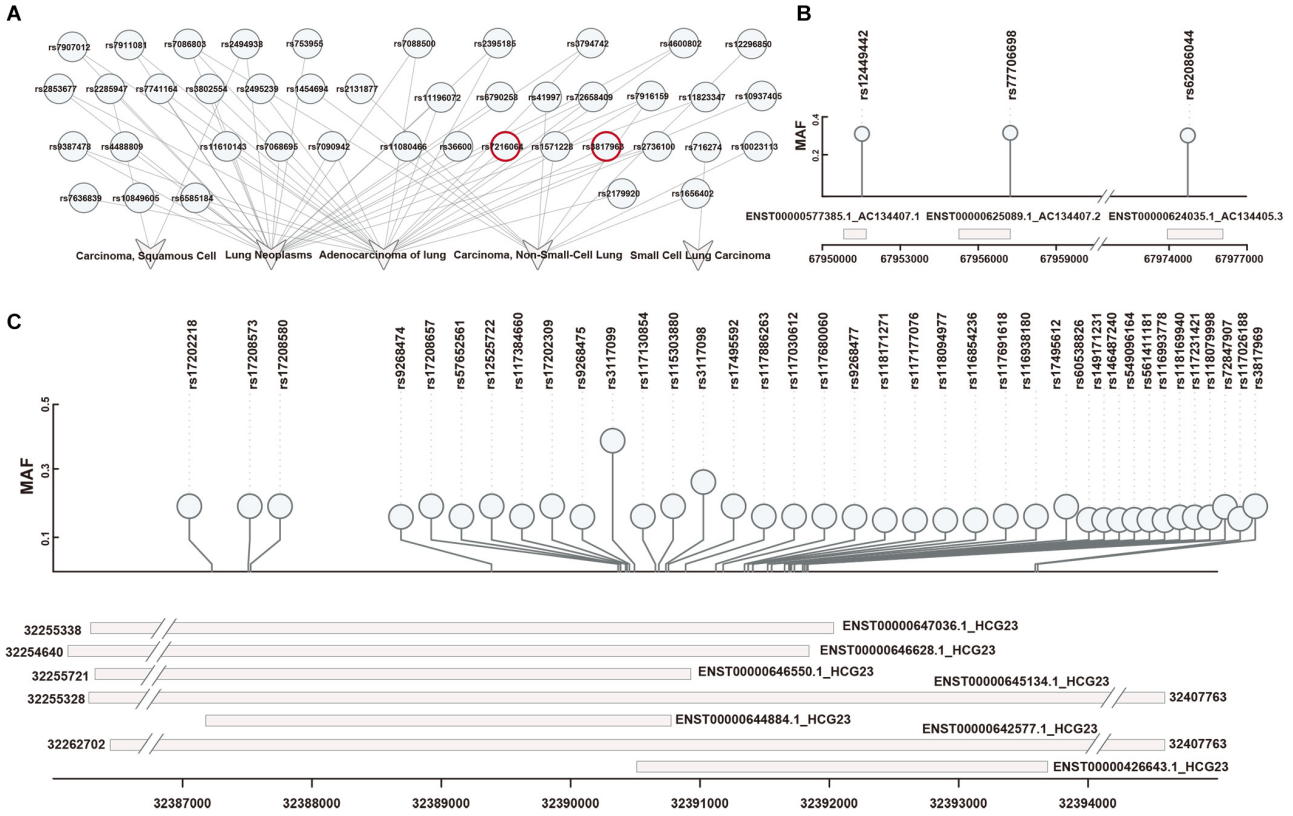


FIGURE 2 | Obtaining and repositioning of LD SNPs. **(A)** Lung-disease-associated SNPs were downloaded in dbGap. SNPs in red represented their linkage disequilibrium SNPs mapped on lncRNAs. **(B)** The one-to-one relationships of LD SNPs and lncRNA transcripts based on a short-sequence alignment algorithm. **(C)** The corresponding locations of LD SNPs mapped onto lncRNA transcripts.

or that diverse regulation of different lncRNA transcripts may be generated by identical SNPs.

Analyzing lncRNA Structural Heterogeneity

RNA secondary structure consists of five conformational sub-structures, namely, the stem loop (S), bulge loop (B), interior loop (I), hairpin (H), and multi-branched loop (M). In the present study, we focused on identifying LD SNPs that had an effect on lncRNA secondary structures. We took full advantage of an algorithmic toolkit, RNAsmc score, to probe lncRNA structural heterogeneity based on comparing these sub-structures. We analyzed 115 items that included 41 SNPs in 10 lncRNA transcripts that affected lncRNA secondary structure. The scores of WT and MT lncRNA transcripts were computed and illustrated as bubble charts in **Figure 3A**, with further information provided in **Supplementary Table S2**. We found that SNPs of 85 items had an effect on the lncRNA structural ensemble with scores under 10 (about 73.91% of SNPs gave rise to secondary structural variations of lncRNA transcripts), whereas all of the other SNPs (about 26.09%) had no impact, as indicated by their scores of 10. This result suggests that changes in sequences that resulted from SNPs may lead to conformational transformations of lncRNAs. In addition, such disturbances may affect the molecular function of lncRNAs within cells. For instance, changes in lncRNA confirmations may disrupt molecular binding, which may then influence epigenetic, transcriptional, and post-transcriptional regulation of lncRNAs. We found large SNP-induced conformational variations in lncRNAs (**Figure 3B**), which allowed us to then compare the extent of these SNP-induced structural changes. As shown in **Figure 3B**, the secondary structures were notably different in WT and MT HCG23 (four different transcripts of HCG23). This result illustrated that the majority of SNPs exhibited an influence on lncRNA secondary structure. Additionally, it is well known that structure often influences function. Therefore, we inferred that LD SNPs not only influence spatial structure, but they also functionally regulate lncRNAs. Furthermore, conformational changes in lncRNA structure may represent a possible cause of lung diseases.

Comparing and Assessing the Significance of lncRNA Structural Disturbances

Since RNAsmc scores alone are not able to determine the significance of lncRNA structural heterogeneity, we next designed two randomized schemes to strictly search for significant SNP-mediated structural changes. The permutation by

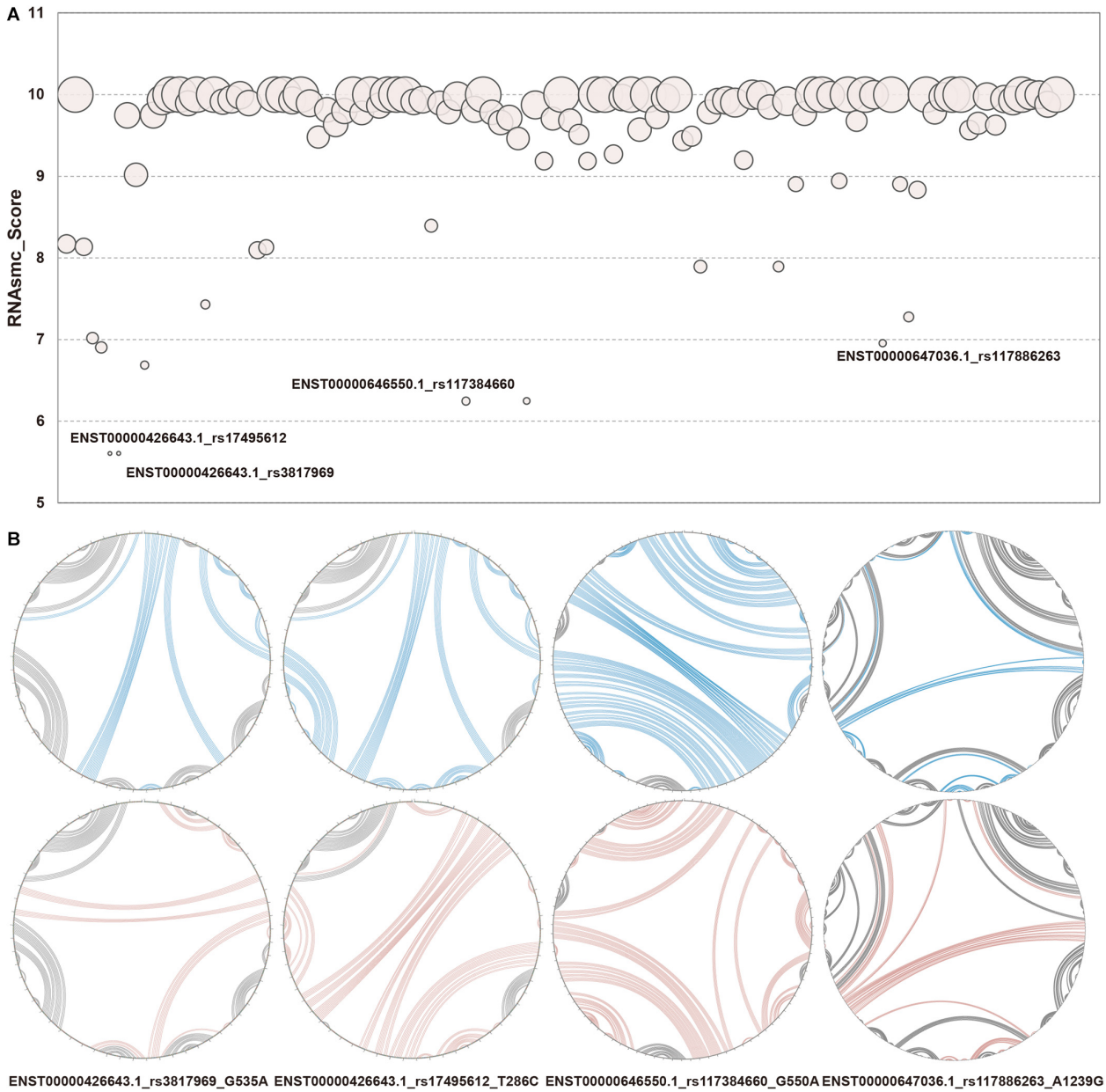


FIGURE 3 | Structural heterogeneity analysis of lncRNA transcripts altered by LD SNPs. **(A)** Quantification of WT and MT lncRNA transcripts. The x axis showed 115 items among LD SNPs, lncRNA transcripts, and lung diseases. The y axis represented RNAsmc scores. The size of each circle indicated the *P* value of the RS1. **(B)** Circular structural comparison of WT and MT lncRNA transcripts. The corresponding relationships between WT and MT lncRNA transcripts and LD SNPs. The lines in blue and red represented the corresponding regions of WT and MT lncRNA transcripts, respectively, altered by LD SNPs. The label under each circle indicated lncRNA transcript, SNP and alleles of SNP. e.g., G535A showed that 535 base G in ENST00000426643.1 change to A.

RS1 and RS2 was illustrated in **Supplementary Figures S2A,B**. The RS1 was used to calculate *P* values by rearranging flanking sequences of SNPs. In addition, RS2 considered all of the possibilities in the overall length of lncRNA sequences. To evaluate the consistency between RS1 and RS2, we selected items which their RNAsmc score were not 10 (10 means no difference among two structures). **Supplementary Figure S2C** indicated RS1 and RS2 had identical tendency in evaluating significance of lncRNA structural heterogeneity. And points in red represented significant items appeared by two methods. As determined by both RS1 and RS2, we identified four SNPs that significantly altered the secondary structures of lncRNA transcripts (**Figure 3B**). Moreover, an additional six SNPs were predicted at a $P < 0.05$ using RS2 (**Supplementary Tables S3, S4**). In **Supplementary Tables S3, S4**, although the outputs of significant *P* values between methods were distinct,

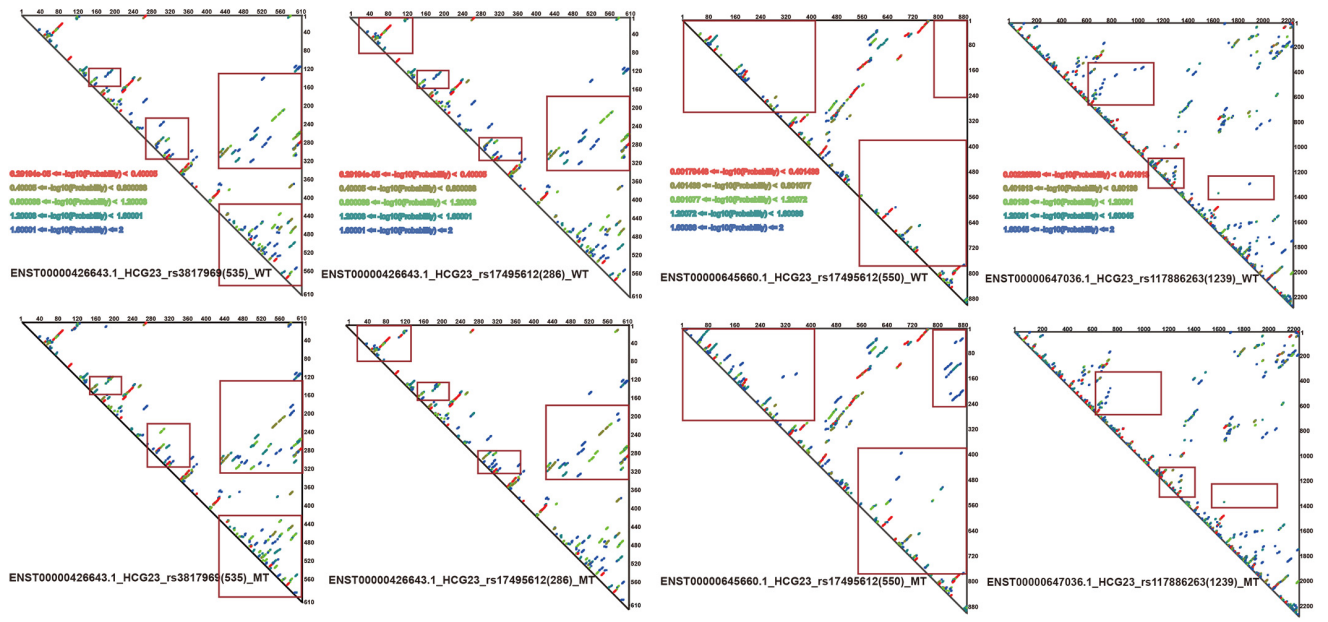


FIGURE 4 | Paired probabilities of WT and MT lncRNA transcripts. Changes in paired probabilities of WT and MT lncRNA transcripts induced by LD SNPs are labeled with red squares.

they exhibited a coherent trend for every item. The RS1 provided an approach to restrict the constitution of each base in lncRNA transcripts; hence, the RS1 was much stricter than the RS2. To ensure reliability of data, we chose common items for evaluation of significance. The base pair probabilities of the four significant WT and MT lncRNA transcripts are shown in **Figure 4**. These lncRNAs were significantly changed by SNPs, as determined by RS1 and RS2 quantitative analyses. **Figure 4** illustrates that a majority of SNPs in lncRNA transcripts only had small effects. Additionally, SNP-induced structural rearrangements often only existed locally (labeled within the red box in **Figure 4**), rather than affecting overall lncRNA architecture.

Probing Combined Effects of Multiple SNPs

Comprehensive annotations of SNPs from the 1000 Genomes Project and PLINK toolkit made it possible to predict combined effects of multiple SNPs. Among the 115 items, there were 41 SNPs located in 10 lncRNAs. Meanwhile, only 3 interactions between lncRNAs and SNPs exhibited a one-to-one relationship. This phenomenon suggests that SNP-mediated changes in lncRNA structure are affected by the combined effects of multiple mutation sites. In order to evaluate structural changes induced by multiple SNPs, we mapped SNPs within one lncRNA transcript and predicted possible LD blocks using PLINK. Ultimately, 44 haplotype blocks existed in seven unique lncRNA transcripts. We quantified the overall structural effect of multiple SNPs within one lncRNA transcript by computing the RNAsmc score. The resultant haplotype blocks, RNAsmc scores, and *P* values are presented in **Supplementary Table S5**. We found that 34 haplotype blocks had an impact on the secondary structures of lncRNA transcripts; however, 10 haplotype blocks had no impact. In addition, we evaluated the significance of lncRNA conformational changes induced by multiple SNPs. Only one haplotype block in HCG23 had a significant effect on lncRNA secondary structure. This haplotype block included 10 SNPs (rs117130854, rs115303880, rs17495612, rs60538826, rs149171231, rs146487240, rs549096164, rs561411181, rs117026188, and rs3817969) mapped onto ENST00000426643.1 (one of the HCG23 transcripts). This result illustrates that a majority of haplotype blocks had only subtle or negligible effects on lncRNA secondary structure. Hence, we inferred that the destructive power of large-span haplotype blocks was very little. In addition, these results demonstrate that the frequency of multiple simultaneous SNP mutations was low.

Scanning SNP-Mediated Disturbances in Molecular Combined Abilities

We identified four SNPs in HCG23 (including four lncRNA transcripts that significantly affected lncRNA secondary structures and that were associated with adenocarcinoma of the lung). Upon searching published papers and the LncMap database, we identified that five TFs—DDX17, STAT1, PPARG, ETS1, and E2F6—were closely associated with adenocarcinoma of the lung and HCG23. Four of these TFs (DDX17, STAT1, PPARG, and ETS1) have previously been verified to interact with other molecules or to participate in specific signaling pathways (Li et al., 2017; Sun et al., 2017; To et al., 2018; Yang et al., 2019). However, only over-expression of E2F6 has been associated with the development of adenocarcinoma of the lung. Next, we

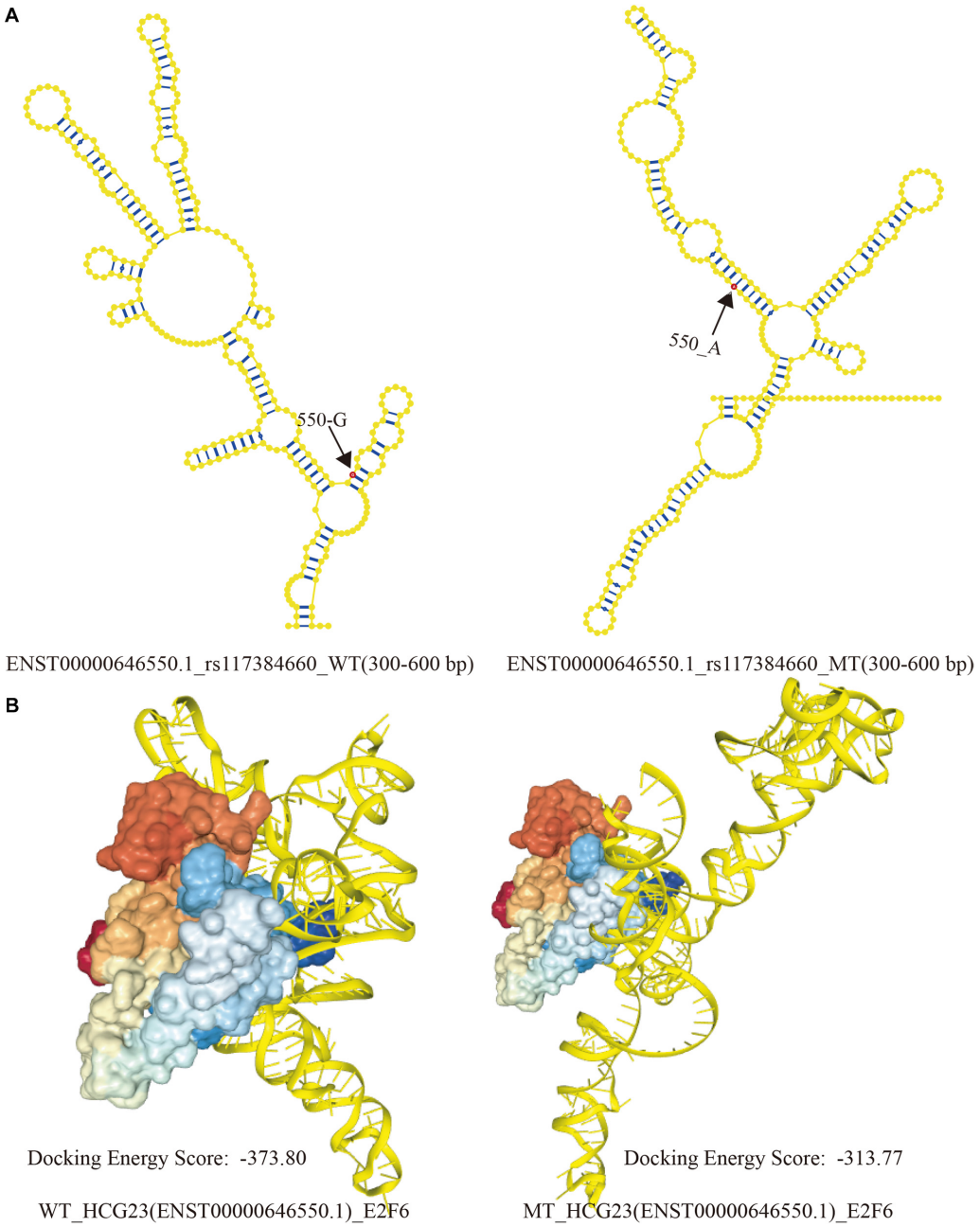


FIGURE 5 | Local structural visualization and prediction of molecular binding. **(A)** The local secondary structures of WT and MT HCG23 (ENST00000646550.1) induced by rs117384660. Bases in red showed the SNP sites in lncRNA transcript, and the numbers indicated location of SNPs. **(B)** Predictions of the structural conformations in interactive regions between WT and MT HCG23 and E2F6 induced by LD SNPs using HDOCK (HDOCK: http://hdock.phys.hust.edu.cn/).

analyzed how perturbations of HCG23 altered binding affinities and structural regulation (Barh et al., 2013). Using catRAPID, the interactive regions between four WT and MT lncRNA transcripts and E2F6 were predicted. Among these predictions, one combination of rs117384660 in ENST00000646550.1 of E2F6 led to noteworthy diversity.

The corresponding intervals of WT and MT lncRNAs were 49–102 bp and 301–352 bp, respectively. However, unique intervals (524–576 bp) arose when base G become A at position 550 of ENST00000646550.1. Based on previous study, Wang et al. proposed that local structural units could be formed within 150–300 bp in a lncRNA transcripts. Then, taking account of binding region predicted by CatRAPID, the interactive region must contain a range of 524–576 bp in lncRNA transcripts. And The SNP of 550 base was exactly located in this region. Therefore, 300 bp (300–600 bp) of the ENST00000646550.1 sequence was

chosen to represent the spatial combination with E2F6, as a result of the limitation of RNAComposer. The local secondary structures (300 bp) of WT and MT ENST00000646550.1 are shown in **Figure 5A**. The visualization of interactive regions was realized by HDOCK. In **Figure 5B**, we found that the docking score was intuitively distinct. Additionally, when E2F6 was kept at the same angle, the conformations and binding sites varied greatly between WT to MT HCG23. This result suggests that LD SNPs affected the structures of lncRNA transcripts and their abilities to bind to corresponding TFs, which may contribute to the occurrence and development of adenocarcinoma of the lung.

DISCUSSION

In the present study, we identified LD SNPs by enlarging lung-disease-associated SNPs. We also determined the positions of LD SNPs within lncRNAs, which provided a foundation for establishing the regulatory relationships of LD SNPs and lncRNAs in lung diseases. The LD SNPs in seven different HCG23 transcripts accounted for approximately 97.39% of all analyzed items (**Supplementary Table S1**). As we known, HCG23 locates at 6p21.32, the HLA locus that is known to be highly enriched for nucleotide polymorphism. Therefore, we developed a strict evaluation system, and set threshold to quantify HCG23 structural heterogeneity induced by single nucleotide mutations. The significance of structural heterogeneity was estimated by RS1 and RS2. RS1 performed 10,000 permutations of the flanking sequences. The *P* value, defined as the Random Score 1 (RS1), was determined by the order of real RNAsmc scores among random scores. However, RS2 mutated each base into three other bases and obtained all of the possible 3N mutations. The significance of scores was computed between the WT sequence and all of the background sequences. Ultimately, only a little SNPs result in significant changes in the structure of lncRNA transcripts. Meanwhile, they might have influence on expression or other functions. These results revealed that HCG23 on chromosome 6 plays a major role in adenocarcinoma of the lung. And according to previous study, HCG23 was also supported participating immune-related diseases (Debiec et al., 2018).

Our analysis of the effects of lung-disease-associated human genetic variation in lncRNAs revealed the extent to which specific SNPs affected lncRNA structure. The RNAsmc score is an algorithm that takes into account the secondary structure of each WT and MT lncRNA. In our present study, 73.91% of SNPs altered the lncRNA structural ensemble. However, we found that a majority of these SNPs exhibited only small or negligible effects on lncRNA structure (Halvorsen et al., 2010; Wan et al., 2014; Zhou et al., 2018). In contrast, only four SNPs had a significant effect on three lncRNA transcripts. These present results are consistent with those of previous studies. In addition, we analyzed the expression of HCG23 which included four significant SNPs. The expression profile in lung adenocarcinoma was derived from The Atlas of ncRNA in Cancer Database (TANRIC) (Li et al., 2015). Using R package-limma, the expression of HCG23 existed significant difference ($P < 0.05$) between normal and lung adenocarcinoma patient. This result can also demonstrated that SNPs affected not only lncRNA second structure, but also gene expression level.

The impact of allelic variants can be determined by analyzing the position and LD block of the associated SNP within an lncRNA sequence; SNPs not only affect gene expression, but they also influence secondary structure (Castellanos-Rubio and Ghosh, 2019). In addition, a previous study demonstrated that a single SNP could alter RNA conformation (Sharma et al., 2019). A similar behavior has been observed for haplotype blocks, the majority of which influence secondary structures of lncRNA transcripts. However, only one analyzed haplotype block significantly affected lncRNA transcripts in our present study. Our results also suggested that LD blocks were not formed by assigning alleles of SNPs randomly, and groups of these LD blocks obeyed specific rules to ensure molecular stability. Hence, we speculate that such conservative metabolic mechanisms for maintaining molecular structure/function may confer self-protection for each individual.

To ascertain whether structural changes affect protein binding, we predicted interactive regions of WT and MT HCG23 with E2F6 using CatRAPID. Compared with that of WT HCG23, MT HCG23 had a distinctive region (524–576 bp). Additionally, we found that binding sites of lncRNAs and proteins changed dramatically (**Figure 5**). This finding suggests that few LD SNPs inducing structural variation affect protein binding with lncRNAs. Furthermore, structural rearrangement of lncRNAs may contribute to regulation of transcription and/or post-transcription, and contribute to lung diseases.

Structural rearrangements of RNAs play crucial roles in adenocarcinoma of the lung. Rs114020893 in NEXN-AS1 has been predicted to change secondary structure and may contribute to lung cancer susceptibility (Yuan et al., 2016). Additionally, a novel ROS1-ADGRG6 rearrangement induced by the fusion of exons 1–33 of ROS1 on chr6 to exons of 2–26 of ADGRG6 on chr6 has been previously reported in lung cancer (Xu et al., 2019). Therefore, it is important to further elucidate the intricate regulatory mechanisms of disease-associated lncRNAs. Although large numbers of mutations exist within lncRNAs, the mechanisms of such mutations remain unclear. However, the interpretation of non-protein-coding mutations will become more accurate as experimental and computational methods improve.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions in this study will be made available by the authors upon reasonable requests.

AUTHOR CONTRIBUTIONS

LX and HW designed the overall concept of the study. XL processed data and wrote the manuscript. YD constructed the graphs. HW and all of the other authors revised the manuscript. All of the authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant numbers 31801098 and 31501062), the China Postdoctoral Science Foundation Project (grant number NO.62), the Key Research and Development Program of Zhejiang Province (grant number 2020C03036), the Fundamental Research Funds for the Provincial Universities (grant number 2017JCZX50), and the Internal Fund Project of Eye Hospital of Wenzhou Medical University (grant numbers YJGG20181001 and KYQD201901010.

ACKNOWLEDGMENTS

The authors wish to thank all of the members of their lab. Additionally, the authors also appreciate the support from the Training Center for Students Innovation and Entrepreneurship Education, Harbin Medical University (Harbin 150081, China), and the School of Ophthalmology and Optometry, Eye Hospital, and School of Biomedical Engineering at Wenzhou Medical University (Wenzhou 325027, China).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.00242/full#supplementary-material

FIGURE S1 | The robust analysis of RNAsmc score. Blue and red lines indicate the differences between WT and MT lncRNA transcripts, respectively.

FIGURE S2 | The permutation rules and significance evaluating of RS1, RS2, respectively. **(A)** Permutating flanking sequence in random way, and keeping WT and MT SNP allele unchanged. The sequence in blue shadow represented the changed content in RS1. **(B)** Building the background distribution for RS2. The base in red shadow showed it altered in each iteration. **(C)** The significance of lncRNA structural heterogeneity using RS1 and RS2. Red dot indicated that the changes were significant both in RS1 and RS2.

TABLE S1 | LD SNPs mapping on lncRNA transcripts.

TABLE S2 | The analysis of lncRNA structural heterogeneity induced by SNPs.

TABLE S3 | The significance of lncRNA structural heterogeneity by RS1.

TABLE S4 | The significance of lncRNA structural heterogeneity by RS2.

TABLE S5 | Quantifying and assessing the structural heterogeneity of lncRNA transcripts indeced by haplotypes.

REFERENCES

Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., and Tartaglia, G. G. (2013). catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics* 29, 2928–2930. doi: 10.1093/bioinformatics/btt495

Barh, D., Jain, N., Tiwari, S., Field, J. K., Padin-Iruegas, E., Ruibal, A., et al. (2013). A novel in silico reverse-transcriptomics-based identification and blood-based validation of a panel of sub-type specific biomarkers in lung cancer. *BMC Genomics* 14(Suppl. 6):S5. doi: 10.1186/1471-2164-14-S6-S5

Biesiada, M., Purzycka, K. J., Szachniuk, M., Blazewicz, J., and Adamiak, R. W. (2016). Automated RNA 3D Structure Prediction with RNAComposer. *Methods Mol. Biol.* 1490, 199–215. doi: 10.1007/978-1-4939-6433-8_13

Castellanos-Rubio, A., and Ghosh, S. (2019). Disease-associated SNPs in inflammation-related lncRNAs. *Front. Immunol.* 10:420. doi: 10.3389/fimmu.2019.00420

Chen, Z., Lei, T., Chen, X., Gu, J., Huang, J., Lu, B., et al. (2019). Long non-coding RNA in lung cancer. *Clin. Chim. Acta* 504, 190–200. doi: 10.1016/j.cca.2019.11.031

Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25, 1974–1975. doi: 10.1093/bioinformatics/btp250

Debiec, H., Dossier, C., Letouze, E., Gillies, C. E., Vivarelli, M., Putler, R. K., et al. (2018). Transethnic, genome-wide analysis reveals immune-related risk alleles and phenotypic correlates in pediatric steroid-sensitive nephrotic syndrome. *J. Am. Soc. Nephrol.* 29, 2000–2013. doi: 10.1681/ASN.2017111185

Eeles, R. A., Olama, A. A., Benlloch, S., Saunders, E. J., Leongamornlert, D. A., Tymrakiewicz, M., et al. (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.* 45, 385–391. doi: 10.1038/ng.2560

Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., et al. (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* 48, 500–509. doi: 10.1038/ng.3547

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Halvorsen, M., Martin, J. S., Broadaway, S., and Laederach, A. (2010). Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* 6:e1001074. doi: 10.1371/journal.pgen.1001074

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111

Hua, J. T., Ahmed, M., Guo, H., Zhang, Y., Chen, S., Soares, F., et al. (2018). Risk SNP-mediated promoter-enhancer switching drives prostate cancer through lncRNA PCAT19. *Cell* 174:564-575.e18. doi: 10.1016/j.cell.2018.06.014

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., et al. (2015). TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.* 75, 3728–3737. doi: 10.1158/0008-5472.CAN-15-0273

Li, K., Mo, C., Gong, D., Chen, Y., Huang, Z., Li, Y., et al. (2017). DDX17 nucleocytoplasmic shuttling promotes acquired gefitinib resistance in non-small cell lung cancer cells via activation of beta-catenin. *Cancer Lett.* 400, 194–202. doi: 10.1016/j.canlet.2017.02.029

Li, Y., Li, L., Wang, Z., Pan, T., Sahni, N., Jin, X., et al. (2018). LncMAP: pan-cancer atlas of long noncoding RNA-mediated transcriptional network perturbations. *Nucleic Acids Res.* 46, 1113–1123. doi: 10.1093/nar/gkx1311

Loewen, G., Jayawickramarajah, J., Zhuo, Y., and Shan, B. (2014). Functions of lncRNA HOTAIR in lung cancer. *J. Hematol. Oncol.* 7:90. doi: 10.1186/s13045-014-0090-4

Martin, J. S., Halvorsen, M., Davis-Neulander, L., Ritz, J., Gopinath, C., Beauregard, A., et al. (2012). Structural effects of linkage disequilibrium on the transcriptome. *RNA* 18, 77–87. doi: 10.1261/rna.029900.111

Pirooz, H. J., Jafari, N., Rastegari, M., Fathi-Roudsari, M., Tasharrofi, N., Shokri, G., et al. (2018). Functional SNP in microRNA-491-5p binding site of MMP9 3'-UTR affects cancer susceptibility. *J. Cell. Biochem.* 119, 5126–5134. doi: 10.1002/jcb.26471

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Ramírez-Bello, J., and Jiménez-Morales, M. (2017). [Functional implications of single nucleotide polymorphisms (SNPs) in protein-coding and non-coding RNA genes in multifactorial diseases]. *Gac. Med. Mex.* 153, 238–250.

Reuter, J. S., and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11:129. doi: 10.1186/1471-2105-11-129

Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Gross, M., et al. (2019). A pan-cancer analysis of synonymous mutations. *Nat. Commun.* 10:2569. doi: 10.1038/s41467-019-10489-2

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81. doi: 10.1038/nature15394

Sun, Q., Jiang, C. W., Tan, Z. H., Hou, L. Y., Dong, H., Liu, K., et al. (2017). MiR-222 promotes proliferation, migration and invasion of lung adenocarcinoma cells by targeting ETS1. *Eur. Rev. Med. Pharmacol. Sci.* 21, 2385–2391.

To, K. K. W., Wu, W. K. K., and Loong, H. H. F. (2018). PPARgamma agonists sensitize PTEN-deficient resistant lung cancer cells to EGFR tyrosine kinase inhibitors by inducing autophagy. *Eur. J. Pharmacol.* 823, 19–26. doi: 10.1016/j.ejphar.2018.01.036

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505, 706–709. doi: 10.1038/nature12946

Wang, C., Li, Y., Li, Y. W., Zhang, H. B., Gong, H., Yuan, Y., et al. (2018). HOTAIR lncRNA SNPs rs920778 and rs1899663 are associated with smoking, male gender, and squamous cell carcinoma in a Chinese lung cancer population. *Acta Pharmacol. Sin.* 39, 1797–1803. doi: 10.1038/s41401-018-0083-x

Wang, H., Zheng, H., Wang, C., Lu, X., Zhao, X., and Li, X. (2017). Insight into HOTAIR structural features and functions as landing pads for transcription regulation proteins. *Biochem. Biophys. Res. Commun.* 485, 679–685. doi: 10.1016/j.bbrc.2017.02.100

Wong, K. M., Langlais, K., Tobias, G. S., Fletcher-Hoppe, C., Krasnewich, D., Leeds, H. S., et al. (2017). The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. *Nucleic Acids Res.* 45, D819–D826. doi: 10.1093/nar/gkw1139

Xu, S., Wang, W., Xu, C., Li, X., Ye, J., Zhu, Y., et al. (2019). ROS1-ADGRG6: a case report of a novel ROS1 oncogenic fusion variant in lung adenocarcinoma and the response to crizotinib. *BMC Cancer* 19:769. doi: 10.1186/s12885-019-5948-y

Yan, Y., Zhang, D., Zhou, P., Li, B., and Huang, S. Y. (2017). HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res.* 45, W365–W373. doi: 10.1093/nar/gkx407

Yang, J., Liu, Y., Mai, X., Lu, S., Jin, L., and Tai, X. (2019). STAT1-induced upregulation of LINC00467 promotes the proliferation migration of lung adenocarcinoma cells by epigenetically silencing DKK1 to activate Wnt/beta-catenin signaling pathway. *Biochem. Biophys. Res. Commun.* 514, 118–126. doi: 10.1016/j.bbrc.2019.04.107

Yang, J., and Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* 43, W174–W181. doi: 10.1093/nar/gkv342

Yuan, H., Liu, H., Liu, Z., Owzar, K., Han, Y., Su, L., et al. (2016). A novel genetic variant in long non-coding RNA gene NEXN-AS1 is associated with risk of lung cancer. *Sci. Rep.* 6:34234. doi: 10.1038/srep34234

Zhou, Z. Y., Hu, Y., Li, A., Li, Y. J., Zhao, H., Wang, S. Q., et al. (2018). Genome wide analyses uncover allele-specific RNA editing in human and mouse. *Nucleic Acids Res.* 46, 8888–8897. doi: 10.1093/nar/gky613

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 © 2020 Lu, Ding, Bai, Li, Zhang, Wang, Gao, Xu 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 Genetics

ORIGINAL RESEARCH
published: 15 April 2020
doi: 10.3389/fgene.2020.00346

Differential DNA Methylation Encodes Proliferation and Senescence Programs in Human Adipose-Derived Mesenchymal Stem Cells

Mark E. Pepin[1†], Teresa Infante[2†], Giuditta Benincasa[2†], Concetta Schiano[2], Marco Miceli[3], Simona Ceccarelli[4], Francesca Megiorni[4], Eleni Anastasiadou[4], Giovanni Della Valle[5], Gerardo Fatone[5], Mario Faenza[6], Ludovico Docimo[7], Giovanni F. Nicoletti[6], Cinzia Marchese[4], Adam R. Wende[1] and Claudio Napoli[3,7*]*

[1] Department of Pathology, Division of Molecular & Cellular Pathology, University of Alabama at Birmingham, Birmingham, AL, United States, [2] Department of Advanced Clinical and Surgical Sciences, University of Campania Luigi Vanvitelli, Naples, Italy, [3] IRCCS SDN, Naples, Italy, [4] Department of Experimental Medicine, Sapienza University of Rome, Rome, Italy, [5] Department of Veterinary Medicine and Animal Productions, University of Naples Federico II, Naples, Italy, [6] Multidisciplinary Department of Medical, Surgical and Dental Sciences, Plastic Surgery Unit, University of Campania Luigi Vanvitelli, Naples, Italy, [7] Clinical Department of Internal Medicine and Specialistics, University of Campania Luigi Vanvitelli, Naples, Italy

Adult adipose tissue-derived mesenchymal stem cells (ASCs) constitute a vital population of multipotent cells capable of differentiating into numerous end-organ phenotypes. However, scientific and translational endeavors to harness the regenerative potential of ASCs are currently limited by an incomplete understanding of the mechanisms that determine cell-lineage commitment and stemness. In the current study, we used reduced representation bisulfite sequencing (RRBS) analysis to identify epigenetic gene targets and cellular processes that are responsive to 5′-azacitidine (5′-AZA). We describe specific changes to DNA methylation of ASCs, uncovering pathways likely associated with the enhancement of their proliferative capacity. We identified 4,797 differentially methylated regions (FDR < 0.05) associated with 3,625 genes, of which 1,584 DMRs annotated to the promoter region. Gene set enrichment of differentially methylated promoters identified "phagocytosis," "type 2 diabetes," and "metabolic pathways" as disproportionately hypomethylated, whereas "adipocyte differentiation" was the most-enriched pathway among hyper-methylated gene promoters. Weighted coexpression network analysis of DMRs identified clusters associated with cellular proliferation and other developmental programs. Furthermore, the ELK4 binding site was disproportionately hyper-methylated within the promoters of genes associated with AKT signaling. Overall, this study offers numerous preliminary insights into the epigenetic landscape that influences the regenerative capacity of human ASCs.

Keywords: Whole-genome DNA methylation, stem cell biology, regenerative medicine, computational biology, 5′-azacitidine, epigenomics and epigenetics, cellular reprogramming

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Nady El Hajj,
Hamad bin Khalifa University, Qatar
Hongyi Zhang,
The University of Texas Southwestern Medical Center, United States

***Correspondence:**
Mark E. Pepin
pepinme@gmail.com
Claudio Napoli
direzione.immunoematologia@unicampania.it

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Epigenomics and Epigenetics, a section of the journal Frontiers in Genetics

Received: *03 December 2019*
Accepted: *23 March 2020*
Published: *15 April 2020*

Citation:
Pepin ME, Infante T, Benincasa G, Schiano C, Miceli M, Ceccarelli S, Megiorni F, Anastasiadou E, Della Valle G, Fatone G, Faenza M, Docimo L, Nicoletti GF, Marchese C, Wende AR and Napoli C (2020) Differential DNA Methylation Encodes Proliferation and Senescence Programs in Human Adipose-Derived Mesenchymal Stem Cells. Front. Genet. 11:346. doi: 10.3389/fgene.2020.00346

INTRODUCTION

Adipose-derived mesenchymal stem cells (ASCs) represent a population of self-renewing and multipotent cells that reside in the vascular stroma of adipose tissue; under specific conditions, they are capable of differentiating into cellular phenotypes that resemble adipocytes, myocytes, chondrocytes, and osteocytes (Patrikoski et al., 2019). This indigenous cell population is known to play a central role in embryologic development, tissue growth, tissue repair, and regeneration due to its pluripotency and immunomodulatory capacity (Scuderi et al., 2013; Dai et al., 2016; Tabatabaei Qomi and Sheykhhasan, 2017). However, its role in end-organ homeostasis and repair following organ damage remains largely undefined.

The central aim of regenerative medicine is to restore function to pathological tissues via cellular regeneration and reimplantation. Although recent advances in gene therapy have enabled researchers to modify the human genome for therapeutic purposes, its utility is limited by the irreversible nature gene editing (Bates, 2016). Furthermore, most human diseases cause myriad transcriptional effects, such that targeting a subset of genomic loci is likely insufficient to successfully regenerate function (Hong, 2018). Regenerative medicine thus turns also to the procurement of adult mesenchymal cells, including ASCs, for *in vitro* expansion, differentiation into end-organ tissues, and eventual reimplantation (Shyam et al., 2017).

Among the biologic mechanisms capable of influencing cellular plasticity, epigenomic modification has been shown to impact both the regenerative capacity and eventual differentiation of adult stem cells (Ceccarelli et al., 2018; Costantino et al., 2018). These molecular marks, including methylation of cytosine residues, have been described as critical features in determining the process of ASC aging and senescence (Munoz-Najar and Sedivy, 2011). Identification of epigenetically active small molecules, termed epidrugs, has enable the exogenous manipulation of DNA methylation to define both the transcriptional and phenotypic components under direct epigenetic control, targeting gene expression in a transcriptome-wide manner.

5′-azacitidine (5′-AZA) is one such epigdrug, that disrupts the methylation of cytosine when incorporated into the newly synthesized DNA of progeny cells (Sajadian et al., 2015). Its use *in vitro* has been shown to trans-differentiate non-osteoblastic cells into an osteocytic lineage (Cho et al., 2014). 5′-AZA has also been shown to attenuate aging-associated impairments in proliferation of ASCs (Kornicka et al., 2017). Similarly, reduction of genome-wide DNA methylation has been linked to enhancement in their self-renewal (Kornicka et al., 2017).

In the current study, we investigate how 5′-AZA-induced alterations in genome-wide DNA methylation likely affect molecular networks involved in proliferation and slowing of senescence processes in ASCs. We use reduced-representation bisulfite sequencing (RRBS) to localize differentially methylated regions (DMRs) and identify candidate methylation-sensitive transcriptional regulators of ASC senescence and proliferation.

Abbreviations: 5′-AZA, 5′-azacitidine; AKT, protein kinase B; ASCs, Adipose-derived mesenchimal stem cells; DMR, differentially methylated 500-base CpG-containing region; ELK4, ETS transcription factor.

MATERIALS AND METHODS

Ethics Statement

Human subcutaneous abdominal adipose tissue was collected from abdominal wall resection of two healthy subjects who underwent cosmetic surgery, male and female Caucasian Italians (BMIs < 25) ages 62 and 52 years, respectively. Written informed consent was obtained, and the clinical protocol was approved by the Institutional Review Board of the Department of Experimental Medicine, Sapienza University of Rome (Italy). All human genome-wide DNA methylation data have been

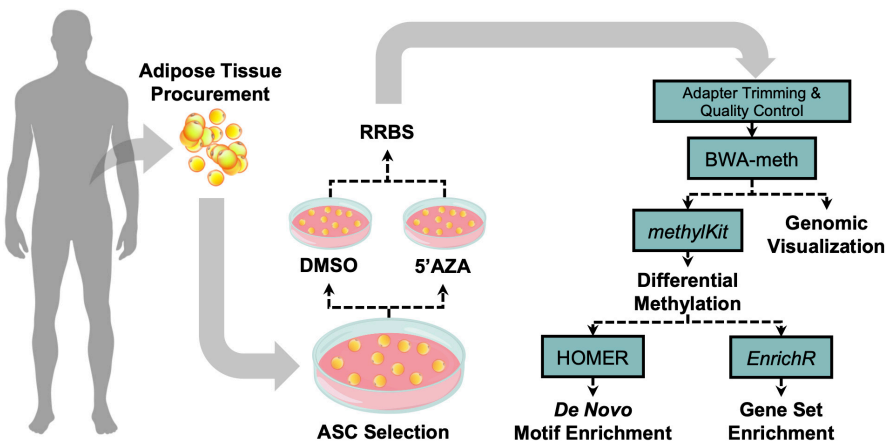


FIGURE 1 | Visual illustration of sample processing and data analysis pipeline. Abdominal white adipose tissue was biopsied from two patients, and subsequently plated for adipose-derived mesenchymal cell selection and treatment with 5′-AZA or vehicle (DMSO) for 24 h. Isolated DNA was bisulfite-reduced and sequenced, followed by bioinformatic analysis of differential CpG methylation.

uploaded to the NCBI Gene Expression Omnibus database (GSE139157): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139157.

ASC Isolation and Culture

The tissue procurement, cellular treatment, and data analysis pipeline was performed as illustrated in **Figure 1**. Adipose tissue was transferred to the laboratory and processed under sterile conditions within 24 h. Isolation of ASCs was performed as previously described (Ceccarelli et al., 2018). Briefly, tissue fragments were washed extensively with sterile phosphate-buffered saline containing 2% penicillin/streptomycin and minced. The extracellular matrix was digested with 0.075% collagenase type I for 30–60 min at 37°C and 5% CO_2. The suspension was filtered to remove debris and centrifuged for 5 min at 2000 rpm. The pellets of stromal vascular fraction (SVF) containing ASCs were washed with PBS, then resuspended in the culture medium and transferred to a culture flask. ASCs were self-selected out of the SVF, since they were adherent to the plastic tissue cultureware. ASC cells were cultured in DMEM-Ham's F-12 medium (vol/vol, 1:1) (DMEM/F12; Gibco) supplemented with 10% FBS, 100 U/ml penicillin, 100 mg/ml streptomycin, and 2 mM L-glutamine, and maintained in a 5% CO_2 incubator at 37°C in a humidified atmosphere, with medium change twice a week. When reaching 80–90% confluence, cells were detached with 0.5 mM EDTA/0.05% trypsin (Gibco) for 5 min at 37°C and then replated. ASCs were expanded and cell viability was assessed by using the trypan blue exclusion assay. Cell morphology was evaluated by phase contrast microscopy. Experiments were conducted between passage numbers 7,8. Absence of mycoplasma contamination was confirmed by PCR with specific primers.

Treatment of ASC With 5′-Azacytidine (5′-AZA)

The DNA methyltransferase inhibitor 5′-AZA was purchased from Sigma-Aldrich (Milan, Italy) and was reconstituted at 10 mM using dimethyl sulfoxide (DMSO). ASCs were seeded at a cell density of 5×10^3 cells/cm^2. After cell attachment, the medium was changed to freshly made culture medium containing 10 μM 5′-AZA. After 24 h, the 5′-AZA containing culture medium was refreshed for another 24 h (total treatment 48 h). Control samples were treated with DMSO alone at 0.1% (v/v) concentration.

DNA Extraction

Pooled genomic DNA was isolated from ASCs treated with or without 24-h of 10 μM 5′-AZA using the GeneJET Genomic DNA Purification Kit (Thermo Scientific). DNA concentration and purity were determined using Qubit 3 Fluorometer

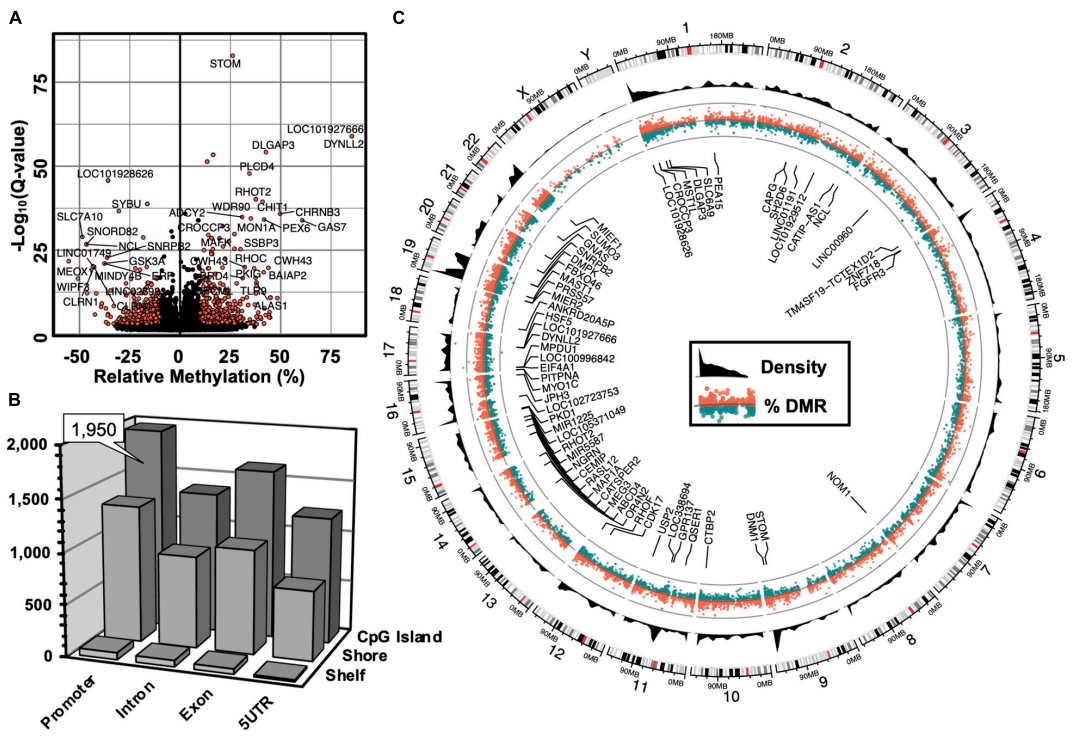


FIGURE 2 | Genome-wide alterations in DNA methylation. **(A)** Volcano plot illustrating -$\log_{10}$(Q-value) as a function of differentially methylated regions (DMRs, in % difference). Genes were labeled which contained robust changes in methylation ($Q < 10^{-15}$), | Methylation| > 25%). **(B)** 3-dimensional bar plot depicting the distribution of DMRs according to both genic annotation (promoter, intron, exon, and 5′ untranslated regions) and proximity to CpG Islands (CpG Island, Shore, and Shelf). **(C)** Circular plot illustrating the genomic distribution of DMRs (red = hyper-methylated, green = hypo-methylated), and DMR density (black). Labeled genes are DMRs (| Methylation| > 10%) that meet genome-wide significance ($Q < 10^{-8}$).

(Invitrogen) and NanoDrop 2000 (Thermo Scientific). DNA integrity was checked on 1% agarose gel.

Reduced-Representation Bisulfite Sequencing

Sequencing was performed at the Genomix4Life S.r.l. with subsequent bioinformatics performed at the University of Alabama at Birmingham (United States). Briefly, 2 μg of genomic DNA were used for each library preparation. Each DNA sample was digested by MspI restriction enzyme. The digested products were purified with the GeneJet PCR Purification Kit (Thermo Fisher Scientific) and libraries were prepared by TruSeq Library Prep Kit (Illumina, Inc., United States). Fragments were bisulfite converted using the EZ DNA Methylation-Gold Kit (Zymo Research, United States). The converted DNA was amplified using PfuTurbo Cx Hotstart DNA Polymerase (Agilent Technologies, United States). The amplified fragments were purified by AMPure XP Beads and further quantified by the Agilent 4200 TapeStation (Agilent Technologies, United States). Each DNA library was analyzed by paired-end sequencing read (2 × 75 cycles) on Illumina Nextseq 500.

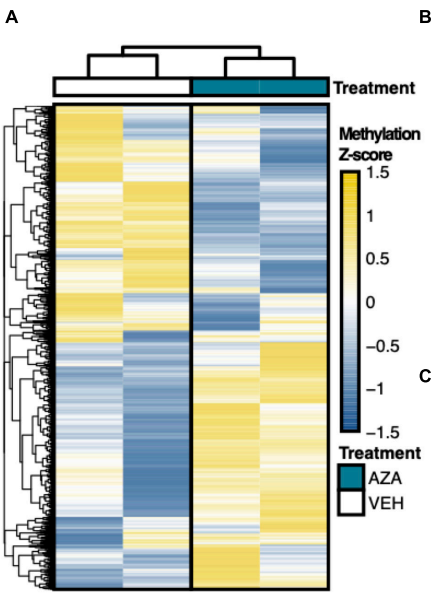
Bioinformatic Analysis and Data Visualization

RRBS Analysis

Details of the R coding scripts and other bioinformatics tools used in the current study are available as online **Supplemental Methods** and GitHub data repository: https://github.com/mepepin/Napoli_ASCs. To evaluate sequencing quality, *FastQC* (0.11.7) was used both before and after adapter trimming via *TrimGalore* (0.4.4). The bisulfite-reduced and sequenced reads were then aligned to the CT- and GA-converted human hg38 (GRCh38.p12) genome assembly via BWA-meth to quantify relative alignment of methylated and unmethylated CpGs, respectively (Sun et al., 2018). We then quantified differential DNA methylation in 500-base windows to exploit the regional CpG methylation analysis afforded by RRBS using the R package *methylKit* (1.8.0). Briefly, alignments were first filtered for those with <99.9% methylated CpGs and sequencing depth >10× to remove PCR-biased and low-coverage CpG sites, respectively. To determine regional methylation, 500-base genomic window was used to perform regional methylation in ASCs treated with 5′-AZA relative to vehicle-treated ASCs. Statistical significance of differential-methylation was assumed based on an over-correction adjusted Fisher's exact test, as recommended by Wreczycka et al. (2017). A sliding linear model (SLIM) method was used to adjust *P*-values for multiple testing (Wang et al., 2011).

Pathway Enrichment Analysis and Functional Network Mapping

Functional gene set enrichment analyses (GSEA) were performed using the interactive web-based platform *Enrichr* (Chen et al., 2013) using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Kanehisa, 2002). Weighted gene network analysis and visualization were performed using Cytoscape (3.7.0) using genes with differentially methylated promoters (DMPs) and a low-stringency statistical threshold of $P < 0.05$. Heatmap and hierarchical clustering generation was performed using *pheatmap* package (1.0.8) within R.



B

GO-Term Pathway	Overlap	P-value	Score
Phagocytosis	22/91	9.1×10^{-05}	22.1
Type II diabetes mellitus	12/46	1.7×10^{-03}	16.3
Endometrial cancer	14/58	1.7×10^{-03}	15.1
Choline metabolism	21/99	8.6×10^{-04}	14.7
carbon metabolism	15/65	1.9×10^{-03}	14.2
Bacterial invasion of epithelial cells	16/74	2.8×10^{-03}	12.5
Insulin Signaling	26/137	1.3×10^{-03}	12.4
Glioma	16/75	3.0×10^{-3}	12.0
Hepatocellular carcinoma	30/168	1.6×10^{-03}	11.3
Thyroid hormone Signaling	22/116	3.0×10^{-03}	10.8

C

GO-Term Pathway	Overlap	P-value	Score
White adipocyte differentiation	16/32	1.8×10^{-05}	32.0
Mammary gland development	15/32	8.7×10^{-05}	25.7
BDNF signaling	42/144	2.05×10^{-04}	14.5
White/Brown Adipocyte Differentiation	12/25	3.39×10^{-04}	22.5

FIGURE 3 | Reduced Representation Bisulfite Sequencing Analysis. **(A)** Hierarchical clustering and heatmap analysis of regional differential CpG methylation (DMRs).* **(B)** GO-term enrichment analysis of hyper-methylated and **(C)** hypo-methylated DMRs demonstrating the % enrichment ("Overlap"), *P*-value of overlap by Fischer exact test, and composite enrichment score.

Statistical Analysis

Unless otherwise indicated, statistical significance was determined via unpaired two-tailed Bonferroni-adjusted *P*-value (Q) < 0.05. Pathway analysis and motif enrichment both employed an unadjusted *P*-value using the Fisher's exact test.

RESULTS

Differential Methylation of ASCs Treated With 5′-AZA

The computation of differentially methylated regions (DMRs) identified 4,797 DMRs (Q-value < 0.05) associated with 3,625 annotated genes (**Supplementary Table S1**), among which was an equivalent proportion of hyper-methylated (2,588) and hypo-methylated (2,209) regions was uncovered. A volcano plot was used to examine the genes associated with differentially methylated promoters with 5′-AZA treatment (**Figure 2A**). Because the transcriptional effects of DNA methylation are highly dependent on the position of DMRs relative to known genomic features, with promoter methylation classically inversely associated with transcriptional activity (Bird, 1986; Jjingo et al., 2012), we examined the distribution of DMRs in relation to genomic locations most impacted by changes in DNA methylation following treatment with 5′-AZA. Annotated DMRs were mapped onto both genic location (promoter, 5′UTR, gene body, and 3′UTR) as well as according to their distance from established CpG Islands (CGIs). Methylation dynamics were modestly concentrated among CpG islands located in the promoter region (**Figure 2B**). Illustrating DMRs using a circular genome plot demonstrated the genome-wide distribution of methylation changes, numerous hyper-dynamic regions, and a broad array of genes with DMRs reaching genome-wide significance ($P < 10^{-8}$) (**Figure 2C**).

Pathway Enrichment Analysis of DMRs

Hierarchical clustering and heatmap visualization was then used to understand the degree of consistency between 5′-AZA and DMSO-treated ASCs, which revealed robust separation of differential promoter methylation between 5′-AZA and vehicle-treated ASCs with no apparent outliers (**Figure 3A**). Gene set enrichment analysis (GSEA) was performed using the genes possessing differentially methylated promoters, separating the hyper-methylated from hypo-methylated DMRs to interpret the possible impact of methylation dynamics on the enriched pathway(s). This approach identified phagocytosis ($P = 10^{-5}$), type 2 diabetes ($P = 10^{-4}$), and numerous metabolic pathways as disproportionately impacted by promoter hypo-methylation (**Figure 3B** and **Supplementary Table S2**), whereas white-adipocyte differentiation was the top most-enriched pathway among genes with hyper-methylated promoters (**Figure 3C** and **Supplementary Table S3**). Altogether, these findings suggest that 5′-AZA-mediated promoter methylation influences the metabolic phenotype of ASCs.

Weighted Functional Network Enrichment Analysis

To identify novel networks affected by 5′-AZA treatment, we used the STRING database (Szklarczyk et al., 2017) to generate a network of DMRs based on known and predicted interactions among their encoded proteins, with Markov Clustering (MCL) to separate DMR networks according

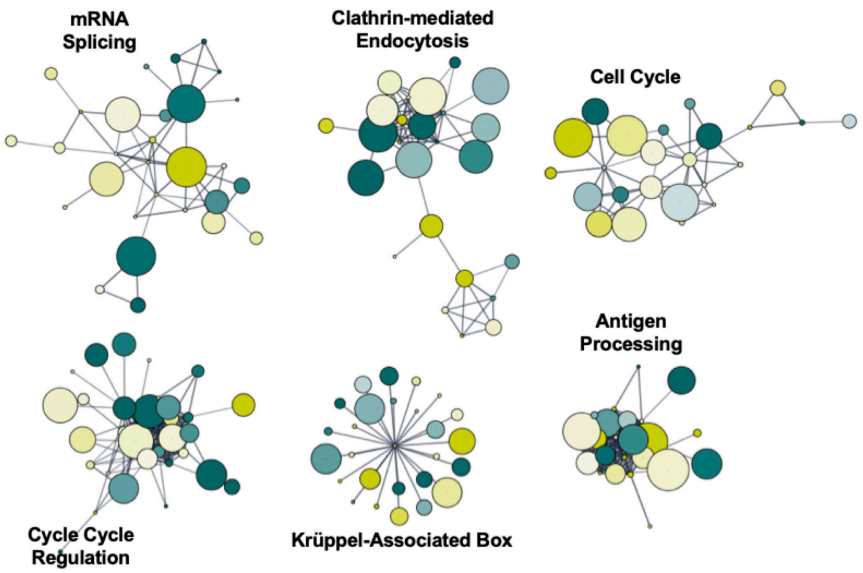


FIGURE 4 | Functional network enrichment analysis. Weighted gene co-expression network analysis of DMRs based on protein-protein interactions found within the STRING database. Network topology was defined according to both the degree of interaction and number of common nodes (minimum of 3) using a Markov cluster algorithm (MCL).

to both the degree and number of adjacent interactions (**Figure 4**). After ranking nodes by degree of interaction, the largest cluster was identified to functionally enrich pathways via the KEGG pathway database associated with cellular proliferation and developmental programs: krüppel-associated box ($P = 2.9 \times 10^{-32}$), antigen processing ($P = 6.9 \times 10^{-26}$), mRNA splicing ($P = 1.1 \times 10^{-23}$), cell cycle regulation ($P = 5.7 \times 10^{-17}$), clathrin-mediated endocytosis ($P = 1.2 \times 10^{-15}$), and cell cycle (BH-adjusted $P = 2.9 \times 10^{-13}$). Because we and others have previously found Krüppel-like factors (KLFs) to possess methylation-sensitive response elements (Pepin et al., 2019), the current analysis further supports that 5′-AZA treatment is sufficient to induce disproportionate differential methylation of KLF target gene promoters.

Motif Interference by DNA Hyper-Methylation

Promoter DNA methylation regulates gene expression in-part by interfering with the binding of transcription factors, particularly those possessing CpG-rich response elements (Medvedeva et al., 2014). Therefore, to identify candidate transcription factors likely disrupted by co-localizing promoter methylation, a *de novo* motif enrichment was performed via HOMER (Heinz et al., 2010). To correct for the CpG density of DMRs, background genomic regions were selected according to CpG content when performing hypergeometric enrichment. From this analysis, we found that *ZNF711* was significantly associated with hypo-methylated promoter DMRs ($Log_{10}(P) = -9.9$) (**Supplementary Figure S1A**), whereas the ELK4 motif disproportionately co-localized among hyper-methylated promoters (**Figure 5A**). The ENCODE dataset (GSE31477) was then used to identify genes associated with hyper-methylated DMRs ($Q < 0.05$, |Methylation| >10%) found in the promoter of ELK4 target genes, revealing (**Figure 5B**). Gene-set enrichment of co-localized DMRs within ELK4 target promoters identified AKT signaling as disproportinately affected by DMR co-localization, supporting prior studies demonstrating the role of ELK4 and HDAC (**Figure 5C**). A preliminary evaluation of ELK4 expression revealed its induction in ASC's treated with 5′AZA (**Supplementary Figure S1B**). Taken together, these observations support the hypothesis that 5′-AZA-induced promoter methylation may interfere with the downstream targets of ELK4 signaling and may therefore disrupt its role as a regulator of cellular proliferation (Boog, 1993).

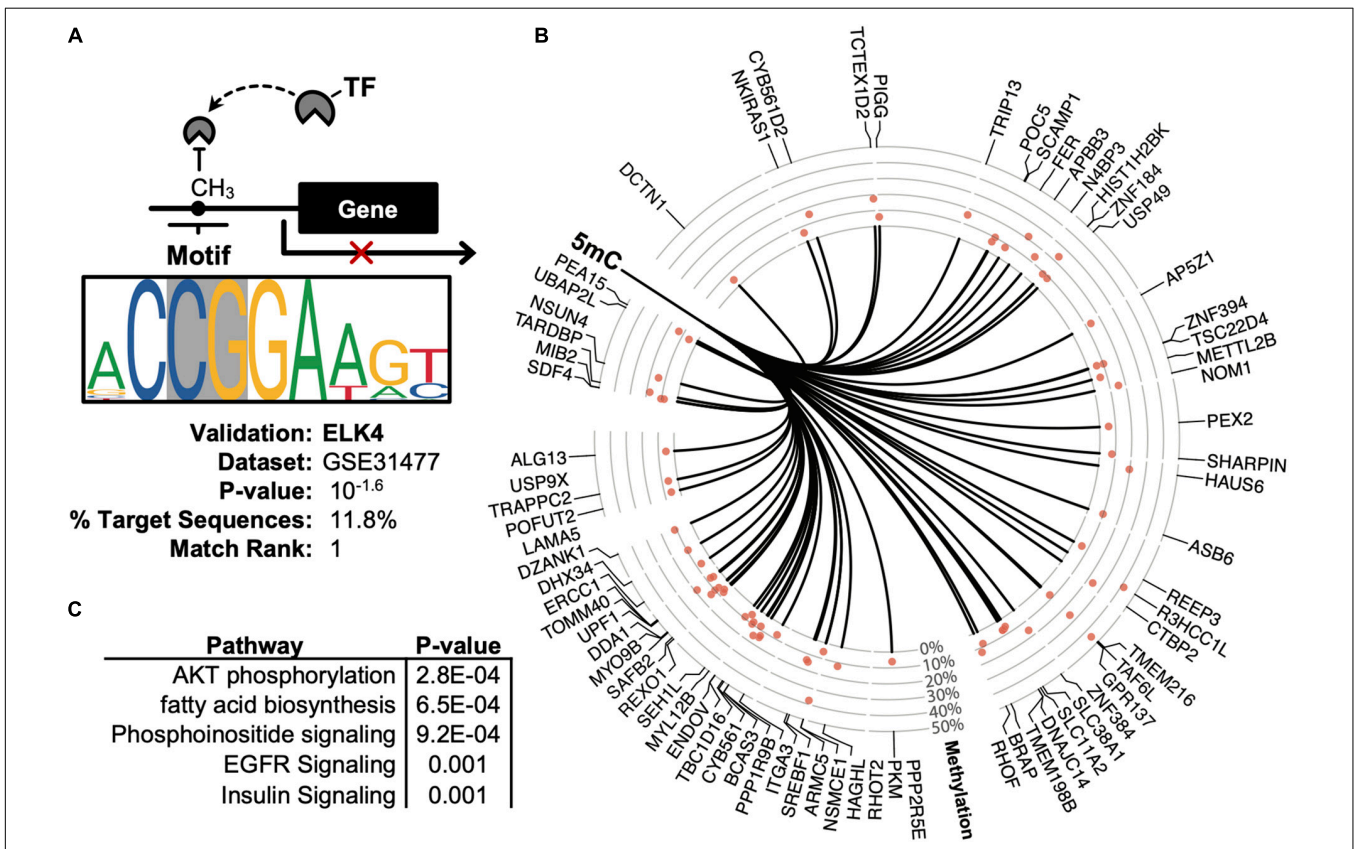


FIGURE 5 | Promoter-based DMR enrichment of known response elements. **(A)** Known motif enrichment of hyper-methylated DMRs localized to the promoter region based on the ENCODE database of ChIP-sequencing datasets. **(B)** Downstream ELK4 targets genes with hyper-methylated ($Q < 0.05$, | methylation| > 5%) promoters based on the ENCODE ChIP-sequencing dataset by Cayting et al. (GSE31477). **(C)** Gene-set enrichment analysis of ELK4 gene targets with hyper-methylated promoters, showing the top-5 disproportionately enriched pathways.

DISCUSSION

The epigenetic programs that define cellular identity offer fundamental insight into the mechanisms of cellular regeneration. These developmental programs, though once considered irreversible, have become a major focus of investigation as researchers seek to understand the pathogenesis of disease, tissue regeneration, and cellular aging. The field of regenerative medicine has begun to embrace a new treatment paradigm wherein senescent end-organ cells may be harvested from patients and re-programmed for therapeutic purposes. However, progress in regenerative medicine is currently limited by an incomplete understanding of the cellular machinery and molecular programs responsible for determining cell fate. In the current study, we employed RRBS to uncover an epigenomic program within ASC's which encodes proliferation and senescence in response to 5′-AZA.

Motif analysis of hyper-methylated DMRs uncovered the disproportionate methylation of promoter targets for ETS domain-containing transcription factor ELK4. As a regulator of numerous proto-oncogenes, ELK4 is believed to regulate malignant transformation in the context of multiple cancer types (Wu et al., 2018). Specifically, ELK4 has been shown to recruit and stabilize epigenetic regulator SIRT7 to promote tumor growth and maintain rapid cellular proliferation (Makkonen et al., 2008; Day et al., 2011). Therefore, we provide novel support that differential DNA methylation of the ELK4 response element modulates the downstream effects of ELK4 signaling.

Overall the current study has identified 4,797 genomic regions differentially methylated in 5′-AZA-treated relative to vehicle-treated ASCs. Furthermore, differential methylation was found to disproportionately affect promoter-associated CpG Islands, regions which have been extensively studied for their role as negative transcriptional regulators (Bird, 1986; Jjingo et al., 2012). Roughly half of these DMRs were hyper-methylated, and therefore cannot be explained by the direct effects of 5′-AZA, a potent inducer of DNA demethylation (Christman, 2002). Numerous potential mechanisms exist capable of mediating these indirect effects, including the compensatory regulation of DNA methyltransferase (DNMT) activity. Furthermore, 5′-AZA has been found to regulate DNA methylation in a targeted manner (Tabolacci et al., 2016). Such genomic restriction could involve the pre-templated epigenetic architecture present prior to 5′-AZA treatment; however, such notions remain speculative, warranting future studies to empirically define.

Although we provide numerous novel findings, we must acknowledge the key limitations of the current analysis. Although genome-wide approaches represent a valuable tool to identify novel candidates and genomic networks, our pooled analysis of ASCs yields a small sample size that limits the statistical generalizability of our DNA methylation analysis. Future work will involve single-cell analytical approaches to appreciate spectrum of response to 5′-AZA treatment. Additionally, future work is needed to correlate epigenomic changes with a cellular phenotype. Furthermore, we cannot exclude that other epigenetic phenomena influence adipocyte aging and proliferation. Lastly, validation and functional studies using larger patient cohorts are needed to assess the *in vivo* efficacy of our epigenetically modified ASCs in preclinical models to assess their usefulness in regenerative medicine.

CONCLUSION

In the current study, we offer several important, albeit preliminary, insights that support the existence of a complex milieu of epigenomic changes able to regulate proliferative gene programs. Although future studies are needed to understand the therapeutic potential of these novel programs in ASCs, we show that 5′-AZA activates a methylation program that likely interferes with *ELK4* downstream signaling, thereby offering one potential mechanism whereby DNA methylation influences transcriptional activity toward adipocyte differentiation pathways.

DATA AVAILABILITY STATEMENT

All data generated or analyzed during this study are included in this article and are found on the NCBI Gene Expression Omnibus (GEO) repository (GSE139157), as well as at the following repository: https://github.com/mepepin/Epigenetics-in-ASCs.

ETHICS STATEMENT

Written informed consent was obtained, and the clinical protocol was approved by the Institutional Review Board of the Department of Experimental Medicine, Sapienza University of Rome (Italy).

AUTHOR CONTRIBUTIONS

MP designed the bioinformatics analysis, interpreted the data, drafted the manuscript, and designed illustrations used in the current work. TI, GB, and CS interpreted the data and drafted the manuscript. MM and EA provided technical experimental support and interpreted the data. SC and FM participated in the design of the study, performed ASC isolation and treatment, and interpreted the data. GD, GF, MF, LD, and GN revised the manuscript critically for important intellectual content. AW provided critical feedback on methods and interpretation, as well as revision of the text. CM and CN designed and coordinated the study and revised the manuscript.

FUNDING

This work was supported by progetto PRIN 2017 (2017F8ZB89), funded by Italian Ministry of Research. Training support was provided to MP by an NIH F30 HL137240.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00346/full#supplementary-material

REFERENCES

Bates, M. (2016). The CRISPR conundrum: as millions of dollars flow into developing the gene-editing tool, scientists continue to discover the technology's potential and pitfalls. *IEEE Pulse* 7, 17–21. doi: 10.1109/MPUL.2016.2514899

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* 321, 209–213. doi: 10.1038/321209a0

Boog, G. (1993). Does a preventive treatment of pregnancy toxemia exist: the role of aspirin. *Rev. Fr. Gynecol. Obstet.* 88, 63–68.

Ceccarelli, S., Nodale, C., Vescarelli, E., Pontecorvi, P., Manganelli, V., Casella, G., et al. (2018). Neuropilin 1 mediates keratinocyte growth factor signaling in adipose-derived stem cells: potential involvement in adipogenesis. *Stem Cells Int.* 2018:1075156. doi: 10.1155/2018/1075156

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

Cho, Y. D., Yoon, W. J., Kim, W. J., Woo, K. M., Baek, J. H., Lee, G., et al. (2014). Epigenetic modifications and canonical wingless/int-1 class (WNT) signaling enable trans-differentiation of nonosteogenic cells into osteoblasts. *J. Biol. Chem.* 289, 20120–20128. doi: 10.1074/jbc.M114.558064

Christman, J. K. (2002). 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene* 21, 5483–5495. doi: 10.1038/sj.onc.1205699

Costantino, S., Camici, G. G., Mohammed, S. A., Volpe, M., Luscher, T. F., and Paneni, F. (2018). Epigenetics and cardiovascular regenerative medicine in the elderly. *Int. J. Cardiol.* 250, 207–214. doi: 10.1016/j.ijcard.2017.09.188

Dai, R., Wang, Z., Samanipour, R., Koo, K. I., and Kim, K. (2016). Adipose-derived stem cells for tissue engineering and regenerative medicine applications. *Stem Cells Int.* 2016:6737345. doi: 10.1155/2016/6737345

Day, B. W., Stringer, B. W., Spanevello, M. D., Charmsaz, S., Jamieson, P. R., Ensbey, K. S., et al. (2011). ELK4 neutralization sensitizes glioblastoma to apoptosis through downregulation of the anti-apoptotic protein Mcl-1. *Neuro Oncol.* 13, 1202–1212. doi: 10.1093/neuonc/nor119

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell.* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Hong, A. (2018). CRISPR in personalized medicine: industry perspectives in gene editing. *Semin. Perinatol.* 42, 501–507. doi: 10.1053/j.semperi.2018.09.008

Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V., and Jordan, I. K. (2012). On the presence and role of human gene-body DNA methylation. *Oncotarget* 3, 462–474.

Kanehisa, M. (2002). The KEGG database. *Novartis Found. Symp.* 247, 91–101. ; discussion 101-103, 119-128, 244-152.

Kornicka, K., Marycz, K., Maredziak, M., Tomaszewski, K. A., and Nicpon, J. (2017). The effects of the DNA methyltranfserases inhibitor 5-Azacitidine on ageing, oxidative stress and DNA methylation of adipose derived stem cells. *J. Cell. Mol. Med.* 21, 387–401. doi: 10.1111/jcmm.12972

Makkonen, H., Jaaskelainen, T., Pitkanen-Arsiola, T., Rytinki, M., Waltering, K. K., Matto, M., et al. (2008). Identification of ETS-like transcription factor 4 as a novel androgen receptor target in prostate cancer cells. *Oncogene* 27, 4865–4876. doi: 10.1038/onc.2008.125

Medvedeva, Y. A., Khamis, A. M., Kulakovskiy, I. V., Ba-Alawi, W., Bhuyan, M. S., Kawaji, H., et al. (2014). Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics* 15:119. doi: 10.1186/1471-2164-15-119

Munoz-Najar, U., and Sedivy, J. M. (2011). Epigenetic control of aging. *Antioxid Redox Signal.* 14, 241–259. doi: 10.1089/ars.2010.3250

Patrikoski, M., Mannerstrom, B., and Miettinen, S. (2019). Perspectives for clinical translation of adipose stromal/stem cells. *Stem Cells Int.* 2019:5858247. doi: 10.1155/2019/5858247

Pepin, M. E., Ha, C. M., Crossman, D. K., Litovsky, S. H., Varambally, S., Barchue, J. P., et al. (2019). Genome-wide DNA methylation encodes cardiac transcriptional reprogramming in human ischemic heart failure. *Lab. Invest.* 99, 371–386. doi: 10.1038/s41374-018-0104-x

Sajadian, S. O., Ehnert, S., Vakilian, H., Koutsouraki, E., Damm, G., Seehofer, D., et al. (2015). Induction of active demethylation and 5hmC formation by 5-azacytidine is TET2 dependent and suggests new treatment strategies against hepatocellular carcinoma. *Clin. Epigenetics* 7:98. doi: 10.1186/s13148-015-0133-x

Scuderi, N., Ceccarelli, S., Onesti, M. G., Fioramonti, P., Guidi, C., Romano, F., et al. (2013). Human adipose-derived stromal cells for cell-based therapies in the treatment of systemic sclerosis. *Cell Transplant.* 22, 779–795. doi: 10.3727/096368912x639017

Shyam, H., Singh, S. K., Kant, R., and Saxena, S. K. (2017). Mesenchymal stem cells in regenerative medicine: a new paradigm for degenerative bone diseases. *Regen. Med.* 12, 111–114. doi: 10.2217/rme-2016-0162

Sun, X., Han, Y., Zhou, L., Chen, E., Lu, B., Liu, Y., et al. (2018). A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. *Bioinformatics* 34, 2715–2723. doi: 10.1093/bioinformatics/bty174

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). 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

Tabatabaei Qomi, R., and Sheykhhasan, M. (2017). Adipose-derived stromal cell in regenerative medicine: a review. *World J. Stem Cells* 9, 107–117. doi: 10.4252/wjsc.v9.i8.107

Tabolacci, E., Mancano, G., Lanni, S., Palumbo, F., Goracci, M., Ferre, F., et al. (2016). Genome-wide methylation analysis demonstrates that 5-aza-2-deoxycytidine treatment does not cause random DNA demethylation in fragile X syndrome cells. *Epigenetics Chromatin* 9:12.

Wang, H. Q., Tuominen, L. K., and Tsai, C. J. (2011). SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics* 27, 225–231. doi: 10.1093/bioinformatics/btq650

Wreczycka, K., Gosdschan, A., Yusuf, D., Gruning, B., Assenov, Y., and Akalin, A. (2017). Strategies for analyzing bisulfite sequencing data. *J. Biotechnol.* 261, 105–115. doi: 10.1016/j.jbiotec.2017.08.007

Wu, D., Li, Y., Zhu, K. S., Wang, H., and Zhu, W. G. (2018). Advances in cellular characterization of the sirtuin isoform, SIRT7. *Front. Endocrinol. (Lausanne)* 9:652. doi: 10.3389/fendo.2018.00652

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 © 2020 Pepin, Infante, Benincasa, Schiano, Miceli, Ceccarelli, Megiorni, Anastasiadou, Della Valle, Fatone, Faenza, Docimo, Nicoletti, Marchese, Wende and Napoli. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 15 April 2020
doi: 10.3389/fbioe.2020.00256

Identification and Validation of Novel Long Non-coding RNA Biomarkers for Early Diagnosis of Oral Squamous Cell Carcinoma

Yue Li[1], Xiaofang Cao[2] and Hao Li[3]*

[1] Department of Orthodontics, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [2] Department of Endodontics, The Second Affiliated Hospital of Endodontics, Harbin, China, [3] Department of Stomatology, The Fourth Affiliated Hospital of Heilongjiang University of Traditional Chinese Medicine, Harbin, China

Long non-coding RNAs (lncRNAs) are recently emerging as a novel promising biomarker for cancer diagnosis and prognosis. Despite these previous investigations, the expression pattern and diagnostic role of lncRNAs in oral squamous cell carcinoma (OSCC) remain unclear. In this study, we identified six novel lncRNA biomarkers (*LINC01697*, *LINC02487*, *LOC105376575*, *AC005083.1*, *SLC8A1-AS1*, and *U62317.1*) from a list of 29 differentially expressed lncRNAs using the least absolute shrinkage and selection operator (LASSO) method in the discovery dataset of 167 OSCC samples and 45 normal oral tissues. Using the logistic regression method, we constructed a six lncRNAs-based diagnostic risk model (6lncRNAScore) which was able to differentiate between OSCC samples and control samples with high performance with AUC of 0.995 and high diagnostic specificity of 88.9% and sensitivity of 98.2% in the discovery dataset. The diagnostic performance of the 6lncRNAScore was further validated in another two independent OSCC dataset with AUC of 0.968 and 1.0. Functional enrichment analysis for lncRNA biomarkers-related mRNAs suggested that lncRNAs biomarkers tended to be involved in the lipid metabolic process. Together, our study highlighted the importance of lncRNAs in OSCC and demonstrated the utility of lncRNA expression as a promising biomarker for early diagnosis of OSCC.

Keywords: biomarker, diagnosis, long non-coding RNAs, oral squamous cell carcinoma, epigenetic

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Xing Chen,
China University of Mining and Technology, China
Lixin Cheng,
Jinan University, China

****Correspondence:***
Yue Li
scullylee@163.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *04 February 2020*
Accepted: *12 March 2020*
Published: *15 April 2020*

Citation:
Li Y, Cao X and Li H (2020) Identification and Validation of Novel Long Non-coding RNA Biomarkers for Early Diagnosis of Oral Squamous Cell Carcinoma. Front. Bioeng. Biotechnol. 8:256. doi: 10.3389/fbioe.2020.00256

INTRODUCTION

Oral cavity and pharynx cancer are one of ten leading cancer types with approximately 53,000 estimated new cases and 10,860 estimated deaths in the United States according to cancer statistics, 2019 (Siegel et al., 2019). Oral squamous cell carcinoma (OSCC) is the most common oral cavity and pharynx cancer accounting for more than 90% of oral cancers (Tandon et al., 2017). Despite the improvement in therapeutic approaches, surgery followed by postoperative radiation or chemoradiation is still standard treatment. The prognosis of OSCC patients remains unfavorable with 5-year survival rates of 40–50% in large part due to diagnosis in advanced stages. Therefore,

identifying biomarkers for early diagnosis is a crucial way to improve the survival rate and quality of life of OSCC patients.

Recent advances in high-throughput sequencing theologies have found that the approximately 2% human genome encodes only ∼20,000 protein-coding genes and the vast majority of the human genome were actively transcribed as non-coding RNAs (ncRNAs) (Wright and Bruford, 2011). Long nc RNAs (lncRNAs) are a newly discovered class of ncRNAs ranging in length from 200 nt to ∼100 kilobases (kb) (Ulitsky and Bartel, 2013). Functional studies in cell lines and animal models have demonstrated that lncRNAs have emerged as key genomic regulators in a wide variety of biological pathways including differentiation and development (Rinn and Chang, 2012; Batista and Chang, 2013). A large number of dysregulated lncRNAs have been discovered in multiple tumor types compared with the corresponding normal tissues, which demonstrated the important roles of lncRNAs in cancer development, progression, and treatment (Zhang et al., 2013; Li et al., 2017). Increasing evidence have suggested that lncRNAs may be promising biomarkers in cancer diagnosis and prognosis compared with protein-coding genes because lncRNAs are expressed in a more highly cell type-, tissue-, and disease type-specific manner than protein-coding genes and their expression may be a better indicator of the tumor status (Hauptman and Glavac, 2013; Zhou et al., 2015, 2017; Salviano-Silva et al., 2018; Bao et al., 2019).

In this study, to evaluate the potential of lncRNAs biomarkers in the diagnosis of OSCC, we compared differential expression profiles of lncRNAs between OSCC samples and controls in a larger OSCC dataset. Then we used multiple statistical methods to identify novel lncRNA biomarkers and developed an lncRNA-based diagnostic prediction model, which was validated in several independent OSCC datasets.

MATERIALS AND METHODS

OSCC Datasets

We retrospectively collected a total of 262 samples from three publicly available datasets from the Gene Expression Omnibus (GEO) database,[1] including 167 OSCC samples and 45 normal oral tissues from GSE30784 dataset[2] (Chen et al., 2008), 26 OSCC samples and 12 control samples from GSE9844 dataset[3] (Ye et al., 2008), and six OSCC samples and six adjacent non-involved oral tissue from GSE74530 dataset[4] (Oghumu et al., 2016). The largest sample dataset GSE30784 was used as a discovery dataset and the other two datasets were used as independent testing datasets.

Acquisition and Analysis of lncRNA Expression Profiles

The raw microarray data files (.CEL files) of three OSCC sample datasets on Affymetrix Human Genome U133 Plus 2.0 (Affymetrix HG-U133 Plus 2.0) were downloaded directly from the GEO database and were processed and normalized using robust multichip average method by R "affy" package. LncRNA expression data of all samples in three datasets were retrieved by repurposing the probes based on the NetAffx annotation files of the probe sets and the annotation files of RefSeq and GENCODE according to previous studies (Zhou et al., 2018, 2019). Finally, expression data of 2466 lncRNAs of three OSCC datasets were obtained for further analysis.

Differential expression analysis of lncRNAs between OSCC samples and control samples was performed using the R package "limma" (version 3.42.0), those lncRNAs with $|\log2(\text{fold change})| > 1$ and false discovery rate (FDR) adjusted p-value < 0.05 was considered as differentially expressed lncRNAs.

Construction of lncRNA-Based Diagnostic Risk Model for Early Detection of OSCC

The least absolute shrinkage and selection operator (LASSO) method was used to select the most useful predictive features from the list of differentially expressed lncRNAs as diagnostic lncRNA biomarkers in the discovery dataset. Then all diagnostic lncRNA biomarkers were fitted a logistic regression model as the covariates, and an lncRNA-based diagnostic risk model was constructed by using the sum of expression value of lncRNA biomarkers weighted by the unbiased coefficients estimates from the logistic regression model (Hao et al., 2017) as follows:

$$\text{Diagnostic risk score} = \frac{e^{w^T * x + b}}{1 + e^{w^T * x + b}}$$

Where x is expression levels of each lncRNA biomarker, w is the coefficients estimates from the logistic regression model, and b is 18.35146902. The risk score range from 0 to 1, and >0.5 was used as the cutoff for the diagnosis of OSCC.

The diagnostic performance of the lncRNA-based diagnostic risk model was evaluated by a receiver operating characteristic (ROC) curve and the area under the ROC (AUC).

Functional Enrichment Analysis

Functional enrichment analysis of GO and Kyoto encyclopedia of genes and genomes (KEGG) pathway was performed to identify significantly enriched GO terms and KEGG pathways of mRNAs correlated with diagnostic lncRNA biomarkers using the R package "clusterprofiler" (Yu et al., 2012).

RESULTS

Analysis of Altered lncRNA Expression Pattern in OSCC

To explore the altered lncRNA expression pattern in OSCC, we compared lncRNA expression profiles between 167 OSCC samples and 45 control samples in the discovery dataset and identified 29 differentially expressed lncRNAs [$|\log2(\text{fold change})| > 1$ and FDR-adjusted p-value < 0.05] (**Supplementary Table S1**). Of them, 18 lncRNAs are downregulated and 11

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

[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30784

[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9844

[4] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74530

lncRNAs are upregulated in OSCC compared with control samples (**Figure 1A**). Results of unsupervised hierarchical clustering analysis showed that all samples in the discovery dataset can could be grouped into two clusters based on the expression pattern of these 29 differentially expressed lncRNAs. 159 of 167 OSCC samples and one control were classified into Cluster 2, and 44 of 45 control samples and eight OSCC samples were classified into Cluster 1. Statistical analysis indicated a significant association between clusters and disease status ($p < 0.001$, chi-squared test) (**Figure 1B**).

Development and Validation of the lncRNA-Based Diagnostic Risk Model for Early Detection of OSCC in the Discovery Dataset

To identify diagnostic lncRNA biomarkers, 29 differentially expressed lncRNAs were analyzed using the LASSO with 10-fold cross-validation and the turning parameter λ of 0.055. As shown in **Figure 2A**, we obtained six lncRNAs (*LINC01697*, *LINC02487*, *LOC105376575*, *AC005083.1*, *SLC8A1-AS1*, and *U62317.1*) from the list of differentially expressed lncRNAs as optimal diagnostic biomarkers considering a balance between classification accuracy and the number of lncRNAs (**Table 1**). Of six lncRNA biomarkers, five lncRNAs biomarkers (*LINC01697*, *LINC02487*, *LOC105376575*, *AC005083.1*, and *SLC8A1-AS1*) seem to be tumor suppressors and downregulated in OSCC, while lncRNA U62317.1 tended to be an oncogene and up-regulated in OSCC (**Figure 2B**). Using a logistic regression method, a six-lncRNAs-based diagnostic risk model was generated (named 6lncRNAScore). When the 6lncRNAScore was applied to samples of the discovery dataset, the 6lncRNAScore correctly classified 164 of 167 OSCC samples and 40 of 45 control samples, achieving an AUC of 0.995 with the sensitivity of 98.2% and the specificity of 88.9% (**Figures 2C,D**).

Independent Validation of the 6lncRNAScore in Two Independent OSCC Datasets

To confirm the robustness of the 6lncRNAScore for early diagnosis, the 6lncRNAScore was tested in an additional independent OSCC dataset (GSE9844) of 38 samples. We first investigated the expression patterns of six lncRNA biomarkers between OSCC samples and control samples in the GSE9844 dataset. As shown in **Figure 3A**, five lncRNA biomarkers (*LINC01697*, *LINC02487*, *LOC105376575*, *AC005083.1*, and *SLC8A1-AS1*) were significantly downregulated and one lncRNA U62317.1 was significantly upregulated in OSCC samples compared with control samples, which is consistent with observed in the discovery dataset. Then each sample of GSE9844 was assigned a risk score according to 6lncRNAScore, and was classified as OSCC-like or normal-like samples. Finally, the 6lncRNAScore correctly classified all 26 OSCC samples and eight of 12 control samples, achieving an AUC of 0.968 with the sensitivity of 100% and the specificity of 66.7% (**Figures 3B,C**).

Further validation of the diagnostic ability of the 6lncRNAScore was conducted using another completely independent OSCC dataset (GSE74530) of 12 samples. The expression pattern of six lncRNA biomarkers between OSCC samples and control samples in the GSE74530 dataset was examined. As shown in **Figure 4A**, five (*LINC01697*, *LINC02487*, *LOC105376575*, *AC005083.1*, and *U62317.1*) of six lncRNA

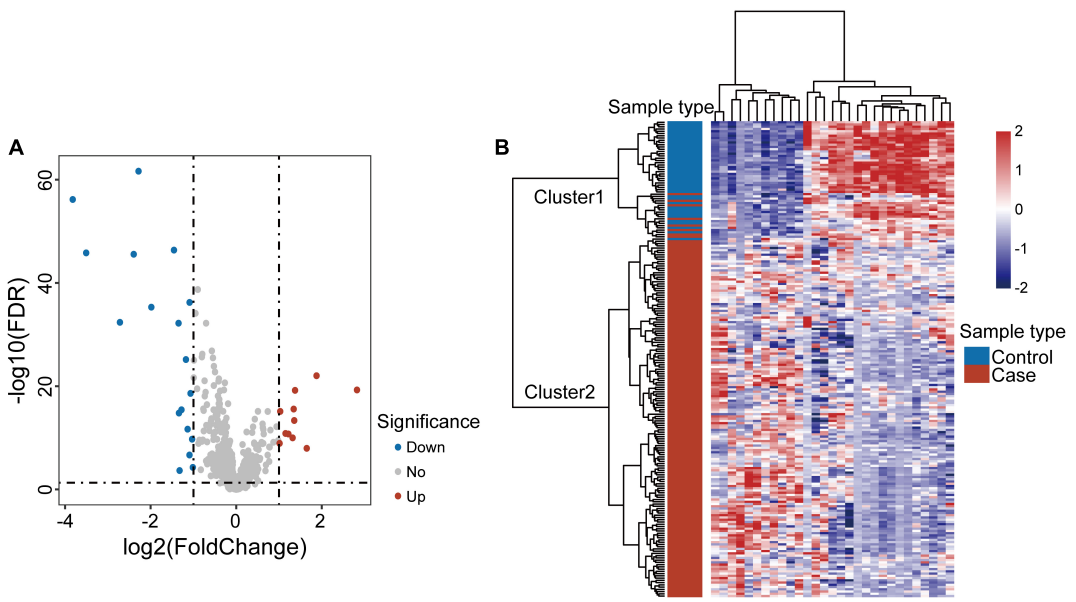


FIGURE 1 | Analysis of differentially expressed genes in oral squamous cell carcinoma. **(A)** Volcano plot displaying differential expressed lncRNAs between OSCC samples and control samples. **(B)** Unsupervised clustering of all samples based on the expression pattern of 29 differentially expressed lncRNAs in the discovery dataset.

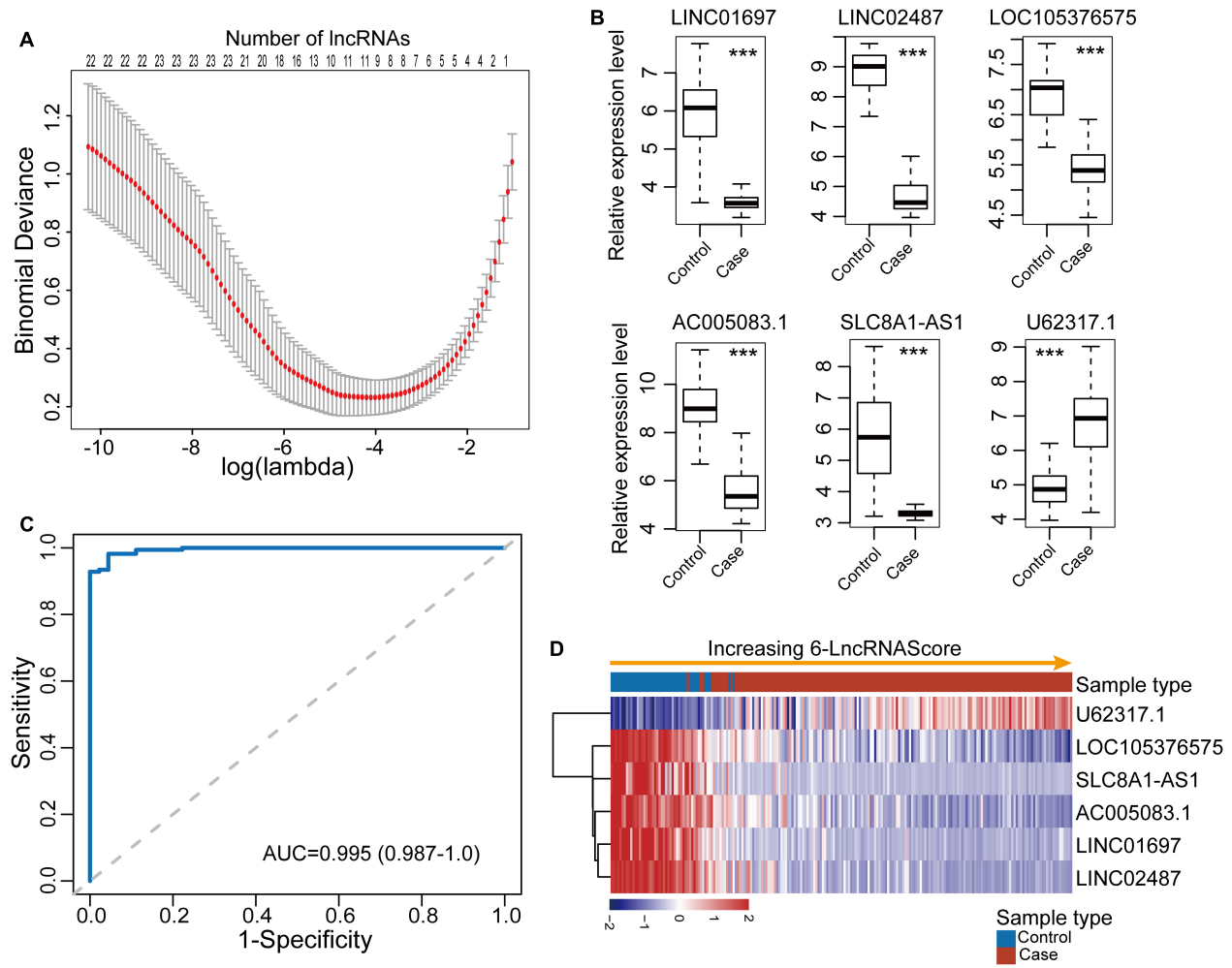


FIGURE 2 | Development and performance evaluation of the lncRNA-based diagnostic risk model in the discovery dataset. **(A)** Feature selection for lncRNA biomarkers using the least absolute shrinkage and selection operator (LASSO) method. **(B)** Expression patterns of six lncRNA biomarkers between OSCC samples and control samples. **(C)** Receiver operating characteristic (ROC) curves for 6lncRNAScore. **(D)** Expression heatmap of six lncRNA biomarkers of samples with increasing 6lncRNAScore. ***$p < 0.001$.

biomarkers revealed consistent expression patterns as observed in the discovery dataset and GSE9844 dataset except for *SLC8A1-AS1*. When the 6lncRNAScore was tested in the GSE74530 dataset, the 6lncRNAScore correctly classified all six OSCC samples and three of six control samples, achieving an AUC of 1.0 with the sensitivity of 100% and the specificity of 50% (**Figures 4B,C**).

Functional Implication of Six lncRNA Biomarkers

To gain insight into the biological function of six newly identified lncRNA biomarkers, we computed the Pearson correlation coefficient (PCC) between expression levels of mRNAs and six lncRNA biomarkers and identified 95 mRNAs correlated with six lncRNA biomarkers ($r > 0.8$ and $p < 0.05$). Then we performed GO and KEGG enrichment analysis for 95 mRNAs correlated with six lncRNA biomarkers using the R package "clusterprofiler." Results of functional enrichment analysis showed that mRNAs correlated with six lncRNA biomarkers might be involved in several lipid metabolic processes, such as membrane lipid metabolic process, cellular lipid catabolic process, sphingolipid metabolic process, fatty acid derivative metabolic process, and so on (**Figure 5**).

DISCUSSION

Although smoking and alcohol use were well-known chief risk factors for OSCC, complex genetic, transcriptional, and epigenetic alteration and their interaction between them have also been reported to indeed contribute to OSCC tumorigenesis (Choi and Myers, 2008; Perez-Sayans et al., 2009). Traditional diagnosis methods, such as the biopsy, endoscopy, and chest X-ray and CT, were inadequate for early diagnosis, and there is the lack of specific and sensitive biomarkers for earlier detection

TABLE 1 | Detailed information of six diagnostic lncRNA biomarkers identified in the discovery dataset.

Ensemble id	Gene name	Genomic location	Coefficient[a]	FDR *p*-value[b]
ENSG00000232079	LINC01697	Chr 21: 28,048,404-28,137,611 (+)	−0.0953	<0.001
ENSG00000203688	LINC02487	Chr 6: 167,679,626-167,696,290 (−)	−0.4317	<0.001
105376575	LOC105376575	NC_000011.10 (17332042.17349973)	−1.8052	<0.001
ENSG00000233834	AC005083.1	Chr 7: 20,217,577-20,221,700 (+)	−0.8215	<0.001
ENSG00000227028	SLC8A1-AS1	Chr 2: 39,786,453-40,255,209 (+)	−0.4150	<0.001
ENSG00000272666	U62317.1	Chr 22: 50,542,305-50,542,906 (−)	0.8571	<0.001

[a]*Derived from the logistic regression model.* [b]*Derived from differential expression analysis.*

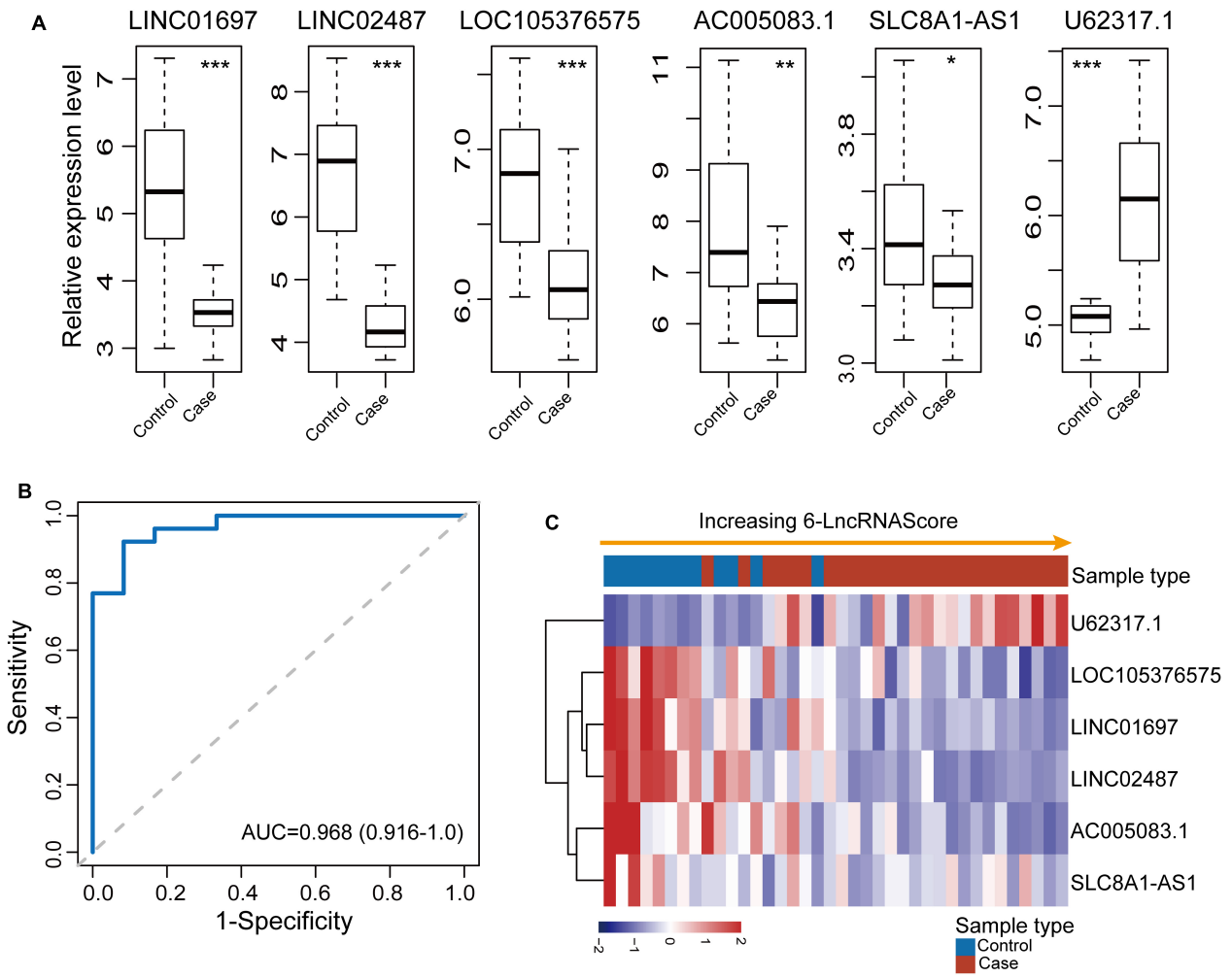


FIGURE 3 | Independent validation of the 6lncRNAScore in 38 samples of GSE9844. **(A)** Expression patterns of six lncRNA biomarkers between OSCC samples and control samples. **(B)** Receiver operating characteristic (ROC) curves for 6lncRNAScore. **(C)** Expression heatmap of six lncRNA biomarkers of samples with increasing 6lncRNAScore. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

of OSCC. Alteration in molecular profiles have been observed in various human cancers compared to corresponding normal tissues, and have been used as novel markers for cancer diagnosis and prognosis (Mehta et al., 2010; Jurgensmeier et al., 2014; Chen et al., 2018, 2019; Sun et al., 2020). lncRNAs are recently emerging as a novel promising biomarker for cancer diagnosis and prognosis (Shi et al., 2016; Bolha et al., 2017). Increasing efforts have been made to focus on the discovery and characterization of lncRNA biomarkers in various human cancers. Recent studies also observed many lncRNAs dysregulation which served an oncogenic or tumor suppression part in OSCC (Lu et al., 2017; Kong et al., 2018; Zhang et al., 2019). Despite these previous investigations, the expression pattern and diagnostic role of lncRNA in OSCC remain unclear.

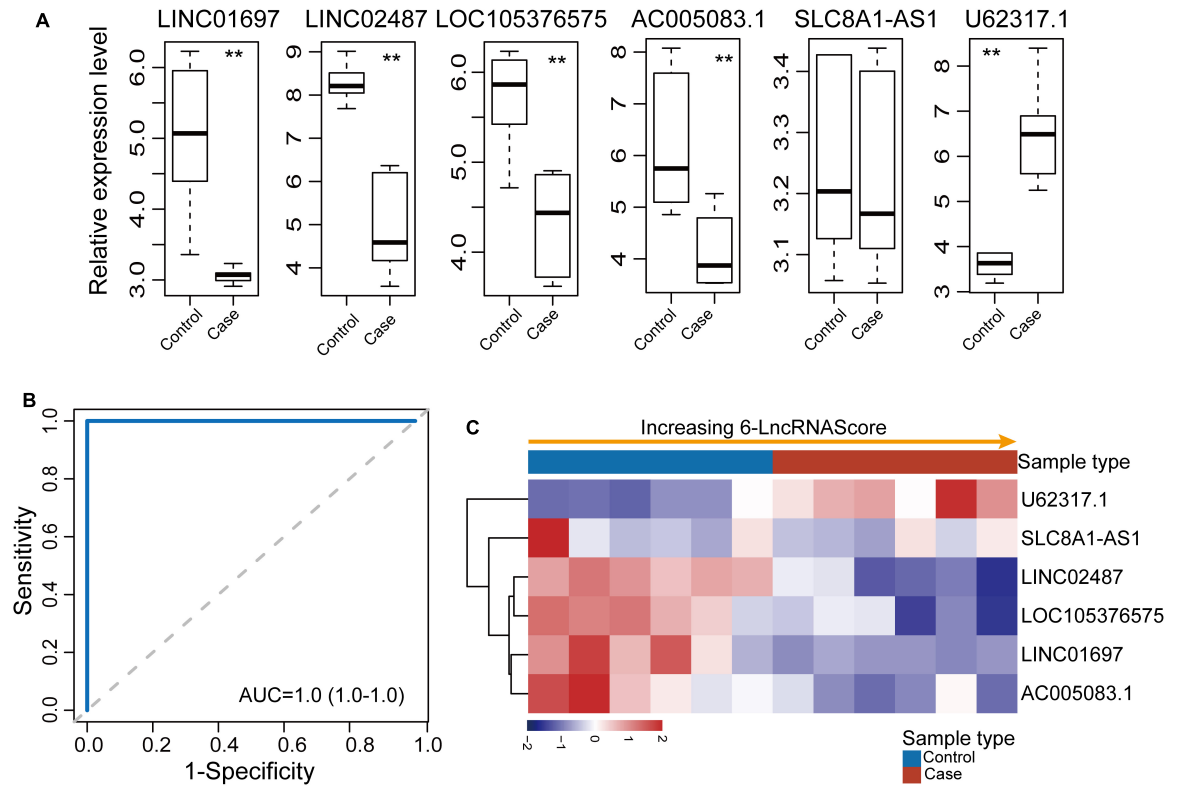


FIGURE 4 | Further confirmation of the 6lncRNAScore in another GSE74530 dataset. **(A)** Expression patterns of six lncRNA biomarkers between OSCC samples and control samples. **(B)** Receiver operating characteristic (ROC) curves for 6lncRNAScore. **(C)** Expression heatmap of six lncRNA biomarkers of samples with increasing 6lncRNAScore. $^{**}p < 0.01$.

In this study, we first obtained lncRNA expression files of a large number of OSCC samples and corresponding control samples by repurposing publicly available microarray data, and compared the expression pattern of lncRNAs between OSCC and control samples as the first step toward identifying diagnostic lncRNA biomarkers. A total of 29 lncRNAs were determined as differentially expressed lncRNAs whose dysregulated expressions are closely associated with OSCC. To identify diagnostic lncRNA biomarkers from the list of differentially expressed lncRNAs, we performed a feature selection procedure to reduce the number of lncRNAs using the lasso binary logistic regression model which is the powerful and versatile regression method for high-dimensional data. Considering a balance between classification accuracy and the number of lncRNAs, six lncRNAs were identified as diagnostic biomarkers including five tumor suppressor lncRNAs (*LINC01697*, *LINC02487*, *LOC105376575*, *AC005083.1*, and *SLC8A1-AS1*) and one oncogenic lncRNAs (*U62317.1*). Although the huge number of lncRNAs have been identified using high-throughput experimental technologies, only a very small fraction of lncRNAs were well functionally characterized. After the literature search, we found that several lncRNA biomarkers identified in our study have been reported to be involved in human diseases. *LINC02487* has been reported as a tumor suppressor to inhibit migration and invasion of oral cancer cells through directly binding protein USP17 (Feng et al., 2019). lncRNA *AC005083.1* was also differentially expressed in lung adenocarcinoma (Peng et al., 2017). The latest study by Guo et al. found that LncRNA *SLC8A1-AS1* protects against myocardial damage through activation of the cGMP-PKG signaling pathway. Previous studies have suggested that lncRNA function can be inferred by using co-expression with the coding genes approach (Ma et al., 2012; Cheng et al., 2016). Therefore, to further explore the potential function of other lncRNA biomarkers, we performed functional enrichment analysis for lncRNA biomarkers-related mRNAs derived from co-expression analysis and found that mRNAs co-correlated with lncRNA biomarkers tended to be involved in lipid metabolic process. The modulation in lipid metabolism has been implicated in increased invasiveness of OSCC cells, and some lipid metabolism-related genes were used to predict poor outcome of OSCC (Beloribi-Djefaflia et al., 2016; Sant'Anna-Silva et al., 2018; Hu et al., 2019).

To accelerate clinical application, we constructed a diagnostic prediction model with these six lncRNA biomarkers using a logistic regression method. The 6lncRNAScore achieved an AUC value of 0.995 for distinguishing OSCC samples and healthy controls in the discovery dataset. To confirm the robustness of the 6lncRNAScore for early diagnosis, we also tested the performance of the 6lncRNAScore in the other two independent OSCC datasets. Validation results showed that

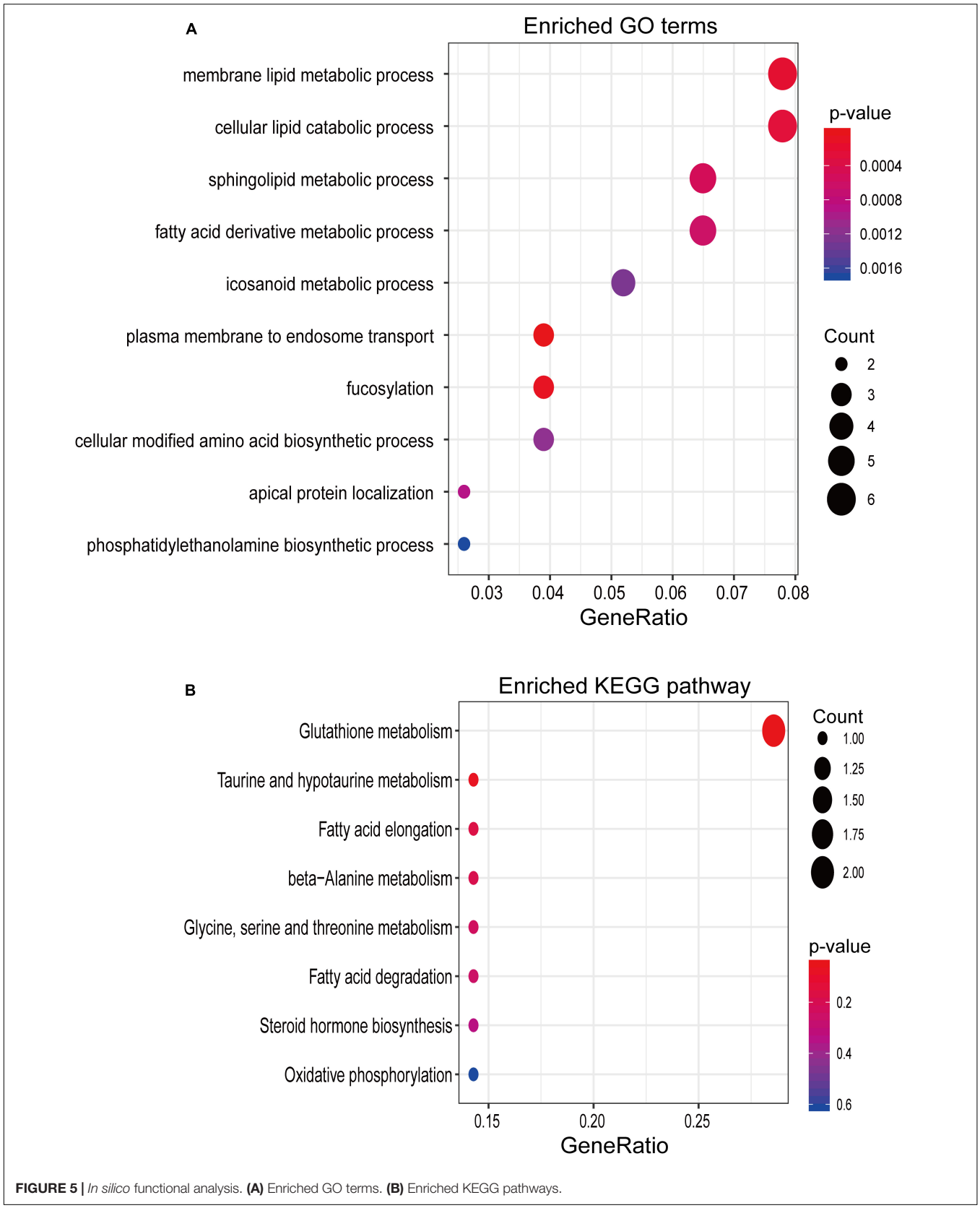


FIGURE 5 | *In silico* functional analysis. **(A)** Enriched GO terms. **(B)** Enriched KEGG pathways.

the 6lncRNAScore also had well predictive performance in effectively discriminating OSCC patients from controls in the other two independent datasets comparable with the discovery dataset. These results demonstrated the reproducible and robust predictive power and general applicability of the 6lncRNAScore in early diagnosis of OSCC.

CONCLUSION

In summary, we evaluated the utility of lncRNA expression in early diagnosis of OSCC and constructed a diagnostic prediction model composing of six lncRNA biomarkers which showed high and robust diagnostic performance in discriminating OSCC patients from controls. However, further experimental studies or independent validations are warranted to fully explore the molecular mechanism and clinical applications of lncRNAs in the diagnosis of OSCC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: Three publicly available OSCC datasets were obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30784, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9844, and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74530).

AUTHOR CONTRIBUTIONS

YL conceived and designed the experiments. YL, XC, and HL performed the experiments and analyzed the data. YL wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the Young and Middle-Aged Innovative Scientific Research Fund of The Second Affiliated Hospital of Harbin Medical University (KYCX2019-18).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00256/full#supplementary-material

REFERENCES

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief Bioinform*. doi: 10.1093/bib/bbz118

Batista, P. J., and Chang, H. Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. *Cell* 152, 1298–1307. doi: 10.1016/j.cell.2013.02.012

Beloribi-Djefaflia, S., Vasseur, S., and Guillaumond, F. (2016). Lipid metabolic reprogramming in cancer cells. *Oncogenesis* 5:e189. doi: 10.1038/oncsis.2015.49

Bolha, L., Ravnik-Glavaè, M., and Glavaè, D. J. (2017). Long noncoding RNAs as biomarkers in cancer. *Dis. Markers* 2017:7243968.

Chen, C., Mendez, E., Houck, J., Fan, W., Lohavanichbutr, P., Doody, D., et al. (2008). Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiol. Biomarkers Prev.* 17, 2152–2162. doi: 10.1158/1055-9965.EPI-07-2893

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. J. B. (2018). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265.

Chen, X., Xie, D., Zhao, Q., and You, Z.-H. J. (2019). MicroRNAs and complex diseases: from experimental results to computational models. *Brief Bioinform.* 20, 515–539. doi: 10.1093/bib/bbx130

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* 7:47864.

Choi, S., and Myers, J. N. (2008). Molecular pathogenesis of oral squamous cell carcinoma: implications for therapy. *J. Dent. Res.* 87, 14–32. doi: 10.1177/154405910808700104

Feng, L., Qiu, F., Zhang, J., Sun, M., Chen, W., and Qiu, W. (2019). *Tumor Suppressor LINC02487 Inhibits Oral Cancer Cell Migration And Invasion Through Directly Binding Protein USP17*. Philadelphia: AACR.

Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., et al. (2017). DNA methylation markers for diagnosis and prognosis of common cancers. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7414–7419. doi: 10.1073/pnas.1703577114

Hauptman, N., and Glavac, D. (2013). Long non-coding RNA in cancer. *Int. J. Mol. Sci.* 14, 4655–4669. doi: 10.3390/ijms14034655

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

Jurgensmeier, J. M., Eder, J. P., and Herbst, R. S. (2014). New strategies in personalized medicine for solid tumors: molecular markers and clinical trial designs. *Clin. Cancer Res.* 20, 4425–4435. doi: 10.1158/1078-0432.ccr-13-0753

Kong, J., Sun, W., Zhu, W., Liu, C., Zhang, H., and Wang, H. (2018). Long noncoding RNA LINC01133 inhibits oral squamous cell carcinoma metastasis through a feedback regulation loop with GDF15. *J. Surg. Oncol.* 118, 1326–1334. doi: 10.1002/jso.25278

Li, S., Li, B., Zheng, Y., Li, M., Shi, L., and Pu, X. (2017). Exploring functions of long noncoding RNAs across multiple cancers through co-expression network. *Sci. Rep.* 7:754.

Lu, M. Y., Liao, Y. W., Chen, P. Y., Hsieh, P. L., Fang, C. Y., Wu, C. Y., et al. (2017). Targeting LncRNA HOTAIR suppresses cancer stemness and metastasis in oral carcinomas stem cells through modulation of EMT. *Oncotarget* 8, 98542–98552. doi: 10.18632/oncotarget.21614

Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J., et al. (2012). Molecular mechanisms and function prediction of long noncoding RNA. *Sci. World J.* 2012:541786. doi: 10.1100/2012/541786

Mehta, S., Shelling, A., Muthukaruppan, A., Lasham, A., Blenkiron, C., Laking, G., et al. (2010). Predictive and prognostic molecular markers for cancer medicine. *Ther. Adv. Med. Oncol.* 2, 125–148. doi: 10.1177/1758834009360519

Oghumu, S., Knobloch, T. J., Terrazas, C., Varikuti, S., Ahn-Jarvis, J., Bollinger, C. E., et al. (2016). Deletion of macrophage migration inhibitory factor inhibits murine oral carcinogenesis: potential role for chronic pro-inflammatory immune mediators. *Int. J. Cancer* 139, 1379–1390. doi: 10.1002/ijc.30177

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.

Perez-Sayans, M., Somoza-Martin, J. M., Barros-Angueira, F., Reboiras-Lopez, M. D., Gandara Rey, J. M., and Garcia-Garcia, A. (2009). Genetic and molecular alterations associated with oral squamous cell cancer (Review). *Oncol. Rep.* 22, 1277–1282. doi: 10.3892/or_00000565

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166. doi: 10.1146/annurev-biochem-051410-092902

Salviano-Silva, A., Lobo-Alves, S. C., Almeida, R. C., Malheiros, D., and Petzl-Erler, M. L. (2018). Besides pathology: long non-coding RNA in cell and tissue homeostasis. *Noncoding RNA* 4. doi: 10.3390/ncrna4010003

Sant'Anna-Silva, A. C. B., Santos, G. C., Campos, S. P. C., Oliveira Gomes, A. M., Perez-Valencia, J. A., and Rumjanek, F. D. (2018). Metabolic profile of oral squamous carcinoma cell lines relies on a higher demand of lipid metabolism in metastatic cells. *Front. Oncol.* 8:13. doi: 10.3389/fonc.2018.00013

Shi, T., Gao, G., and Cao, Y. J. (2016). Long noncoding RNAs as novel biomarkers have a promising future in cancer diagnostics. *Dis. Markers* 2016:9085195.

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Sun, J., Zhang, Z., Bao, S., Yan, C., Hou, P., Wu, N., et al. (2020). Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *J. Immunother. Cancer* 8:e000110. doi: 10.1136/jitc-2019-000110

Tandon, P., Dadhich, A., Saluja, H., Bawane, S., and Sachdeva, S. (2017). The prevalence of squamous cell carcinoma in different sites of oral cavity at our Rural Health Care Centre in Loni, Maharashtra - a retrospective 10-year study. *Contemp. Oncol.* 21, 178–183. doi: 10.5114/wo.2017.68628

Ulitsky, I., and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26–46. doi: 10.1016/j.cell.2013.06.020

Wright, M. W., and Bruford, E. A. J. H. (2011). Naming'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Hum. Genomics* 5:90. doi: 10.1186/1479-7364-5-2-90

Ye, H., Yu, T., Temam, S., Ziober, B. L., Wang, J., Schwartz, J. L., et al. (2008). Transcriptomic dissection of tongue squamous cell carcinoma. *BMC Genomics* 9:69. doi: 10.1186/1471-2164-9-69

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, C., Bao, C., Zhang, X., Lin, X., Pan, D., and Chen, Y. (2019). Knockdown of lncRNA LEF1-AS1 inhibited the progression of oral squamous cell carcinoma (OSCC) via Hippo signaling pathway. *Cancer Biol. Ther.* 20, 1213–1222. doi: 10.1080/15384047.2019.1599671

Zhang, H., Chen, Z., Wang, X., Huang, Z., He, Z., and Chen, Y. (2013). Long non-coding RNA: a new player in cancer. *J. Hematol. Oncol.* 6:37. doi: 10.1186/1756-8722-6-37

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2019). Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Brief Bioinform.* 20, 598–608. doi: 10.1093/bib/bby021

Zhou, M., Zhao, H., Wang, Z., Cheng, L., Yang, L., Shi, H., et al. (2015). Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. *J. Exp. Clin. Cancer Res.* 34:102.

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

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 © 2020 Li, Cao 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: 16 April 2020
doi: 10.3389/fcell.2020.00252

Integrative Analysis for Elucidating Transcriptomics Landscapes of Glucocorticoid-Induced Osteoporosis

Xiaoxia Ying[1†], Xiyun Jin[2†], Pingping Wang[2†], Yuzhu He[1], Haomiao Zhang[1], Xiang Ren[1], Songling Chai[1], Wenqi Fu[1], Pengcheng Zhao[1], Chen Chen[1], Guowu Ma[1] and Huiying Liu[1]**

[1] School of Stomatology, Dalian Medical University, Dalian, China, [2] School of Life Sciences and Technology, Harbin Institute of Technology, Harbin, China

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Liang Yu,
Xidian University, China
Yanshuo Chu,
The University of Texas MD Anderson Cancer Center, United States

***Correspondence:**
Guowu Ma
Mgw64024@163.com
Huiying Liu
lhy04512000@dmu.edu.cn

† These authors have contributed equally to this work

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

Received: 07 February 2020
Accepted: 25 March 2020
Published: 16 April 2020

Citation:
Ying X, Jin X, Wang P, He Y, Zhang H, Ren X, Chai S, Fu W, Zhao P, Chen C, Ma G and Liu H (2020) Integrative Analysis for Elucidating Transcriptomics Landscapes of Glucocorticoid-Induced Osteoporosis. Front. Cell Dev. Biol. 8:252. doi: 10.3389/fcell.2020.00252

Osteoporosis is the most common bone metabolic disease, characterized by bone mass loss and bone microstructure changes due to unbalanced bone conversion, which increases bone fragility and fracture risk. Glucocorticoids are clinically used to treat a variety of diseases, including inflammation, cancer and autoimmune diseases. However, excess glucocorticoids can cause osteoporosis. Herein we performed an integrated analysis of two glucocorticoid-related microarray datasets. The WGCNA analysis identified 3 and 4 glucocorticoid-related gene modules, respectively. Differential expression analysis revealed 1047 and 844 differentially expressed genes in the two datasets. After integrating differentially expressed glucocorticoid-related genes, we found that most of the robust differentially expressed genes were up-regulated. Through protein-protein interaction analysis, we obtained 158 glucocorticoid-related candidate genes. Enrichment analysis showed that these genes are significantly enriched in the osteoporosis related pathways. Our results provided new insights into glucocorticoid-induced osteoporosis and potential candidate markers of osteoporosis.

Keywords: osteoporosis, glucocorticoid, microarray, protein-protein interaction, enrichment analysis

INTRODUCTION

Osteoporosis is the most common bone disease in the world, which is characterized by low bone mass, microstructure degeneration of bone tissue and decreased bone strength (Coughlan and Dockery, 2014). Osteoporosis is mainly divided into two categories, primary and secondary (Glaser and Kaplan, 1997). Primary osteoporosis further divided into three subtypes, postmenopausal osteoporosis, age-related osteoporosis and idiopathic osteoporosis. Secondary osteoporosis is a metabolic bone disease caused by disease or drugs, including glucocorticoid-induced osteoporosis. Excessive glucocorticoids can affect the differentiation and maturation of osteoblasts, leading to a decrease in their number and function, and can also promote the apoptosis of osteoblasts, thus further reduce bone formation (Whittier and Saag, 2016).

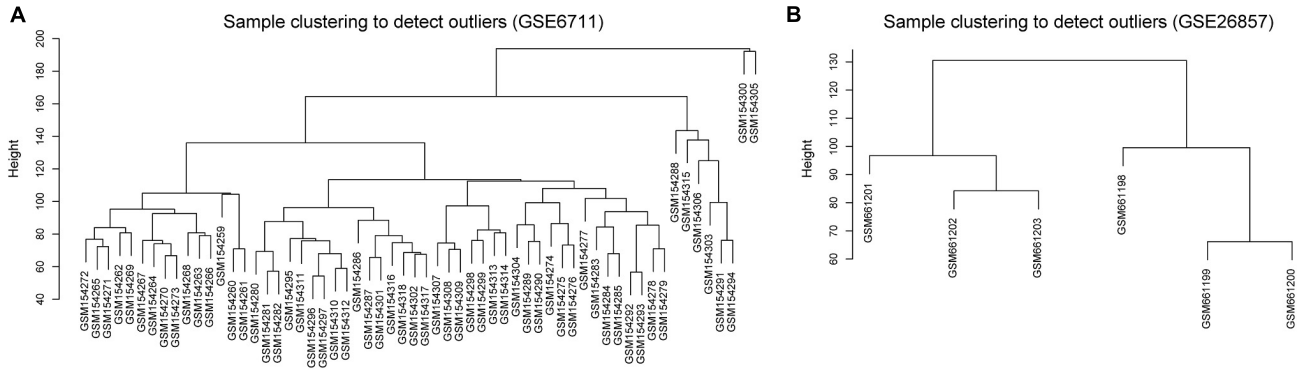


FIGURE 1 | Hierarchical clustering of samples. **(A)** GSE6711 **(B)** GSE26857.

Since the successful application of microarray technology, it has been widely used for expression profiling analysis in almost all fields of biological research (Sun et al., 2018). Due to its high-throughput characteristics, microarray technology has greatly advanced many areas of biological research by transforming the study of biology from a single gene level to the whole transcriptome-wide level (Canales et al., 2006; Cheng et al., 2019). Based on genome-wide microarray expression data, Xiao et al. (2008) studied osteoporosis-related B cells and emphasized the role of B cells in the pathogenesis of osteoporosis. Liu et al. (2005) compared gene expression in circulating monocytes from high and low bone mineral density samples based on the microarray data, revealing the role of monocytes in the pathophysiological mechanism of osteoporosis. In addition, some studies have focused on glucocorticoids and explored the mechanism by which they induce bone cell apoptosis (Lu et al., 2007; Jewell et al., 2012). However, these studies were only conducted in a limited number of single datasets one by one, and there was heterogeneity between different datasets. Therefore, more robust results will be obtained by integrative analysis of the multiple datasets.

In this study, we performed integrative analysis of two glucocorticoid-related osteosarcomas microarray datasets. Glucocorticoid-related gene modules were identified firstly. Then differential expression analysis was performed to obtain glucocorticoid-related differentially expressed genes in each dataset. In order to obtain robust results, the differentially expressed glucocorticoid-related genes of two datasets were intersected. Network analysis revealed that there were 158 robust glucocorticoid-related differentially expressed genes with interacting protein partners in protein-protein interaction network. Finally, gene function enrichment analysis showed that differentially expressed genes under glucocorticoid conditions were enriched in pathways associated with osteoporosis.

MATERIALS AND METHODS

GEO Datasets

The glucocorticoid-related osteosarcoma U-2 OS bone cells microarray datasets of GSE6711 and GSE26857 were downloaded from Gene Expression Omnibus (GEO) database[1] (Edgar et al., 2002), which included a total of 66 samples (48 glucocorticoid-treated and 18 untreated).

Identification of Glucocorticoid-Related Gene Modules

Glucocorticoid-related gene modules of two datasets were identified by R package WGCNA, respectively (Langfelder and Horvath, 2008). The soft power applied for gene modules identification was 7 (GSE6711) and 24 (GSE26857). Correlation coefficients between the module Eigengenes and traits were calculated using Pearson's method. Glucocorticoid-related gene modules were defined as those with correlation coefficients greater than 0.5 (Cheng et al., 2018).

Identification of Differentially Expressed Genes (DEGs)

The DEGs of glucocorticoid-related gene modules were identified using the R package limma with a threshold of $|\log_2\text{FoldChange}| > 1$ and $P < 0.05$. Then the genes up-regulated/down-regulated in both two datasets were considered as the robust up-regulated/down-regulated glucocorticoid-related DEGs.

The Construction of Protein-Protein Interaction (PPI) Network

For the final up- or down-regulated glucocorticoid-related DEGs, we constructed PPI network by STRING database[2], respectively (von Mering et al., 2003). Then Cytoscape (V.3.7.2) was used for network visualization (Shannon et al., 2003).

Pathway Enrichment Analysis

We performed pathway enrichment analysis of DEGs included in PPI network using Enrichr[3], an online enrichment analysis

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

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

[3]http://amp.pharm.mssm.edu/Enrichr/

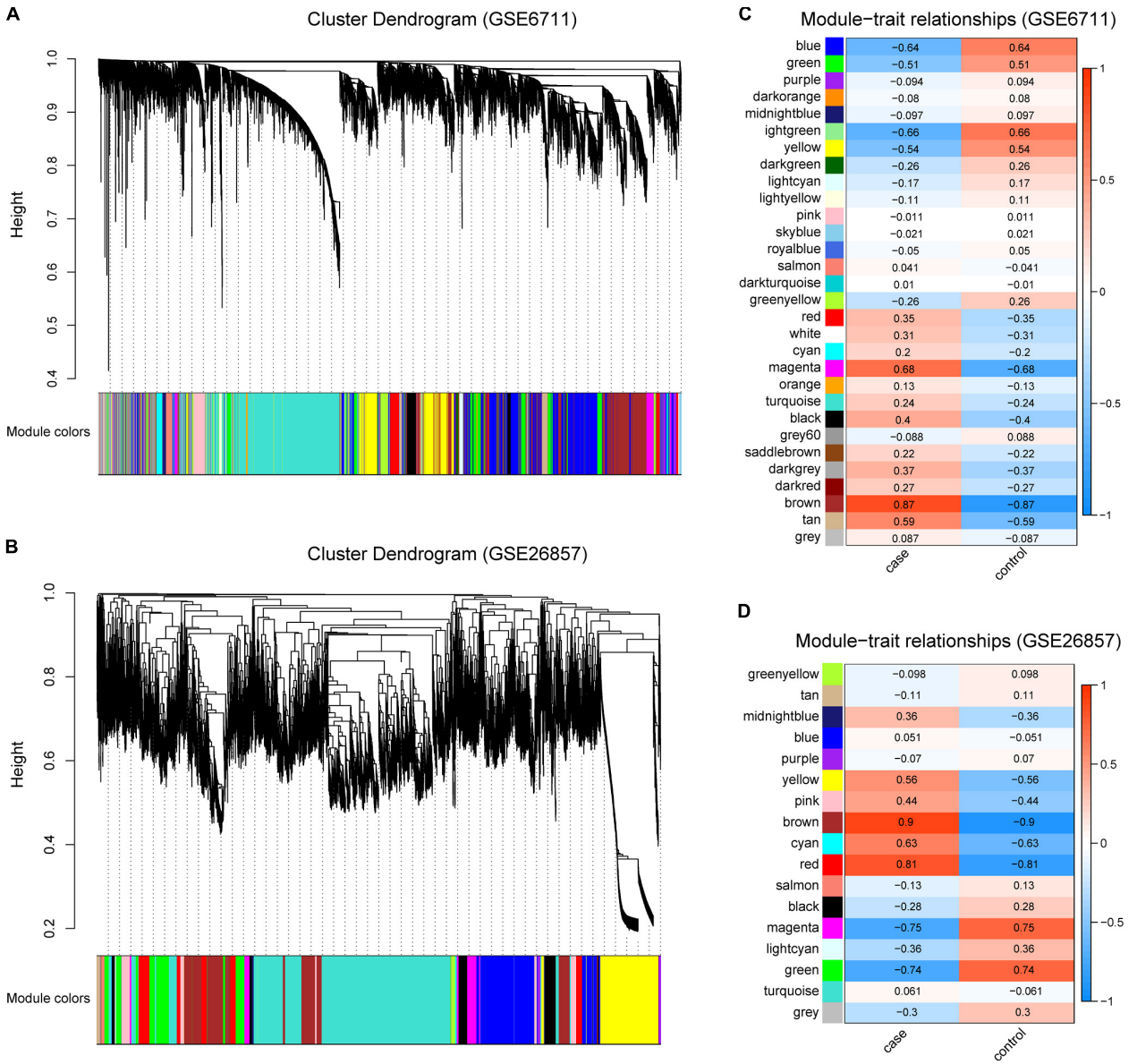


FIGURE 2 | Identification of Glucocorticoid-related Gene Modules. **(A,B)** Gene modules identified by WGCNA in GSE6711 **(A)** and GSE26857 **(B)**. **(C,D)** The correlation between gene modules and trait in GSE6711 **(C)** and GSE26857 **(D)**, case represented glucocorticoid-treated and control represented untreated.

tool (Chen et al., 2013; Kuleshov et al., 2016). The significance threshold is $P < 0.05$.

RESULTS

Identification of Glucocorticoid-Related Gene Modules

Glucocorticoid-induced osteoporosis is the most common form of secondary osteoporosis (Whittier and Saag, 2016; Coskun Benlidayi, 2018). To explore the correlation between glucocorticoid-induced gene expression disorder and osteoporosis, we collected two microarray datasets of glucocorticoid-related osteosarcoma cell lines from GEO database. The two datasets contained 66 samples, 48 of which were treated with glucocorticoids. According to the result of sample hierarchical clustering, we removed two outliers in GSE6711 (**Figure 1**).

In order to obtain glucocorticoid-related genes, R package WGCNA was utilized. For all gene expression profiles, a total of 29 and 16 gene modules were identified in GSE6711 (**Figure 2A**) and GSE26857 (**Figure 2B**), respectively. The correlation between gene modules and traits suggested that three and four gene modules were correlated with glucocorticoids ($R > 0.5$, **Figures 2C,D**). Of these, brown was the most correlated module, followed by magenta module (GSE6711) and red module

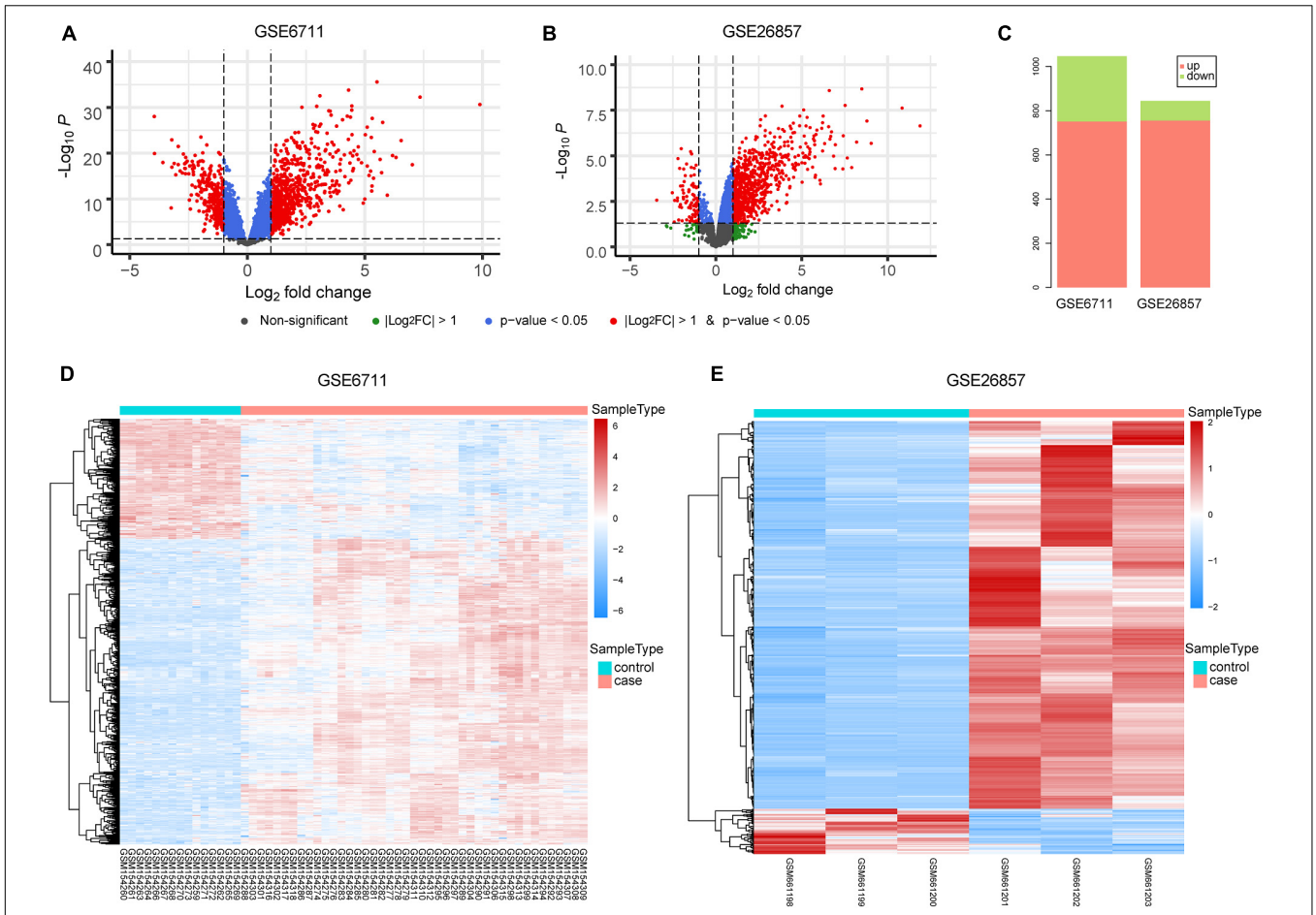


FIGURE 3 | Differential Expression Analysis of Glucocorticoid-related Genes. **(A,B)** The volcano plot of glucocorticoid-related genes in GSE6711 **(A)** and GSE26857 **(B)**. **(C)** Barplot of differentially expressed glucocorticoid-related genes. **(D,E)** Heatmap of differentially expressed glucocorticoid-related gene expression profile.

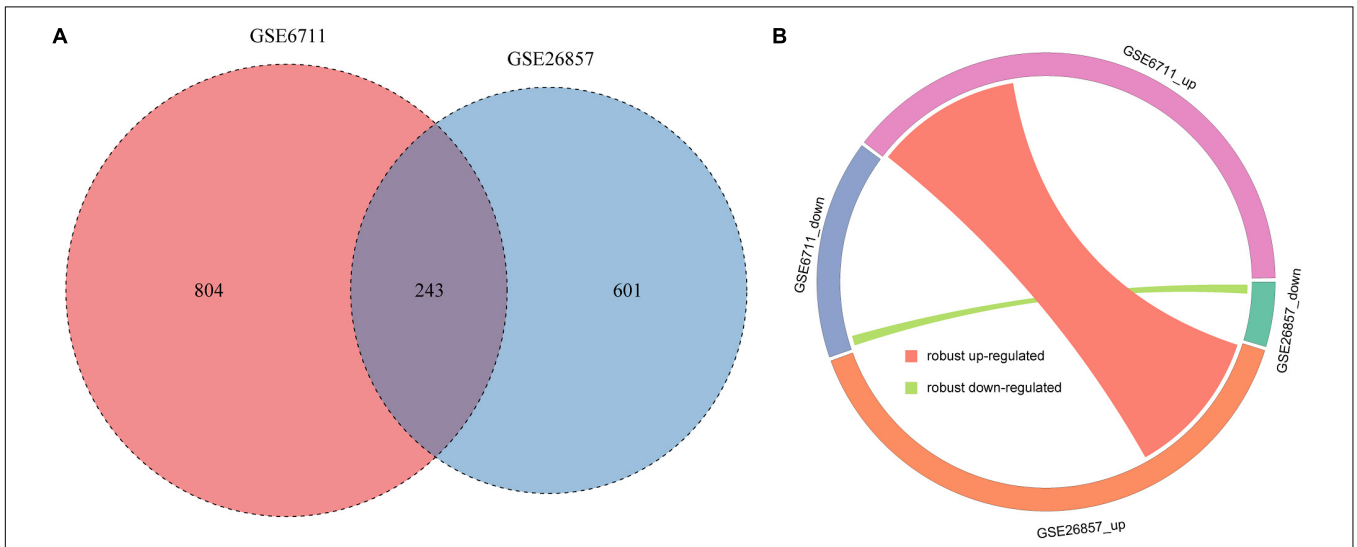


FIGURE 4 | Integration of glucocorticoid-related DEGs. **(A)** The intersection of glucocorticoid-related DEGs in two datasets. **(B)** Circos plot of robust glucocorticoid-related DEGs.

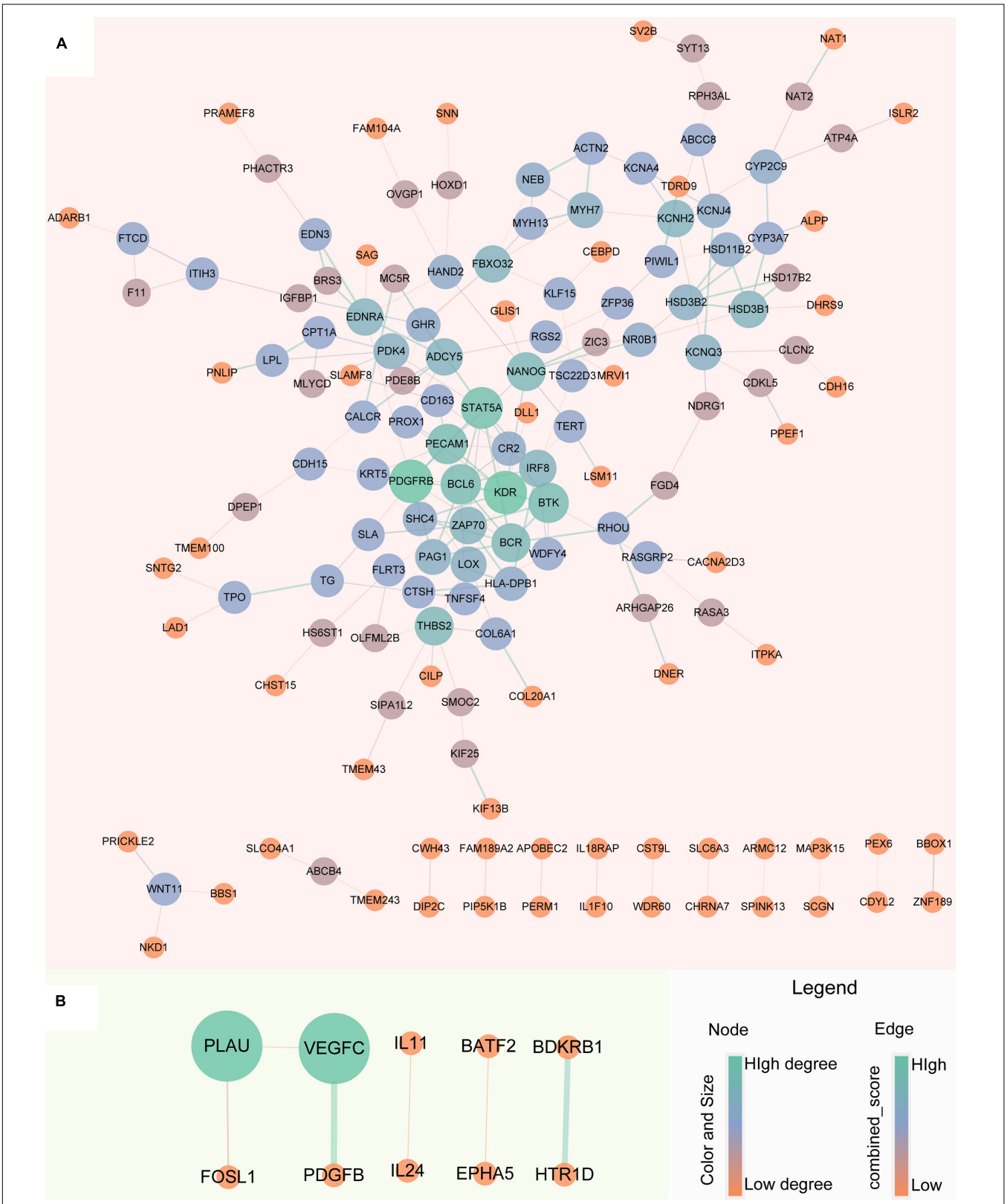


FIGURE 5 | Protein-Protein Interaction network of robust glucocorticoid-related DEGs. **(A)** PPI network of robust up-regulated glucocorticoid-related DEGs. **(B)** PPI network of robust down-regulated glucocorticoid-related DEGs. The size and color of the node depends on the degree, the larger the degree, the larger the node. The width of the edge is determined by the combined score of STRING, the larger the score, the wider the edge.

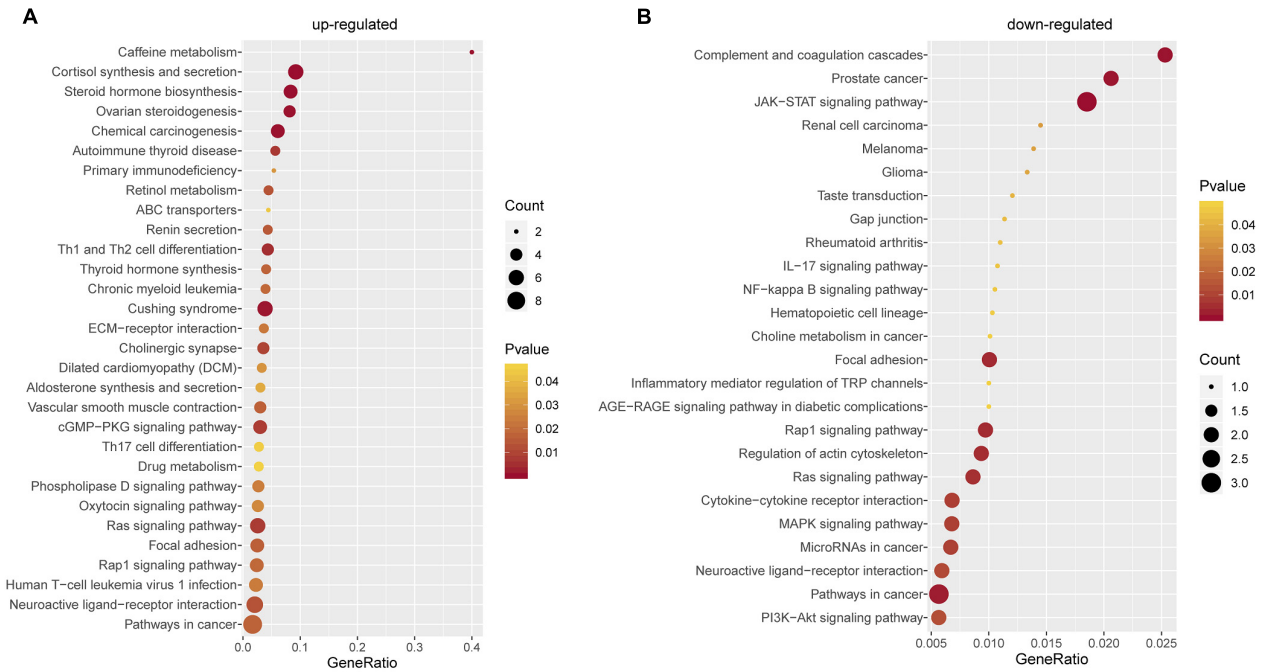


FIGURE 6 | Pathway Enrichment Analysis. **(A)** Pathway enrichment analysis of up-regulated glucocorticoid-related DEGs in PPI network. **(B)** Pathway enrichment analysis of down-regulated glucocorticoid-related DEGs in PPI network.

(GSE26857). In total, 3318 (GSE6711) and 5880 (GSE26857) genes were identified as glucocorticoid-related genes.

Differential Expression Analysis of Glucocorticoid-Related Genes

The occurrence of diseases is often accompanied by gene expression disorders. We further analyzed the differential expression of glucocorticoid-related genes (**Figures 3A,B**). In GSE6711 dataset, 31.6% (1047/3318) genes were differentially expressed, including 751 up-regulated and 296 down-regulated (**Figure 3C**). In GSE26857 dataset, 14.4% (844/5880) genes were differentially expressed, including 756 up-regulated and 88 down-regulated (**Figure 3C**). It can be found that most of glucocorticoid-related DEGs were differentially up-regulated. Furthermore, the heatmap showed that DEGs were able to group the samples by sample types, which is glucocorticoid-treated (case) and untreated (control) (**Figures 3D,E**).

Integration of Glucocorticoid-Related DEGs

Due to the data heterogeneity, there are some differences in the analysis results of different datasets. Therefore, integrative analysis of different datasets can get more robust results. We integrated the common glucocorticoid-related DEGs of two datasets. A total of 243 robust glucocorticoid-related DEGs were obtained, which accounted for 23.2 and 28.8% of glucocorticoid-related DEGs in two datasets (**Figure 4A**). Of the robust glucocorticoid-related DEGs, 242 genes were consistent in their deregulation directions between two datasets, including 227 up-regulated and 15 down-regulated genes (**Figure 4B**).

Protein-Protein Interaction Network of Robust Glucocorticoid-Related DEGs

The robust up-regulated and down-regulated DEGs were used to construct PPI networks, respectively. The results suggested that there are 148 up-regulated and 10 down-regulated DEGs have PPI relationship (**Figure 5**). The degree distribution of up-regulated DEGs PPI network ranges from 1 to 10 and the top 16 hub genes which degree greater than 5 can be found in **Supplementary Figure S1**.

To better understand the biological characteristics of robust glucocorticoid-related DEGs, we performed pathway enrichment analysis using up or down-regulated genes included in PPI networks by Enrichr, an online enrichment analysis tool. The up-regulated genes were significantly enriched in 30 KEGG pathways, including cushing syndrome, cortisol synthesis and secretion, Th1 and Th2 cell differentiation and Th17 cell differentiation (**Figure 6A**). Cushing's syndrome is an endocrine disorder characterized by excessive cortisol secretion (Jiang et al., 2014; Pivonello et al., 2016). Many studies have shown that increased cortisol levels are associated with decreased bone mineral density (van Schoor et al., 2007; Osella et al., 2012; Toth and Grossman, 2013). The evidence for a relationship between the immune and the skeletal systems has long been recognized (Takayanagi, 2010; Yuan et al., 2010, 2011). Activated Th cells are important source of osteoclasts under inflammatory conditions (Rifas and Weitzmann, 2009). It is well known that osteoclasts

are responsible for bone resorption and play an important role in osteoporosis (Detsch and Boccaccini, 2015; Kylmaoja et al., 2016; Zhang et al., 2020).

There were 25 pathways significantly enriched in down-regulated glucocorticoid-related DEGs (**Figure 6B**). Akt is the key element in osteoblast differentiation (Sugatani and Hruska, 2005). PI3K-Akt signaling pathway is one of the significant pathways, which can inhibit osteoporosis by promoting proliferation, differentiation and osteogenesis of osteoblasts (Xi et al., 2015). In addition, down-regulated genes were also enriched in multiple cancer-related pathways, it has been found that glucocorticoids play a role in the treatment of cancer (Pufall, 2015; Hu and Chen, 2017; McNamara et al., 2018).

DISCUSSION

In this study, we performed an integrative analysis of glucocorticoid-related microarray datasets. The WGCNA method enables the identification of glucocorticoid-related genes. Then differential expression analysis screened out the glucocorticoid-related dysregulated genes. By further integration, the robust dysregulated genes were obtained. Through protein-protein interaction network analysis of robust dysregulated genes, 158 candidate genes were obtained. And they significantly enriched the osteoporosis-related pathway.

In summary, we used microarray data to identify glucocorticoid-related dysregulated genes, which are associated with osteoporosis-related pathway. Our findings elucidate the expression mechanism of glucocorticoid-related genes and provide new guides for the diagnosis and treatment of glucocorticoid-induced osteoporosis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The glucocorticoid-related osteosarcoma U-2 OS bone cells microarray datasets of GSE6711 and GSE26857 were downloaded from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/).

AUTHOR CONTRIBUTIONS

HL and GM designed the experiments. XJ obtained data from GEO. XY, XJ, PW, YH, HZ, XR, SC, WF, PZ, CC, and GM analyzed the data. XJ and HL wrote the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61871068 and 81671827).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.00252/full#supplementary-material

FIGURE S1 | The degree distribution of up-regulated DEGs PPI network.

REFERENCES

Canales, R. D., Luo, Y., Willey, J. C., Austermiller, B., Barbacioru, C. C., Boysen, C., et al. (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* 24, 1115–1122. doi: 10.1038/nbt1236

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 Bioinformatics* 14:128. doi: 10.1186/1471-2105-14-128

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., 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

Coskun Benlidayi, I. (2018). Denosumab in the treatment of glucocorticoid-induced osteoporosis. *Rheumatol. Int.* 38, 1975–1984. doi: 10.1007/s00296-018-4106-1

Coughlan, T., and Dockery, F. (2014). Osteoporosis and fracture risk in older people. *Clin. Med.* 14, 187–191. doi: 10.7861/clinmedicine.14-2-187

Detsch, R., and Boccaccini, A. R. (2015). The role of osteoclasts in bone tissue engineering. *J. Tissue Eng. Regen. Med.* 9, 1133–1149. doi: 10.1002/term.1851

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

Glaser, D. L., and Kaplan, F. S. (1997). Osteoporosis. definition and clinical presentation. *Spine* 22(24 Suppl.), 12S–16S. doi: 10.1097/00007632-199712151-00003

Hu, J., and Chen, Q. (2017). The role of glucocorticoid receptor in prostate cancer progression: from bench to bedside. *Int. Urol. Nephrol.* 49, 369–380. doi: 10.1007/s11255-016-1476-8

Jewell, C. M., Scoltock, A. B., Hamel, B. L., Yudt, M. R., and Cidlowski, J. A. (2012). Complex human glucocorticoid receptor dim mutations define glucocorticoid induced apoptotic resistance in bone cells. *Mol. Endocrinol.* 26, 244–256. doi: 10.1210/me.2011-1116

Jiang, Q., Wang, J., Wang, Y., Ma, R., Wu, X., and Li, Y. (2014). TF2LncRNA: identifying common transcription factors for a list of lncRNA genes from ChIP-Seq data. *Biomed Res. Int.* 2014:317642. doi: 10.1155/2014/317642

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

Kylmaoja, E., Nakamura, M., and Tuukkanen, J. (2016). Osteoclasts and remodeling based bone formation. *Curr. Stem Cell Res. Ther.* 11, 626–633. doi: 10.2174/1574888x10666151019115724

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

Liu, Y. Z., Dvornyk, V., Lu, Y., Shen, H., Lappe, J. M., Recker, R. R., et al. (2005). A novel pathophysiological mechanism for osteoporosis suggested by an in vivo gene expression study of circulating monocytes. *J. Biol. Chem.* 280, 29011–29016. doi: 10.1074/jbc.M501164200

Lu, N. Z., Collins, J. B., Grissom, S. F., and Cidlowski, J. A. (2007). Selective regulation of bone cell apoptosis by translational isoforms of the glucocorticoid receptor. *Mol. Cell Biol.* 27, 7143–7160. doi: 10.1128/MCB.00253-07

McNamara, K. M., Kannai, A., and Sasano, H. (2018). Possible roles for glucocorticoid signalling in breast cancer. *Mol. Cell Endocrinol.* 466, 38–50. doi: 10.1016/j.mce.2017.07.004

Osella, G., Ventura, M., Ardito, A., Allasino, B., Termine, A., Saba, L., et al. (2012). Cortisol secretion, bone health, and bone loss: a cross-sectional and prospective study in normal non-osteoporotic women in the early postmenopausal period. *Eur. J. Endocrinol.* 166, 855–860. doi: 10.1530/EJE-11-0957

Pivonello, R., Isidori, A. M., De Martino, M. C., Newell-Price, J., Biller, B. M., and Colao, A. (2016). Complications of Cushing's syndrome: state of the art. *Lancet Diabetes Endocrinol.* 4, 611–629. doi: 10.1016/S2213-8587(16)00086-3

Pufall, M. A. (2015). Glucocorticoids and Cancer. *Adv. Exp. Med. Biol.* 872, 315–333. doi: 10.1007/978-1-4939-2895-8_14

Rifas, L., and Weitzmann, M. N. (2009). A novel T cell cytokine, secreted osteoclastogenic factor of activated T cells, induces osteoclast formation in a RANKL-independent manner. *Arthritis Rheum* 60, 3324–3335. doi: 10.1002/art.24877

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

Sugatani, T., and Hruska, K. A. (2005). Akt1/Akt2 and mammalian target of rapamycin/Bim play critical roles in osteoclast differentiation and survival, respectively, whereas Akt is dispensable for cell survival in isolated osteoclast precursors. *J. Biol. Chem.* 280, 3583–3589. doi: 10.1074/jbc.M410480200

Sun, M. A., Shao, X., and Wang, Y. (2018). Microarray data analysis for transcriptome profiling. *Methods Mol. Biol.* 1751, 17–33. doi: 10.1007/978-1-4939-7710-9_2

Takayanagi, H. (2010). New immune connections in osteoclast formation. *Ann. N.Y. Acad. Sci.* 1192, 117–123. doi: 10.1111/j.1749-6632.2009.05303.x

Toth, M., and Grossman, A. (2013). Glucocorticoid-induced osteoporosis: lessons from Cushing's syndrome. *Clin .Endocrinol.* 79, 1–11. doi: 10.1111/cen.12189

van Schoor, N. M., Dennison, E., Lips, P., Uitterlinden, A. G., and Cooper, C. (2007). Serum fasting cortisol in relation to bone, and the role of genetic variations in the glucocorticoid receptor. *Clin. Endocrinol.*67, 871–878. doi: 10.1111/j.1365-2265.2007.02978.x

von Mering, C., 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

Whittier, X., and Saag, K. G. (2016). Glucocorticoid-induced osteoporosis. *Rheum Dis. Clin. North Am.* 42, 177–189. doi: 10.1016/j.rdc.2015.08.005

Xi, J. C., Zang, H. Y., Guo, L. X., Xue, H. B., Liu, X. D., Bai, Y. B., et al. (2015). The PI3K/AKT cell signaling pathway is involved in regulation of osteoporosis. *J. Recept. Signal Transduct. Res.* 35, 640–645. doi: 10.3109/10799893.2015.1041647

Xiao, P., Chen, Y., Jiang, H., Liu, Y. Z., Pan, F., Yang, T. L., et al. (2008). In vivo genome-wide expression study on human circulating B cells suggests a novel ESR1 and MAPK3 network for postmenopausal osteoporosis. *J. Bone Miner. Res.* 23, 644–654. doi: 10.1359/jbmr.080105

Yuan, F. L., Hu, W., Lu, W. G., Li, X., Li, J. P., Xu, R. S., et al. (2011). Targeting interleukin-21 in rheumatoid arthritis. *Mol. Biol. Rep.* 38, 1717–1721. doi: 10.1007/s11033-010-0285-x

Yuan, F. L., Li, X., Lu, W. G., Xu, R. S., Zhao, Y. Q., Li, C. W., et al. (2010). Regulatory T cells as a potent target for controlling bone loss. *Biochem. Biophys. Res. Commun.* 402, 173–176. doi: 10.1016/j.bbrc.2010.09.120

Zhang, Y., Cui, Y., Wang, L., and Han, J. (2020). Autophagy promotes osteoclast podosome disassembly and cell motility athrough the interaction of kindlin3 with LC3. *Cell Signal.* 67:109505. doi: 10.1016/j.cellsig.2019.109505

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 © 2020 Ying, Jin, Wang, He, Zhang, Ren, Chai, Fu, Zhao, Chen, Ma 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 17 April 2020
doi: 10.3389/fbioe.2020.00266

Characterization of lncRNA-Associated ceRNA Network to Reveal Potential Prognostic Biomarkers in Lung Adenocarcinoma

*Yang Wang, Ruyi He and Lixin Ma**

State Key Laboratory of Biocatalysis and Enzyme Engineering, Hubei Key Laboratory of Industrial Biotechnology, Hubei University, Wuhan, China

Lung adenocarcinoma (LUAD) is one of the most fatal malignant tumors harmful to human health. The complexity and behavior characteristics of long-non-coding RNA (lncRNA)-associated competing endogenous RNA (ceRNA) network in LUAD patients are still unclear. The purpose of this study was to elucidate the regulatory networks of dysregulated RNAs, view, and identify potential prognosis signatures involved in LUAD. The expression profiles of mRNAs, lncRNAs, and miRNAs were obtained from the TCGA database. In total, 2078 DEmRNAs, 257 DElncRNAs, and 101 DEmiRNAs were sorted out. A PPI network including 45 DEmRNAs was constructed. Ten hub genes in the PPI network associated with cell cycle-related pathways were identified and they played key roles in regulating cell proliferation. A total of three DEmiRNAs, seven DElncRNAs, and six DEmRNAs were enrolled in the ceRNA network. Except for certain genes without any published study reports, all the genes in the ceRNA network played an essential role in controlling tumor cell proliferation and were associated with prognosis in LUAD. Finally, based on step regression and Cox regression survival analysis, we identified four candidate biomarkers, including miR490, miR1293, LINC01740, and IGF2BP1, and established a risk model based on the four genes. Our study provided a global view and systematic dissection of the lncRNA-associated ceRNA network, and the identified four genes might be novel important prognostic factors involved in LUAD pathogenesis.

Keywords: lung adenocarcinoma, signature, ceRNA, lncRNA, prognosis

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Shan Cong,
University of Pennsylvania, United States
Tao Han,
The Ohio State University, United States

***Correspondence:**
Lixin Ma
malixing@hubu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *27 December 2019*
Accepted: *13 March 2020*
Published: *17 April 2020*

Citation:
Wang Y, He R and Ma L (2020) Characterization of lncRNA-Associated ceRNA Network to Reveal Potential Prognostic Biomarkers in Lung Adenocarcinoma. Front. Bioeng. Biotechnol. 8:266. doi: 10.3389/fbioe.2020.00266

INTRODUCTION

Despite medical advances, lung cancer remains the leading cause of cancer deaths. Lung cancer is usually recognized late in its natural history and have a poor prognosis, with an overall 5-year survival rate of 10–15% (Cagle et al., 2013). The recognition of histologic subtypes of non-small cell lung carcinoma (NSCLC), namely, adenocarcinoma, squamous cell carcinoma, and large cell lung carcinoma as the most frequent subtypes, has become important as a determinant of therapy in this disease (Kerr et al., 2014). In addition, in recent years, the identification of molecular abnormalities in a large proportion of patients with lung cancer has allowed the emergence of personalized targeted therapies and has opened new horizons and created new expectations for these patients (Ezeife and Leighl, 2018). The use of predictive biomarkers to identify tumors that

could respond to targeted therapies has meant a change in the paradigm of lung cancer diagnosis (Majeed and Amir, 2018).

Currently, the rapid advancement of high-throughput technologies offers great opportunities for biomarker identification (Yu et al., 2018). Non-coding RNAs as biomarker and therapeutic targets play a significant role in human disease (Zhou et al., 2018b,c). Among which, long-non-coding RNA (lncRNAs) are a class of RNA molecules with more than 200 nucleotides in length and have no evident open reading frames (Fatica and Bozzoni, 2014). These long molecules are dysregulated among cancers (Yan et al., 2015) and play key roles in gene regulation and carcinogenesis, including proliferation, survival, migration, and genomic stability (Gutschner et al., 2013; Castro-Oropeza et al., 2018). It is believed that the clinical value of lncRNA is not confined to candidate biomarkers for diagnostic and prognostic purposes (Shi et al., 2018).

In Salmena et al. (2011) put forward a competing endogenous RNA (ceRNA) hypothesis. Subsequently, several studies also mentioned that there is an interplay between lncRNAs and miRNAs during the tumorigenic process, among which lncRNAs serve as molecular sponges for miRNAs (Liz and Esteller, 2016). For example, KCNQ1OT1 promotes cell proliferation and autophagy and inhibits cell apoptosis via regulating miR204-5p/ATG3 axis, providing a promising target for NSCLC therapy (Kang et al., 2019). Guo et al. reported that LINC00173 up-regulated Etk through functioning as a ceRNA by "sponging" miRNA-218 and led to the up-regulation of GSKIP and NDRG1 in small cell lung cancer (Zeng et al., 2019). LncRNA AGAP2-AS1 up-regulates ANXA11 expression by sponging miR16-5p and promotes proliferation and metastasis in hepatocellular carcinoma (Liu et al., 2019). Thus, the discovery of lncRNA–miRNA–mRNA networks may lead to a more comprehensive understanding of the etiology and metastasis mechanism of cancer. However, the complexity and behavior of lncRNA-associated ceRNA network remain poorly characterized in lung adenocarcinoma (LUAD).

In this study, by comprehensively integrating gene and miRNA expression data of LUAD, the LUAD-related lncRNA–miRNA–mRNA competitive network was established. We analyzed and predicted the functions of ceRNA and PPI networks and established a Cox regression model to predict the overall survival of patients with lung cancer. Finally, four predictive genes were identified, including LINC01740, mir1293, mir490, and IGF2BP1, which could contribute to LUAD. This study will contribute to understanding the molecular mechanism and provide new therapeutic targets for LUAD.

MATERIALS AND METHODS

Data Preparation and Differentially Expressed Gene Analysis

All primitive data of LUAD from The Cancer Genome Atlas (TCGA) database[1] were download through GDC Data Transfer Tool, including RNA-seq and miRNA-seq of Transcriptome profiling and Clinical data. EdgeR package (3.3.3 version) (Robinson et al., 2010) in R software was used to analyze and identify differentially expressed RNAs (DERNAs, including DEmRNAs and DElncRNAs) and differentially expressed microRNA (DEmiRNAs) with the thresholds of $|\log2FoldChange| > 2.0$ and FDR (adjusted p value) < 0.01. Then, biomart in R package was used to annotate DEmRNAs and DElncRNAs. The heatmap and volcano plot were constructed by the ggplot2 package in R software (Zhou et al., 2017).

Functional Enrichment Analysis

clusterProfiler (Yu et al., 2012) package in R was used to make the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology (GO) enrichment analysis, including biological process (BP), the cellular component (CC), and molecular function (MF). Pathview (Luo and Brouwer, 2013) and enrichplot packages (Ito and Murphy, 2013) were used to visualize the enrichment results. A significance level of adjusted $p < 0.05$ was set as the cutoff criteria.

Protein–Protein Interaction Analysis

The DEmRNAs were enrolled in a protein–protein interaction (PPI) network through the STRING (version 11.0) database[2] with a confidence score >0.9. Furthermore, genes with degree ≥ 25 were selected as hub genes, and we focused the interaction types among proteins only on physical interaction and co-expression (Sun et al., 2019). Subsequently, GO and KEGG analyses of the PPI network modules were carried out using clusterProfiler package in R.

Construction of the ceRNA Network

According to the hypothesis of ceRNA, a lncRNA–miRNA–mRNA network was constructed (Zhou et al., 2018a). Relevant miRNA-target data were obtained from the miRTarBase, and the support types of targeting were only focused on experiments, including luciferase reporter assay, Western blot, Northern blot, or qRT-PCR. Only the miRNA targets that were differentially expressed between tumor and normal tissue were considered for the next analysis step. Furthermore, the candidate DElnRNA–DEmiRNA interactions were selected based on miRcode database and the following model:

$$Y_{\text{target}} = \beta_0 + \beta_1 * miRNA + \beta_2 * lncRNA + \beta_3 * miRNA * lncRNA + \varepsilon \quad (1)$$

where$miRNA$, $lncRNA$, and Y_{target} are the gene expression of miRNA, lncRNA, and miRNA targets, respectively. $\beta_1$and β_2 represent the effect of miRNA and lncRNA, respectively, on target by themselves alone (main effects), while β_3 represents the effect of miRNA–lncRNA interaction. If a lncRNA and miRNA interaction has effects on target expression outcomes, we expect $\beta_3$to be non-zero.Here, all the miRNAs andlncRNAs and miRNA targets should be differentially expressed between tumor and normal tissue.

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

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

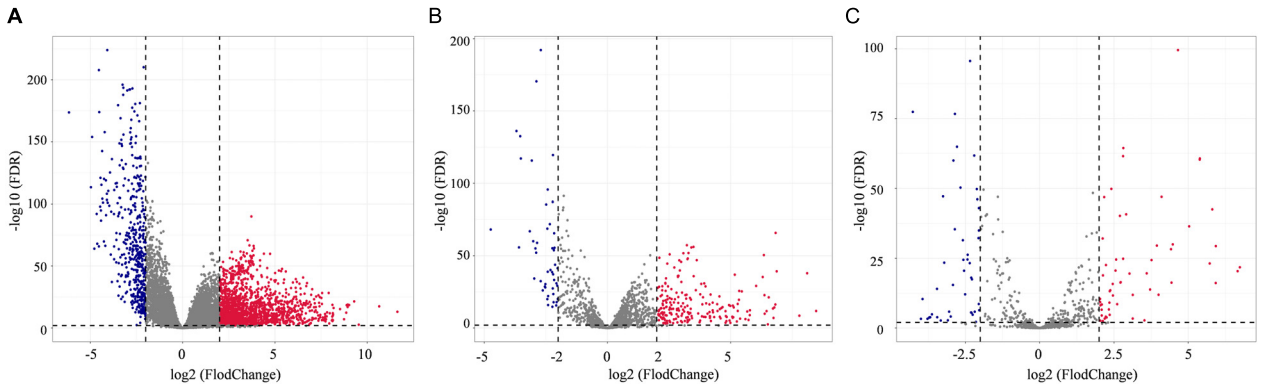


FIGURE 1 | Distribution of differentially expressed genes in lung adenocarcinoma (LUAD) (| log2FoldChange| > 2.0 and adjusted *p* value < 0.01) between 533 tumor tissues and 59 normal tissues. The volcano plots described 2378 DEmRNAs **(A)**, 357 DElncRNAs **(B)**, and 101 DEmiRNAs **(C)**. Red stands for up-regulations, blue stands for down-regulations, and gray stands for normal expression in volcanoes. Each point represents a gene.

Biomarkers Screening and Validation

The status and survival time of LUAD patients were extracted from the TCGA clinical dataset. Subsequently, the DEmRNAs, DElncRNAs, and DEmiRNAs identified in ceRNAs were selected for screening biomarkers. We used univariate Cox regression to screen prognostic factors ($p < 0.05$), and those prognostic factors whose expression levels were significantly relevant to patients' overall survival ($p < 0.05$) were selected as primitive biomarkers (Zhou et al., 2018b; Bao et al., 2019).

Cox Risk Regression Establishment and Validation

The lncRNAs, mRNAs, and miRNAs raw data were transformed and normalized in a log2[cpm(x) + 1] manner. Univariate cox regression was used to select prognosis-associated genes ($p < 0.05$) (Zhou et al., 2018a). Subsequently, we performed Cox regression analysis combined with stepwise regression to establish a Cox risk model (Zhou et al., 2018a). Finally, a validation set and Kaplan–Meier survival curves along with a logrank p test were applied to validate its accuracy (Zhou et al., 2017; Sun et al., 2019).

RESULTS

Identification of Differentially Expressed Genes

RNA expression profiles and corresponding clinical data of 533 cohort LUAD patients and 59 normal controls were downloaded from the TCGA database. Meanwhile, miRNA-seq data corresponding to 561 patients' clinical information, including 515 cohort LUAD patients and 46 normal controls, were obtained from TCGA. In total, 60,483 transcripts and 1046 miRNAs were obtained. With the cutoff criteria unified, CPM(gene) > 1, rowSum(CPM) ≥ 2, 32,495 transcripts and 613 miRNAs were selected for the differentially expressed analysis. After filtering, 5624 DERNAs and 673 DEmiRNAs were identified with the thresholds of | log2FoldChange| > 2.0 and FDR (adjusted p value) < 0.01.

In total, 2078 DEmRNAs (1612 up-regulated and 466 down-regulated, **Figure 1A**), 257 DElncRNAs (209 up-regulated and 48 down-regulated, **Figure 1B**), and 101 DEmiRNAs (56 up-regulated and 45 down-regulated, **Figure 1C**) were sorted out.

Functional Analysis of DERNAs

Gene ontology and KEGG enrichment analyses were used to explore the potential function of DERNAs. Ten biological pathways were highly enriched within cutoff p value < 0.05. Among them, 12% DERNAs were enriched in GPCR ligand binding process, and 9.5% DERNAs were enriched in Class A/1 (Rhodopsin-like receptors) pathway, and 6.7% DERNAs were enriched in peptide ligand binding receptor pathways (**Figure 2A**). Detailed information of these enriched pathways and associated genes is summarized in **Table 1**. The GO functional enrichment analysis results of DERNAs including MF, CC, and BP were described in **Figures 2B–D**. The results show that the genes mainly focused on receptor ligand activity function, extracellular matrix, and morphogenesis of an epithelium process.

Furthermore, KEGG pathway enrichment analysis results demonstrated that the most significantly enriched pathways were neuroactive ligand–receptor interaction, alcoholism, and systemic lupus erythematosus pathways (**Figure 2E**). The pathway–pathway interaction network (PPIN) was constructed based on the DERNAs enriched in same pathway (**Figure 2F**). Four pathways were identified in the PPIN, including alcoholism, maturity onset diabetes of the young, neuroactive ligand–receptor interaction, and systemic lupus erythematosus pathway. We noticed that, all the DERNAs enriched in systemic lupus erythematosus and alcoholism pathways were up-regulated, except for gene 2354, whose gene symbol is "FOSB." Gene annotation of FOSB shows that it was a proto-oncogene, and it has been implicated as regulators of cell proliferation, differentiation, and transformation. Similarly,

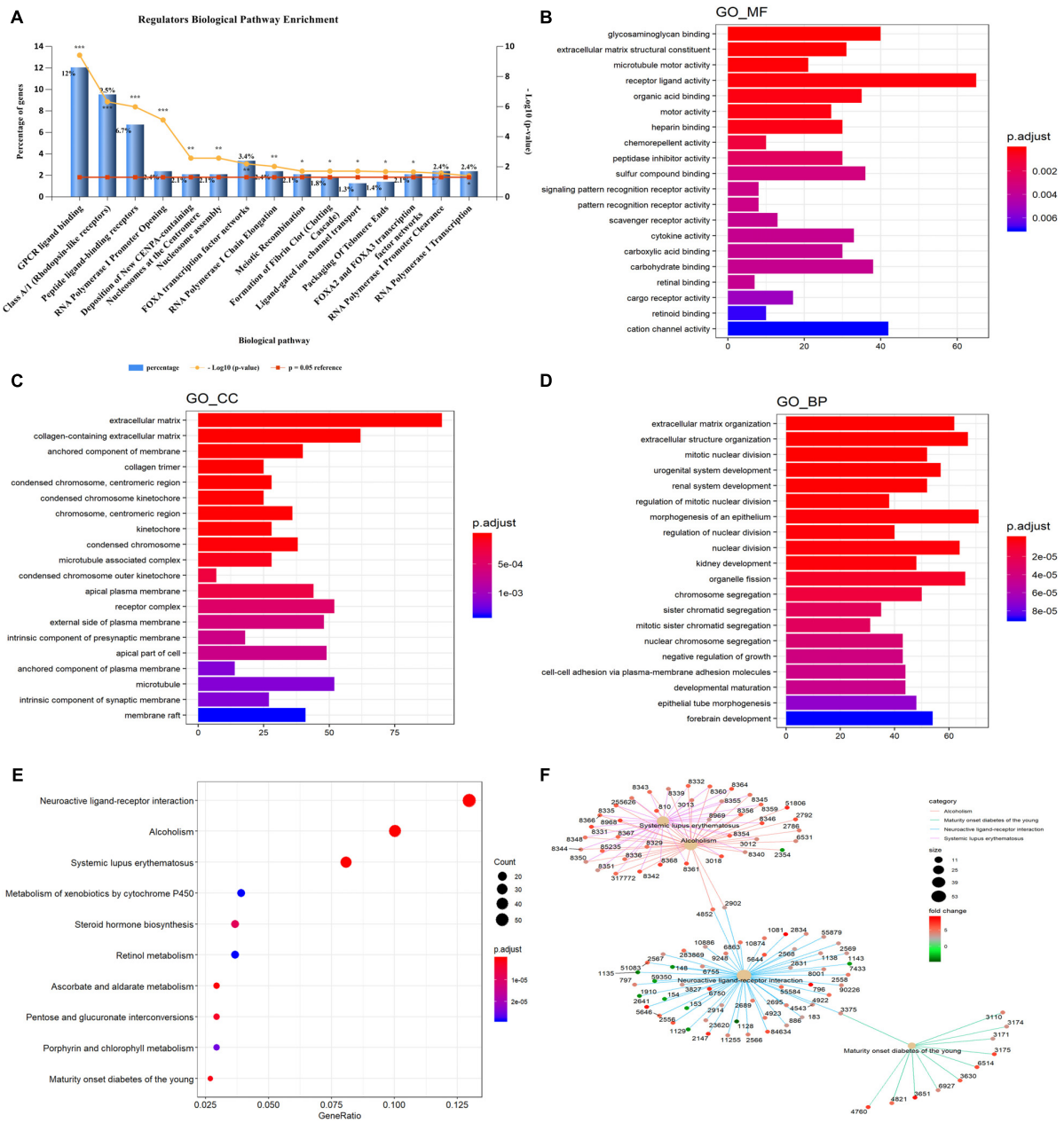


FIGURE 2 | GO and KEGG pathway enrichment analysis of DERNAs. **(A)** The statistical results of genes enriched in biological pathways. The *y* axis on the left represents the percentage of genes in each biological pathway, the *y* axis on the right is –log10(*p* value) of each enrichment pathway, and the *x* axis represents the pathways categories. GO analysis contains the molecular function **(B)**, cellular component **(C)**, biological process **(D)**, the *y* axis represents the number of target genes, and the *x* axis represents the GO categories. **(E)** The most important KEGG pathways in DERNAs. The *y* axis represents the pathways, and the *x* axis represents enriched gene numbers. The circle size represents the counts of genes in each pathway and the color means adjusted *p* value. **(F)** The netplot of KEGG pathways means enrichment of genes in different pathways. The number adjacent to nodes stands for gene ID. The color bar represents the fold change of genes in different pathways. $^{*}p < 0.05$, $^{**}p < 0.005$, and $^{***}p < 0.0005$.

all the genes enriched in maturity onset diabetes of young pathway were all up-regulated in LUAD. Furthermore, the results showed that gene 4852 and 2092 were both enriched in alcoholism pathway and neuroactive ligand–receptor interaction pathway, and played vital roles in connecting the two pathways.

PPI Network Analysis

A total of 55 proteins and 453 edges, including 45 DEmRNAs, were selected in the PPI network. A total of 10 hub genes, including CDK1, TOP2A, PBK, CDCA8, CDC20, KIF20A, DLGAP5, NDC80, NCAPG, and CCNA2, were selected from the PPI network with degree ≥25 and combined score

TABLE 1 | Enriched biological pathways and associated genes.

Biological pathway	No. of genes in DERNAs	No. of genes in background dataset	Percentage of genes	Fold enrichment	*p* value	*p*-adjusted (Bonferroni)
GPCR ligand binding	86	336	12.0	2.2	5.05E-14	8.42E-11
Class A/1 (rhodopsin-like receptors)	68	274	9.5	2.1	1.26E-10	2.1E-07
Peptide ligand-binding receptors	48	167	6.7	2.5	4.13E-10	6.89E-07
RNA polymerase I promoter opening	17	31	2.3	4.8	4.09E-09	6.83E-06
Deposition of new CENPA-containing nucleosomes at the centromere	15	35	2.1	3.7	2.11E-06	0.003
Nucleosome assembly	15	35	2.1	3.7	2.11E-06	0.003
FOXA transcription factor networks	24	81	3.3	2.6	5.93E-06	0.009
RNA polymerase I chain elongation	17	48	2.3	3.1	1.00E-05	0.016
Meiotic recombination	15	42	2.1	3.1	2.99E-05	0.049
Formation of fibrin clot (clotting cascade)	13	33	1.8	3.4	3.05E-05	0.051
Ligand-gated ion channel transport	9	17	1.2	4.6	3.11E-05	0.051
Packaging of telomere ends	10	21	1.4	4.1	3.65E-05	0.060
FOXA2 and FOXA3 transcription factor networks	15	43	2.1	3.0	4.12E-05	0.068
RNA polymerase I promoter clearance	17	54	2.3	2.7	5.79E-05	0.096
RNA polymerase I transcription	17	56	2.3	2.6	9.67E-05	0.161

>0.9 (**Figure 3A**). Furthermore, the association among these interacted proteins should be physical interaction or co-expressed with each other (**Figure 3B**). We noticed that eight RNA expression levels were significantly associated with overall survival outcomes except for CDCA8 and CDC20 (**Figures 3C,D**). Pathway enrichment analysis results of the 10 hub genes are summarized in **Table 2**.

Construction of the ceRNA Network in LUAD

A total of seven DElncRNAs, six DEmRNAs, and three DEmiRNAs were enrolled in the ceRNA network (**Figure 4**). miRTarBase was used to predict the miRNA–mRNA pairs (**Table 3**). We only focused on those miRNA–mRNA pairs whose interaction evidence was validated by experiments, including luciferase reporter assay, Western blot, Northern blot, or qRT-PCR.

Then, we employed a simple linear regression model combined with miRcode database to predict the potential miRNA target by DElncRNAs [see Methods 2.4 model (1)]. In the model, we specified that the input of lncRNAs should be (i) differentially expressed between tumor and normal tissues; (ii) lncRNA expression is associated with overall survival outcomes (logrank p value < 0.05). Finally, 7 of 53 DElncRNAs, 6 of 340 DEmRNAs, and 3 of 9 DEmiRNAs formed the ceRNA network. Detailed information about their expression and association with overall survival outcomes is listed in **Table 4**.

Screen Biomarkers and Construction Risk Model

Three DEmiRNAs, seven DElncRNAs, and six DEmRNAs in the ceRNA network were selected as candidate biomarkers for the following step analysis. Subsequently, combined univariate Cox regression with a logrank test analysis with p value < 0.05 and 12 variables (miR1293, miR196b, miR490, C20orf197, SCAT1, C11orf44, MALAT1, VPS9D-AS1, LINC02473, LINC01740, HOXB7, and IGF2BP1) were identified. Furthermore, a stepwise regression was performed according to the 12 variables. Consequently, four variables including miR490, miR1293, LINC01740, and IGF2BP1 were harvested in the Cox regression. Risk score = -0.455*miR490 + 0.037*miR1293 + 0.034*LINC01740 + 0.005IGF2BP1 (**Figure 5A**).

Afterward, the LUAD patients were divided into two groups based on the median value of Cox regression model. The distribution of the risk score along with the corresponding survival data and the four protective gene expression demonstrated that the high-risk LUAD patients tended to experience shorter survival time, and low-risk LUAD patients were opposite (**Figure 5B**). Results show that miR490 and LINC01740 in the high-risk group expression level were lower

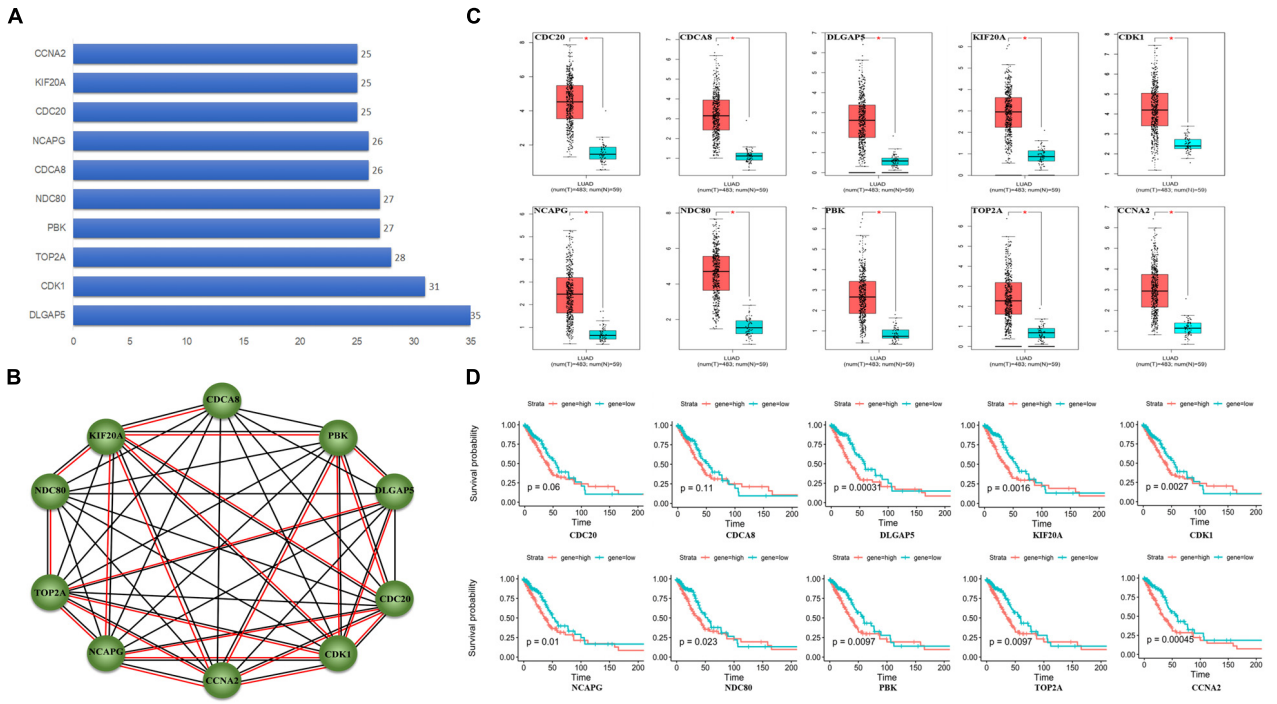


FIGURE 3 | Protein–protein interaction (PPI) network analysis. **(A)** Ten hub genes in PPI based on the DEmRNAs with a combined score of >0.9 and degree ≥25. **(B)** Ten hub genes interaction network. Circles indicate the genes in the PPI network, and the connection indicates the potential interaction between different mRNAs. The red line means physical interaction, and the black line means co-expression with each other. **(C)** Gene expression of 10 hub genes between LUAD tumor and normal tissues. **(D)** Overall survival curves of the 10 hub genes in LUAD. *$p < 0.05$.

than that in the low-risk group; meanwhile, miR1293 and IGF2BP1 were opposite (**Figures 5C,D**).

DISCUSSION

In this study, a total of 2078 DEmRNAs, 257 DElncRNAs, and 101 DEmiRNAs were identified. GO analysis revealed that the function of DERNAs is mainly associated with receptor ligand activity, ligand-gated ion channel transport, morphogenesis of an epithelium process, and cell–cell adhesion, which play vital roles in tumorigenesis (Valley et al., 2015; Inoue et al., 2016). Biological pathway annotation of DERNAs showed that the GPCR ligand binding process accumulated the largest number of dysregulated genes (86 DERNAs), which indicated that the pathway may play an important role in the development and progression of tumors.

In addition, KEGG pathways analysis showed that DERNAs are mainly enriched in neuroactive ligand–receptor interaction, alcoholism, systemic lupus erythematosus, metabolism of xenobiotics by cytochrome P450, and steroid hormone biosynthesis pathways, which are related to the progression of many cancers, including lung cancer (Ashton et al., 2010; Bulk et al., 2017). Among these enriched pathways, the neuroactive ligand–receptor interaction pathway accumulated the most dysregulated genes (53 DERNAs), indicating that they were associated with lung cancer progression. Multiple DERNAs were enriched in both alcoholism pathway and systemic lupus erythematosus pathway. To our surprise, all DERNAs enriched in these two pathways were up-regulated, except for gene 2354, whose gene symbol is "FOSB." Gene annotation of FOSB shows that it is a proto-oncogene, and it has been considered as a regulator of cell proliferation, differentiation, and transformation (Liu et al., 2018; Na and Kim, 2018; Park et al., 2019).

In this study, a total of 55 proteins (including 45 DEmRNAs) were enrolled in the PPI network, and pathway enrichment analysis was performed based on the 10 hub genes. Most of the 10 hub genes were associated to cell cycle-related pathways, including M Phase (Zhang et al., 2018; Chung et al., 2019), Cell Cycle Checkpoints (Lee et al., 2016; Wenzel and Singh, 2018), TP53 Regulates Transcription of Cell Cycle Genes (Ni et al., 2018), and Viral carcinogenesis pathways (Villa, 2013; Shibata et al., 2016), which play an important role in occurrence and development of tumors (Florez et al., 2017; Niculescu, 2019). Cell cycle disorder and cell overgrowth are common biological characteristic of tumors, leading to increased cell proliferation and decreased apoptosis (Hsiao et al., 2014; Kebsa et al., 2018). It should be noted that the cell cycle is a tightly regulated process, which is frequently aberrant in lung cancer (Shcherba et al., 2014). By inhibiting the unrestricted cell division and growth of lung cancer cells, cell cycle-related genes have emerged as new targets for the treatment of lung cancer (Girek et al., 2019).

Among the 10 hub genes, 9 were significantly associated with muscle invasive bladder cancer, including CCNA2, CDC20, CDCA8, DLGAP5, KIF20A, NCAPG, NDC80, PBK, and TOP2A

TABLE 2 | Reactome and KEGG pathway enrichment results.

Term ID	Term description	Observed counts	Background count	FDR
HSA-69278	Cell cycle, mitotic	8	483	8.61E-10
HSA-68886	M phase	6	343	2.74E-07
HSA-68877	Mitotic prometaphase	5	190	7.61E-07
HSA-69620	Cell cycle checkpoints	5	265	3.09E-06
HSA-2500257	Resolution of sister chromatid cohesion	4	118	6.59E-06
HSA-1538133	G0 and early G1	3	27	7.46E-06
HSA-170145	Phosphorylation of proteins involved in the G2/M transition by Cyclin A: Cdc2 complexes	2	3	3.96E-05
HSA-174184	Cdc20: Phospho-APC/C mediated degradation of cyclin A	3	68	8.08E-05
HSA-176408	Regulation of APC/C activators between G1/S and early anaphase	3	76	8.08E-05
HSA-176417	Phosphorylation of Emi1	2	6	8.08E-05
HSA-141444	Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal	3	90	9.66E-05
HSA-2514853	Condensation of prometaphase chromosomes	2	11	0.00013
HSA-1362300	Transcription of E2F targets under negative control by p107 (RBL1) and p130 (RBL2) in complex with HDAC1	2	16	0.00023
HSA-5663220	RHO GTPases activate formins	3	131	0.00023
HSA-174048	APC/C: Cdc20 mediated degradation of cyclin B	2	22	0.00036
HSA-2467813	Separation of sister chromatids	3	178	0.00046
HSA-6804757	Regulation of TP53 degradation	2	35	0.00072
HSA-4615885	SUMOylation of DNA replication proteins	2	40	0.00087
HSA-6791312	TP53 regulates transcription of cell cycle genes	2	49	0.0012
HSA-5688426	Deubiquitination	3	279	0.0013
HSA-597592	Post-translational protein modification	5	1366	0.0013
HSA-69206	G1/S transition	2	128	0.0068
hsa04110 (KEGG)	Cell cycle	3	123	0.00045
hsa05203 (KEGG)	Viral carcinogenesis	3	183	0.00072
hsa04914 (KEGG)	Progesterone-mediated oocyte maturation	2	94	0.0052
hsa04114 (KEGG)	Oocyte meiosis	2	116	0.0059
hsa04218 (KEGG)	Cellular senescence	2	156	0.0084

(Lee et al., 2012). It indicated that there may exist relative risk between muscle invasive bladder cancer and LUAD. Furthermore, we found that all these 10 hub genes were up-regulated in LUAD tumor tissue. The PPI network showed that almost all the 10 hub genes could interact with each other, and DLGAP5, CDK1, and KIF20A play a key role in connecting the network. Among them, DLGAP5 could physically interact with PBK, TOP2A, and CDK1, and all mitosis-associated proteins correlated with poor prognosis for non-small cell lung cancer patients (Shih et al., 2012; Schneider et al., 2017). In addition, a previous study reported that CCNA2, CDC20, PBK, and TOP2A that interacted with CDK1 play vital roles in survival outcomes in human lung cancer. Loss of cytoplasmic CDK1 could predict poor survival in human lung cancer and confers chemotherapeutic resistance (Zhang et al., 2011). Hence, we concluded that these 10 hub genes play key roles in regulating cell proliferation in LUAD.

A total of three DEmiRNAs, seven DElncRNAs, and six DEmRNAs were enrolled in the ceRNA network. In ceRNA network, we found that MALAT1 as a highly conserved lncRNA whose overexpression has been shown in various cancers, such as breast, prostate, colon, and liver, especially in early stage metastasizing patients (Lin et al., 2007; Guffanti et al., 2009; Xu et al., 2011; Ren et al., 2013). In addition, Ping et al. have reported that MALAT1 can predict metastasis in early stage NSCLC (Ji et al., 2003). Consistent with Ping et al., Lars et al. verified that MALAT1 stimulates migration, invasion, and tumor growth (Schmidt et al., 2011), although the underlying mechanism is poorly understood. In our ceRNA network, the expression of miR490 is down-regulated while MALAT1 and HMGA2 expression is up-regulated in LUAD. One possible explanation is that aberrant expression of MALAT1 acts as a ceRNA for miR-490, and high-expression MALAT1 inhibits miR490 and then increased expression of HMGA2 (the target of miR490), finally accelerating to tumor progression.

Many homeobox genes, including HOXC8, HOXB7, and HOXA9, are also "members" of the ceRNA network. A previous study reported that mis-expression of homeobox genes can lead to abnormal differentiation and proliferation, leading to a change in cell identity or homeotic transformation, therefore playing an important role in carcinogenesis (Samuel and Naora, 2005). In cancer, homeobox genes function as "tumor

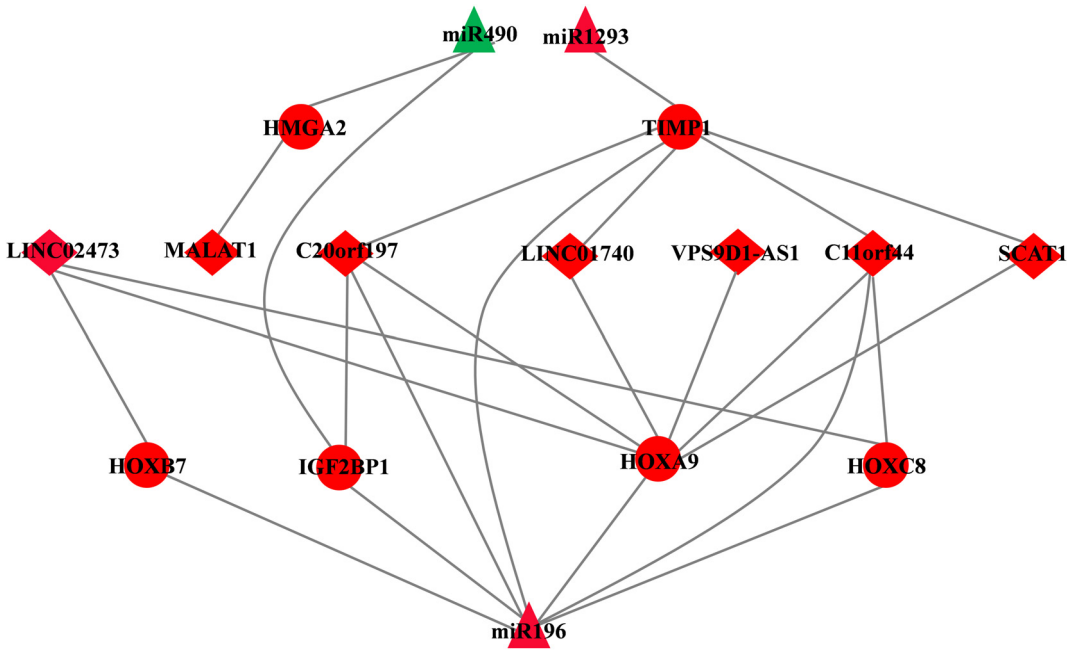


FIGURE 4 | CeRNA network of LUAD. The triangles indicate miRNAs, circles mean mRNAs, and diamonds represent lncRNAs. Red means up-regulated, and green means down-regulated.

TABLE 3 | The miRTarBase database revealed interactions between miRNA and mRNAs.

miRNA	mRNA
miR1293	TIMP1
miR196b	HOXB8, HOXC8, CD8A, HOXA9, MEIS1, FAS, ETS2, RDX, HOXB7, GATA6, TGFBR2, PIK3CG, AKT1, MTOR, FOS, IGF2BP1,
miR490	ERGIC3, FOS, SMARCD1, CCND1, PIK3CA, PAPPA, ABCC2, TGFBR1, HMGA2, TGFA, RHOA, PCBP1, TNKS2, BMPR2, HNRNPA1

modulators" as their deregulation normally involve either up-regulation of genes expressed in undifferentiated cells or down-regulation of genes expressed in differentiated tissue, thus acting either as oncogenes or tumor suppressor genes (Abate-Shen, 2002). Almost all the genes in the ceRNA network have reported that they enrolled or associated with tumor progression, except for LINC02473, LINC0170, VPS9D1-AS1, C11orf44, and SCAT1. Hence, taking all these genes in the ceRNA network into consideration, we combined step regression and Cox regression analysis and identified four genes as prognostic biomarkers in LUAD, including miR490, miR1293, LINC01740, and IGF2BP1.

By searching these genes in PubMed, we found that miR490 and IGF2BP1 have been studied for their mechanism in or association with tumor progression. Gain- and loss-of-function studies of miR490 showed that it regulates cell proliferation and is required for induction of *in vitro* migration and invasion (Zhao and Zheng, 2016). miR490 overexpression reduced proliferation, promoted G1 arrest and apoptosis, and suppressed migration and invasion (Sun et al., 2016). In our study, miR490 expression was significantly lower in lung cancer than in normal tissues, and survival analysis result showed that the lower expression miR490 predicted poor survival in lung cancer. Opposite to miR490, IGF2BP1 expression is up-regulated, and the high expression level of IGF2BP1 showed poor overall survival outcomes in lung cancer. Studies reported that IGF2BP1 has been traditionally regarded as an oncogene and potential therapeutic target for cancers (Huang et al., 2018). It plays essential roles in embryogenesis and carcinogenesis, regulating the expression of some essential mRNA targets required for the control of tumor cell proliferation, growth, and invasion, and associating with a poor overall survival and metastasis in various types of human cancers (Gong et al., 2016). However, there is no public report on miR1293 and LINC01740 according to a PubMed search. Univariate Cox regression analysis showed that high expression of miR1293 tended to show poor survival outcomes (logrank_*p*value < 0.0001), and high expression of LINC01740 tended to show good survival outcomes (logrank_value = 0.048). Our results suggest that the four predictive genes may play crucial roles in the pathomechanism of LUAD and act as potential prognostic biomarkers.

Although a four-predictive gene signature was constructed and appears to be potential prognostic biomarkers in clinical application, there are some limitations. First, the prognostic value of LINC01740 is not very satisfactory. Second, the binding affinities between lncRNA and miRNA were predicted by simple linear regression model and miRcode and should be further experimentally investigated. Third, the function and mechanism

TABLE 4 | Information about differentially expressed RNAs and miRNAs in ceRNA network.

DERNAs/DEmiRNAs	Log2FC	*p* value	FDR	logrank_*p*value
miR1293	4.00	2.2e-13	1.20e-12	9.48e-05
miR196b	3.94	1.91e-31	2.80e-30	3.45e-02
miR490	−2.31	7.58e-07	2.63e-06	3.53e-02
C20orf197	3.06	9.22e-19	7.80e-18	0.003
SCAT1	2.87	3.25e-23	3.89e-22	0.006
C11orf44	2.32	9.35e-07	2.45e-06	0.019
MALAT1	2.09	5.96e-10	2.22e-09	0.023
VPS9D1-AS1	2.86	3.57e-38	9.82e-37	0.029
LINC02473	4.18	5.18e-25	6.98e-24	0.045
LINC01740	3.32	6.55e-07	1.74e-06	0.049
HMGA2	5.77	1.22e-22	1.39e-21	0.152
TIMP1	1.53	5.33e-25	7.16e-24	0.961
HOXB7	2.09	2.99e-17	2.24e-16	0.006
IGF2BP1	6.41	4.26e-21	4.34e-20	0.042
HOXA9	1.77	1.51e-07	4.33e-07	0.371
HOXC8	2.67	3.23e-13	1.66e-12	0.315

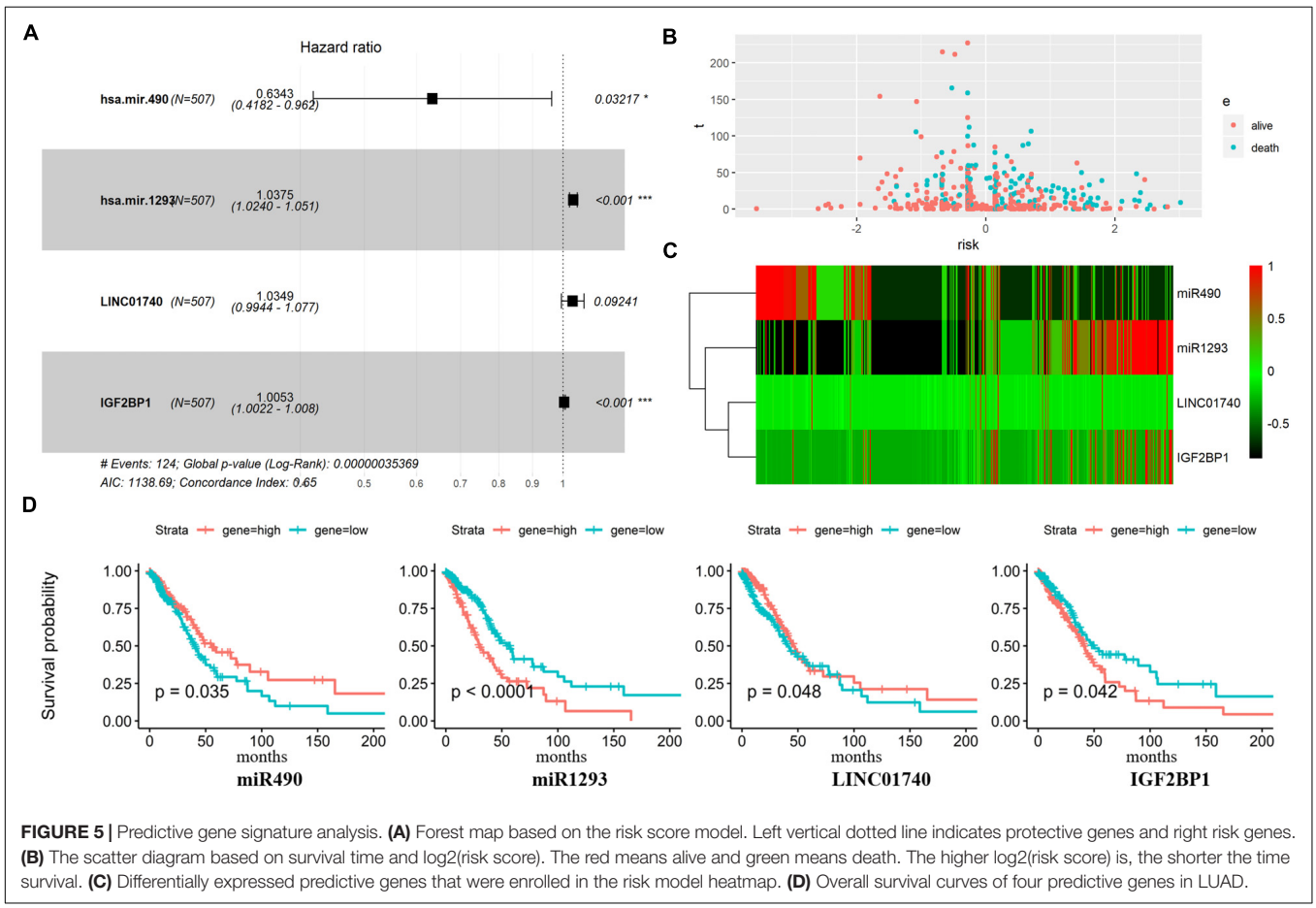


FIGURE 5 | Predictive gene signature analysis. **(A)** Forest map based on the risk score model. Left vertical dotted line indicates protective genes and right risk genes. **(B)** The scatter diagram based on survival time and log2(risk score). The red means alive and green means death. The higher log2(risk score) is, the shorter the time survival. **(C)** Differentially expressed predictive genes that were enrolled in the risk model heatmap. **(D)** Overall survival curves of four predictive genes in LUAD.

of the four predictive genes in LUAD need to be further studied by experiments.

In conclusion, we established the disordered ceRNA network, which is beneficial to understanding the relationship among lncRNA–miRNA–mRNA and provides efficient strategies for subsequent functional studies of them. In addition, we identified that miR1293, miR490, LINC01740, and IGF2BP1 might be novel important prognostic factors involved in LUAD pathogenesis, and the risk score model is helpful in studying the overall survival outcome in LUAD.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this study are publicly available from The Cancer Genome Atlas (TCGA) database, and can be accessed here: https://portal.gdc.cancer.gov/).

AUTHOR CONTRIBUTIONS

YW developed the theory and performed the computations. RH verified the analytical methods and results. LM conceived the original idea and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was supported by the China Postdoctoral Research Fund (2018, 080-090436) and Key Technology Research and Development of Chiral Drug Intermediate Biocatalysts, a Major Technological Innovation Project in Hubei Province (2017ACA174).

REFERENCES

Abate-Shen, C. (2002). Deregulated homeobox gene expression in cancer: cause or consequence? *Nat. Rev. Cancer* 2, 777–785. doi: 10.1038/nrc907

Ashton, K. A., Proietto, A., Otton, G., Symonds, I., McEvoy, M., Attia, J., et al. (2010). Polymorphisms in genes of the steroid hormone biosynthesis and metabolism pathways and endometrial cancer risk. *Cancer Epidemiol.* 34, 328–337. doi: 10.1016/j.canep.2010.03.005

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief Bioinform.* doi: 10.1093/bib/bbz118 [Epub ahead of print]

Bulk, E., Kramko, N., Liashkovich, I., Glaser, F., Schillers, H., Schnittler, H. J., et al. (2017). KCa3.1 channel inhibition leads to an ICAM-1 dependent increase of cell-cell adhesion between A549 lung cancer and HMEC-1 endothelial cells. *Oncotarget* 8, 112268–112282. doi: 10.18632/oncotarget.22735

Cagle, P. T., Allen, T. C., and Olsen, R. J. (2013). Lung cancer biomarkers: present status and future developments. *Arch. Pathol. Lab. Med.* 137, 1191–1198. doi: 10.5858/arpa.2013-0319-CR

Castro-Oropeza, R., Melendez-Zajgla, J., Maldonado, V., and Vazquez-Santillan, K. (2018). The emerging role of lncRNAs in the regulation of cancer stem cells. *Cell Oncol.* 41, 585–603. doi: 10.1007/s13402-018-0406-4

Chung, K. S., Hong, J. Y., Lee, J. H., Lee, H. J., Park, J. Y., Choi, J. H., et al. (2019). beta-Caryophyllene in the essential oil from chrysanthemum boreale induces G1 phase cell cycle arrest in human lung cancer cells. *Molecules* 24:E3754. doi: 10.3390/molecules24203754

Ezeife, D. A., and Leighl, N. B. (2018). Personalized medicine for non-small cell lung cancer: where are we now and where can we go? *Expert. Rev. Respir. Med.* 12, 81–82. doi: 10.1080/17476348.2018.1411805

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606

Florez, M. M., Feo, H. B., da Silva, G. N., Yamatogi, R. S., Aguiar, A. J., Araujo, J. P., et al. (2017). Cell cycle kinetics, apoptosis rates and gene expressions of MDR-1, TP53, BCL-2 and BAX in transmissible venereal tumour cells and their association with therapy response. *Vet. Comp. Oncol.* 15, 793–807. doi: 10.1111/vco.12220

Girek, M., Klosinski, K., Grobelski, B., Pizzimenti, S., Cucci, M. A., Daga, M., et al. (2019). Novel tetrahydroacridine derivatives with iodobenzoic moieties induce G0/G1 cell cycle arrest and apoptosis in A549 non-small lung cancer and HT-29 colorectal cancer cells. *Mol. Cell. Biochem.* 460, 123–150. doi: 10.1007/s11010-019-03576-x

Gong, F., Ren, P., Zhang, Y., Jiang, J., and Zhang, H. (2016). MicroRNAs-491-5p suppresses cell proliferation and invasion by inhibiting IGF2BP1 in non-small cell lung cancer. *Am. J. Transl. Res.* 8, 485–495.

Guffanti, A., Iacono, M., Pelucchi, P., Kim, N., Solda, G., Croft, L. J., et al. (2009). A transcriptional sk+etch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics* 10:163. doi: 10.1186/1471-2164-10-163

Gutschner, T., Hammerle, M., and Diederichs, S. (2013). MALAT1 – a paradigm for long noncoding RNA function in cancer. *J. Mol. Med.* 91, 791–801. doi: 10.1007/s00109-013-1028-y

Hsiao, C. J., Hsiao, G., Chen, W. L., Wang, S. W., Chiang, C. P., Liu, L. Y., et al. (2014). Cephalochromin induces G0/G1 cell cycle arrest and apoptosis in A549 human non-small-cell lung cancer cells by inflicting mitochondrial disruption. *J. Nat. Prod.* 77, 758–765. doi: 10.1021/np400517g

Huang, X., Zhang, H., Guo, X., Zhu, Z., Cai, H., and Kong, X. (2018). Insulin-like growth factor 2 mRNA-binding protein 1 (IGF2BP1) in cancer. *J. Hematol. Oncol.* 11:88. doi: 10.1186/s13045-018-0628-y

Inoue, T., Adachi, K., Kawana, K., Taguchi, A., Nagamatsu, T., Fujimoto, A., et al. (2016). Cancer-associated fibroblast suppresses killing activity of natural killer cells through downregulation of poliovirus receptor (PVR/CD155), a ligand of activating NK receptor. *Int. J. Oncol.* 49, 1297–1304. doi: 10.3892/ijo.2016.3631

Ito, K., and Murphy, D. (2013). Application of ggplot2 to Pharmacometric Graphics. *CPT Pharmacometr. Syst. Pharmacol.* 2:e79. doi: 10.1038/psp.2013.56

Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P. M., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041. doi: 10.1038/sj.onc.1206928

Kang, Y., Jia, Y., Wang, Q., Zhao, Q., Song, M., Ni, R., et al. (2019). Long Noncoding RNA KCNQ1OT1 promotes the progression of non-small cell lung cancer via regulating miR-204-5p/ATG3 Axis. *Oncol. Targets Ther.* 12, 10787–10797. doi: 10.2147/OTT.S226044

Kebsa, W., Lahouel, M., Rouibah, H., Zihlif, M., Ahram, M., Abu-Irmaileh, B., et al. (2018). Reversing multidrug resistance in chemo-resistant human lung Adenocarcinoma (A549/DOX) cells by Algerian Propolis through direct inhibiting the P-gp Efflux-pump, G0/G1 cell cycle arrest and apoptosis induction. *Anticancer Agents Med. Chem.* 18, 1330–1337. doi: 10.2174/1871520618666180808100800

Kerr, K. M., Bubendorf, L., Edelman, M. J., Marchetti, A., Mok, T., Novello, S., et al. (2014). Second ESMO consensus conference on lung cancer: pathology and molecular biomarkers for non-small-cell lung cancer. *Ann. Oncol.* 25, 1681–1690. doi: 10.1093/annonc/mdu145

Lee, S. J., Lee, E. J., Kim, S. K., Jeong, P., Cho, Y. H., Yun, S. J., et al. (2012). Identification of pro-inflammatory cytokines associated with muscle invasive bladder cancer; the roles of IL-5, IL-20, and IL-28A. *PLoS One* 7:e40267. doi: 10.1371/journal.pone.0040267

Lee, Y. S., Doonan, B. B., Wu, J. M., and Hsieh, T. C. (2016). Combined metformin and resveratrol confers protection against UVC-induced DNA damage in A549 lung cancer cells via modulation of cell cycle checkpoints and DNA repair. *Oncol. Rep.* 35, 3735–3741. doi: 10.3892/or.2016.4740

Lin, R., Maeda, S., Liu, C., Karin, M., and Edgington, T. S. (2007). A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* 26, 851–858. doi: 10.1038/sj.onc.1209846

Liu, S., Luan, J., and Ding, Y. (2018). miR-144-3p Targets FosB Proto-oncogene, AP-1 transcription factor subunit (FOSB) to suppress proliferation, migration, and invasion of PANC-1 pancreatic cancer cells. *Oncol. Res.* 26, 683–690. doi: 10.3727/096504017X14982585511252

Liu, Z., Wang, Y., Wang, L., Yao, B., Sun, L., Liu, R., et al. (2019). Long non-coding RNA AGAP2-AS1, functioning as a competitive endogenous RNA, upregulates ANXA11 expression by sponging miR-16-5p and promotes proliferation and metastasis in hepatocellular carcinoma. *J. Exp. Clin. Cancer Res.* 38:194. doi: 10.1186/s13046-019-1188-x

Liz, J., and Esteller, M. (2016). lncRNAs and microRNAs with a role in cancer development. *Biochim. Biophys. Acta* 1859, 169–176. doi: 10.1016/j.bbagrm.2015.06.015

Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29, 1830–1831. doi: 10.1093/bioinformatics/btt285

Majeed, H., and Amir, E. (2018). Personalized medicine in immuno-oncology: a novel prognostic index in non-small cell lung cancer. *J. Thorac. Dis.* 10(Suppl. 9), S995–S998. doi: 10.21037/jtd.2018.03.153

Na, H. H., and Kim, K. C. (2018). SETDB1-mediated FosB regulation via ERK2 is associated with an increase in cell invasiveness during anticancer drug treatment of A549 human lung cancer cells. *Biochem. Biophys. Res. Commun.* 495, 512–518. doi: 10.1016/j.bbrc.2017.10.176

Ni, Z., Wang, X., Zhang, T., Li, L., and Li, J. (2018). Comprehensive analysis of differential expression profiles reveals potential biomarkers associated with the cell cycle and regulated by p53 in human small cell lung cancer. *Exp. Ther. Med.* 15, 3273–3282. doi: 10.3892/etm.2018.5833

Niculescu, V. F. (2019). The reproductive life cycle of cancer: hypotheses of cell of origin, TP53 drivers and stem cell conversions in the light of the atavistic cancer cell theory. *Med. Hypotheses* 123, 19–23. doi: 10.1016/j.mehy.2018.12.006

Park, J. A., Na, H. H., Jin, H. O., and Kim, K. C. (2019). Increased expression of FosB through reactive oxygen species accumulation functions as pro-apoptotic protein in Piperlongumine treated MCF7 breast cancer cells. *Mol. Cells* 42, 884–892. doi: 10.14348/molcells.2019.0088

Ren, S., Liu, Y., Xu, W., Sun, Y., Lu, J., Wang, F., et al. (2013). Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. *J. Urol.* 190, 2278–2287. doi: 10.1016/j.juro.2013.07.001

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, 139–140. doi: 10.1093/bioinformatics/btp616

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, 353–358. doi: 10.1016/j.cell.2011.07.014

Samuel, S., and Naora, H. (2005). Homeobox gene expression in cancer: insights from developmental regulation and deregulation. *Eur. J. Cancer* 41, 2428–2437. doi: 10.1016/j.ejca.2005.08.014

Schmidt, L. H., Spieker, T., Koschmieder, S., Schaffers, S., Humberg, J., Jungen, D., et al. (2011). The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. *J. Thorac. Oncol.* 6, 1984–1992. doi: 10.1097/JTO.0b013e3182307eac

Schneider, M. A., Christopoulos, P., Muley, T., Warth, A., Klingmueller, U., Thomas, M., et al. (2017). AURKA, DLGAP5, TPX2, KIF11 and CKAP5: five specific mitosis-associated genes correlate with poor prognosis for non-small cell lung cancer patients. *Int. J. Oncol.* 50, 365–372. doi: 10.3892/ijo.2017.3834

Shcherba, M., Liang, Y., Fernandes, D., Perez-Soler, R., and Cheng, H. (2014). Cell cycle inhibitors for the treatment of NSCLC. *Expert Opin. Pharmacother.* 15, 991–1004. doi: 10.1517/14656566.2014.902935

Shi, X., Tan, H., Le, X., Xian, H., Li, X., Huang, K., et al. (2018). An expression signature model to predict lung adenocarcinoma-specific survival. *Cancer Manag. Res.* 10, 3717–3732. doi: 10.2147/CMAR.S159563

Shibata, M., Minami, S., Oe, S., Miyagawa, K., Honma, Y., Onitsuka, R., et al. (2016). Recognition of infection, status of outpatient treatment, and treatment history before carcinogenesis in patients with viral hepatitis-associated liver cancer. *Nihon Shokakibyo Gakkai Zasshi* 113, 1734–1742. doi: 10.11405/nisshoshi.113.1734

Shih, M. C., Chen, J. Y., Wu, Y. C., Jan, Y. H., Yang, B. M., Lu, P. J., et al. (2012). TOPK/PBK promotes cell migration via modulation of the PI3K/PTEN/AKT pathway and is associated with poor prognosis in lung cancer. *Oncogene* 31, 2389–2400. doi: 10.1038/onc.2011.419

Sun, J., Zhao, H., Lin, S., Bao, S., Zhang, Y., Su, J., et al. (2019). Integrative analysis from multi-centre studies identifies a function-derived personalized multi-gene signature of outcome in colorectal cancer. *J. Cell Mol. Med.* 23, 5270–5281. doi: 10.1111/jcmm.14403

Sun, K. X., Chen, Y., Chen, S., Liu, B. L., Feng, M. X., Zong, Z. H., et al. (2016). The correlation between microRNA490-3p and TGFalpha in endometrial carcinoma tumorigenesis and progression. *Oncotarget* 7, 9236–9249. doi: 10.18632/oncotarget.7061

Valley, C. C., Arndt-Jovin, D. J., Karedla, N., Steinkamp, M. P., Chizhik, A. I., Hlavacek, W. S., et al. (2015). Enhanced dimerization drives ligand-independent activity of mutant epidermal growth factor receptor in lung cancer. *Mol. Biol. Cell* 26, 4087–4099. doi: 10.1091/mbc.E15-05-0269

Villa, L. L. (2013). Viral carcinogenesis: virus implicated in cancer. *BMC Proc.* 7(Suppl. 2):K11. doi: 10.1186/1753-6561-7-S2-K11

Wenzel, E. S., and Singh, A. T. K. (2018). Cell-cycle checkpoints and aneuploidy on the path to cancer. *In Vivo* 32, 1–5. doi: 10.21873/invivo.11197

Xu, C., Yang, M., Tian, J., Wang, X., and Li, Z. (2011). MALAT-1: a long non-coding RNA and its important 3' end functional motif in colorectal cancer metastasis. *Int. J. Oncol.* 39, 169–175. doi: 10.3892/ijo.2011.1007

Yan, X., Hu, Z., Feng, Y., Hu, X., Yuan, J., Zhao, S. D., et al. (2015). Comprehensive genomic characterization of long non-coding RNAs across human cancers. *Cancer Cell* 28, 529–540. doi: 10.1016/j.ccell.2015.09.006

Yu, F., Quan, F., Xu, J., Zhang, Y., Xie, Y., Zhang, J., et al. (2018). Breast cancer prognosis signature: linking risk stratification to disease subtypes. *Brief Bioinform.* 20, 2130–2140. doi: 10.1093/bib/bby073

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

Zeng, F., Wang, Q., Wang, S., Liang, S., Huang, W., Guo, Y., et al. (2019). Linc00173 promotes chemoresistance and progression of small cell lung cancer by sponging miR-218 to regulate Etk expression. *Oncogene* 39, 293–307. doi: 10.1038/s41388-019-0984-2

Zhang, C., Elkahloun, A. G., Robertson, M., Gills, J. J., Tsurutani, J., Shih, J. H., et al. (2011). Loss of cytoplasmic CDK1 predicts poor survival in human lung cancer and confers chemotherapeutic resistance. *PLoS One* 6:e23849. doi: 10.1371/journal.pone.0023849

Zhang, L. L., Feng, Z. L., Su, M. X., Jiang, X. M., Chen, X., Wang, Y., et al. (2018). Downregulation of Cyclin B1 mediates nagilactone E-induced G2 phase cell cycle arrest in non-small cell lung cancer cells. *Eur. J. Pharmacol* 830, 17–25. doi: 10.1016/j.ejphar.2018.04.020

Zhao, L., and Zheng, X. Y. (2016). MicroRNA-490 inhibits tumorigenesis and progression in breast cancer. *Oncol. Targets Ther.* 9, 4505–4516. doi: 10.2147/OTT.S100037

Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018a). Recurrence-Associated Long Non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. *Mol. Ther. Nucleic Acids* 12, 518–529. doi: 10.1016/j.omtn.2018.06.007

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018b). An Immune-related Six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. *Mol. Neurobiol.* 55, 3684–3697. doi: 10.1007/s12035-017-0572-9

Zhou, M., Zhang, Z., Zhao, H., Bao, S., and Sun, J. (2018c). A novel lncRNA-focus expression signature for survival prediction in endometrial carcinoma. *BMC Cancer* 18:39. doi: 10.1186/s12885-017-3983-0

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

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 © 2020 Wang, He 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/fbioe.2020.00274

Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications

*Rao Zeng and Minghong Liao**

Department of Software Engineering, School of Informatics, Xiamen University, Xiamen, China

DNA N4-methylcytosine modification (4mC) plays an essential role in a variety of biological processes. Therefore, accurate identification the 4mC distribution in genome-scale is important for systematically understanding its biological functions. In this study, we present Deep4mcPred, a multi-layer deep learning based predictive model to identify DNA N4-methylcytosine modifications. In this predictor, we for the first time integrate residual network and recurrent neural network to build a multi-layer deep learning predictive system. As compared to existing predictors using traditional machine learning, our proposed method has two advantages. First, our deep learning framework does not need to specify the features when training the predictive model. It can automatically learn the high-level features and capture the characteristic specificity of 4mC sites, benefiting to distinguish true 4mC sites from non-4mC sites. On the other hand, our deep learning method outperforms the traditional machine learning predictors in performance by benchmarking comparison, demonstrating that the proposed Deep4mcPred is more effective in the DNA 4mC site prediction. Moreover, via experimental comparison, we found that attention mechanism introduced into the deep learning framework is useful to capture the critical features. Additionally, we develop a webserver implementing the proposed method for the academic use of research community, which is now available at http://server.malab.cn/Deep4mcPred.

Keywords: DNA N4-methylcytosine, deep learning, site prediction, webserver, feature representation

OPEN ACCESS

Edited by:
Yongchun Zuo,
Inner Mongolia University, China

Reviewed by:
Balachandran Manavalan,
Ajou University, South Korea
Bin Liu,
Beijing Institute of Technology, China

****Correspondence:***
Minghong Liao
liao@xmu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *15 January 2020*
Accepted: *16 March 2020*
Published: *21 April 2020*

Citation:
Zeng R and Liao M (2020) Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications. Front. Bioeng. Biotechnol. 8:274. doi: 10.3389/fbioe.2020.00274

INTRODUCTION

Epigenetics refers to the heritable phenotype changes in the function of genes that do not involve alterations in DNA sequence. DNA methylation refers to the binding of a methyl group on the nucleotide of DNA (Liu et al., 2019a) under the action of DNA methyltransferases (Dnmt). As one of the earliest discovered and most in-depth epigenetic regulation mechanisms, it is associated with normal development and plays an essential role in key biological processes including regulating gene expression, regulating mammalian growth and development, mediating X chromosome inactivation, and participating in gene imprinting (Jin et al., 2011). It can be divided into three categories according to the position of methylation modification: N6-methyladenine (6mA), 5-Methylcytosine (5mC) and N4-methylcytosine (4mC) (Chen et al., 2017; Wei et al., 2019a). The most prevalent methylation modification in eukaryotes is 5mC (Luo et al., 2015; Xiao et al., 2018)

which consists of methylation at the fifth position of the cytosine pyrimidine ring and has focused on epigenetic markers in mammals and plants (Liu et al., 2019a), while 6mA (methylations on the sixth position of the adenine purine ring) (Liu et al., 2019a) is the most predominant DNA modification in prokaryote and has been found to be related to the regulation of restriction-modification (R-M) system, DNA mismatch repair, gene expression, and other aspects (Luo et al., 2015; Xiao et al., 2018). With the development of high-throughput techniques, the 4mC (methylations on the fourth position of the cytosine pyrimidine ring) was discovered in bacteria, and found to play an important role in protecting genome from invasion in restriction-modification (R-M) system. Developing methods to explore more biological functions of 4mC is of significance.

Single-molecule real time sequencing (SMRT) technology has been proposed to detect the 4mC and 6mA sites from the whole genome (Flusberg et al., 2010). However, using SMRT techniques to analyze the genome is costly inefficient. Therefore, Yu et al. (2015) proposed 4mC-Tet-assisted bisulfite-sequencing (4mC-TABseq) as a new generation of sequencing technology (Illumina sequencing systems) to identify the genome-wide locations of 4mC for bacterial species more rapidly and cost efficiently. Although the prediction of 4mC sites by this sequencing technique has been improved to some extent, recent studies focus more on the recognition of 4mC sites using machine learning, which is capable of predicting 4mC sites based on genome sequences, without any prior experimental knowledge. There are currently four methods available in literature to identify 4mC sites, including iDNA4mC (Chen et al., 2017), 4mCPred (Su et al., 2018), 4mcPred-SVM (Wei et al., 2018a), and 4mcPred-IFL (Wei et al., 2019a). iDNA4mC, as the first machine learning predictor, encodes sequences by nucleotide chemical properties and nucleotide frequency to features and trains support vector machine (SVM) models for prediction (Liang et al., 2018). Although this method has the ability to distinguish between 4mC and non-4mC sites, the prediction accuracy is relatively low overall. Afterwards, He et al. proposed 4mCPred, an SVM-based predictive model trained with position-specific trinucleotide propensity (PSTNP) and electron-ion interaction potential features. More recently, 4mcPred-SVM and 4mcPred-IFL, proposed by Wei et al., further improve the predictive performance on the same golden benchmark datasets. The former employs a two-step feature optimization strategy to improve the feature representation ability, while the latter uses an iterative feature representation algorithm to learn critical information from several sequential feature models. Even though the above methods have improved the performance for identifying 4mC sites, too few data sets have been adopted to fully reflect the whole genome and to build robust models. Consequently, it is eager and indispensable to develop a robust and strong model to more accurately identify 4mC sites.

In recent years, deep learning is not only developed as a new research direction in machine learning, but also has made a lot of achievements in data mining (Lan et al., 2018), speech recognition (Amodei et al., 2014), machine translation (Sutskever et al., 2014), natural language processing (Collobert and Weston, 2008; Young et al., 2018), and other related fields (Hong et al., 2019; Li and Liu, 2019; Liu et al., 2019b; Yang et al., 2019; Zeng et al., 2019a,b). In the field of computational biology, deep learning has been widely applied, especially in solving the problems of genome sequence-based by convolutional neural networks (CNN) (Nie et al., 2018; Peng et al., 2018; Lv et al., 2019a; Wang et al., 2019; Zhang et al., 2019a; Zou et al., 2019). In this paper, we proposed Deep4mcPred, a multi-layer deep learning based predictive model to identify DNA N4-methylcytosine modifications. In this predictor, we for the first time integrate residual network (He et al., 2016) and recurrent neural network, together with attention mechanism, to build a multi-layer deep learning predictive system. We evaluated and compared our predictor with existing predictors. The comparative results demonstrate that our proposed model can more accurately identify 4mC sites than the state-of-the-art predictors. In addition, the proposed method is implemented by the simple and easy-to-use webserver which is freely available on http://server.malab.cn/Deep4mcPred.

METHODS AND MATERIALS

Dataset Collection

Previous study has demonstrated that a stringent dataset is essential for building a robust predictive model (Zeng et al., 2016, 2017a; Liu et al., 2017; Wei et al., 2017a, 2018b,c; Jin et al., 2019; Liu, 2019; Su et al., 2019). In existing studies, there is one golden benchmark dataset proposed by Chen et al. for performance evaluation and comparison. However, the size of the dataset is too small to train a deep learning model. Accordingly, we constructed a larger dataset in this study. We strictly followed the data processing procedure as introduced in Chen's study. By doing so, we can guarantee our dataset the most representative.

Positive Samples Collection

Specifically, there are three main steps for collecting the positive samples. Firstly, we collected all 41bp long sequences centered with true 4mC sites from the MethSMRT database (Ye et al., 2016). Next, we removed the sequences with Modification QV (modQV) score not <30 as it is the default threshold for invoking the modification location according to the Methylome Analysis Technical Note. Next, we used CD-HIT software (with the threshold of 80%) (Fu et al., 2012) to reduce the identity of the positives, avoiding the potential of performance biased-estimation. Ultimately, following the procedure, we collected the positive samples from three species: *Arabidopsis thaliana* (*A. thaliana*), *Caenorhabditis elegans* (*C. elegans*), and *Drosophila melanogaster* (*D. melanogaster*). The details of the positive samples in the three species are presented in **Table 1**. Note that we randomly picked 20,000 positive samples for model training.

Negative Samples Collection

The negative samples were also cytosine-centered sequences with a length of 41bp but are not recognized by the SMRT sequencing technology. In this case, the number of negative samples per species are much larger than the corresponding positive samples. To avoid the data imbalance problem, we randomly selected

TABLE 1 | Summary of benchmark datasets in three species.

Species	Positives	Negatives	Total
A. thaliana	20,000	20,000	40,000
C. elegans	20,000	20,000	40,000
D. melanogaster	20,000	20,000	40,000

the same number of negative samples with that of the positive samples in corresponding species for model training.

The Framework of the Proposed Deep Learning Method

Figure 1 illustrates the overall predictive framework of the proposed multi-layer deep learning network. For given DNA sequences, neural network is composed of four layers: the input layer, the ResNet layer, the LSTM layer and the attention layer, as seen in **Figure 1**. The first layer is the input layer. The sequences of the dataset are encoded by one-hot method and the obtained features are fed into the subsequent ResNet layer. Through this residual network model, deeper networks can be built than plain CNN models for extracting effective global features. The output feature vectors are utilized as inputs of the LSTM layer. In the LSTM layer, the bidirectional LSTM model is utilized to gather feature information from two directions which has been proven to be more effective than the unidirectional LSTM model. In the last attention layer, the attention mechanism is introduced to integrate the output of the LSTM layer for more relevant feature information. Finally, a fully-connected neural network (FC) is attached after the attention model and the softmax activation function is performed to make predictions.

Sequence Representation Using One-Hot Encoding

Genomic sequences are consisting of four nucleotides: "A" (adenine), "G" (guanine), "C" (cytosine), and "T" (thymine). Undetermined bases are annotated as "N." The nucleotides are represented using one-hot encoding over four bits. For example, "A" is represented as the binary vector (1,0,0,0); "G" is encoded as (0,1,0,0); "C" is encoded as (0,0,1,0); "T" is encoded as (0,0,0,1); and "N" is (0,0,0,0).

Deep Learning Model Architecture

We developed a novel prediction method, namely Deep4mcPred, that integrates Long Short Term Memory (LSTM) recurrent neural network and the attention mechanism into the Residual Networks (ResNet). The overall architecture of our proposed model is shown in **Figure 1**.

Residual Networks (ResNet)

Studies have showed that the overall performance of the network is greatly affected by the number of network layers when it comes to convolutional neural network (CNN). To be specific, the accuracy of the network increases as the depth increases, but when the depth reaches a certain level, the accuracy begins to drop rapidly. This is called the degradation problem, making it difficult to generate very deep neural networks.

To address this, ResNet introduces a residual learning framework to improve the degradation, which has achieved great success in the areas of image classification and item identification in recent studies. The internal residual blocks of ResNet utilize jump connections, alleviating the problem of gradient disappearance caused by the increase of depth in convolutional neural networks.

For an input x, ResNet learns a specific residual function $F(x) = H(x) - x$, whereas $F(x) = H(x)$ for plain CNN. Supposing the residual $F(x) = 0$, then it occurs identity mapping "shortcut." The residual block is performed as follows:

$$y = F(X, \{W_i\}) + x$$

where

$$F = W_2\sigma(x, W_i)$$

where the function F denotes the learned residual mapping and σ represents relu. F and x are added element by element under the premise of shortcut connections.

But in fact, the residual $F(x)$ will not be zero, so the dimensions of F and x will be different. The output of the ResNet layer can be formulated as follows:

$$y = F(X, \{W_i\}) + W_s x$$

where W_s is introduced to perform a linear mapping to match the dimensions. Taking consider of ResNet, it allows the stacked layer to extract more distinct features of the input x, resulting in better performance.

Long Short Term Memory (LSTM)

Recurrent Neural Network (RNN) is a powerful neural network for processing sequential data. The parameter learning of the RNN is performed by the back-propagation algorithm over time. When the input sequence is long, a gradient disappearance or gradient explosion problem occurs, which is termed as long-term dependency problem.

LSTM is one type of RNN, which introduces the conception of self-loop to generate a path of continuous gradient flow for a long time and gating mechanism to control the information flow, solving the long-term dependency problem. It was firstly proposed by Hochreiter and Schmidhuber in 1997. From then on, LSTM has achieved considerable success and has been widely used in the fields of handwriting recognition, machine translation, and speech recognition, etc.

The stacked architecture of LSTM is shown in **Figure 1**. The output from the ResNet layers is fed into the subsequent LSTM layer as the input. Then, the LSTM components are updated by the following formulations:

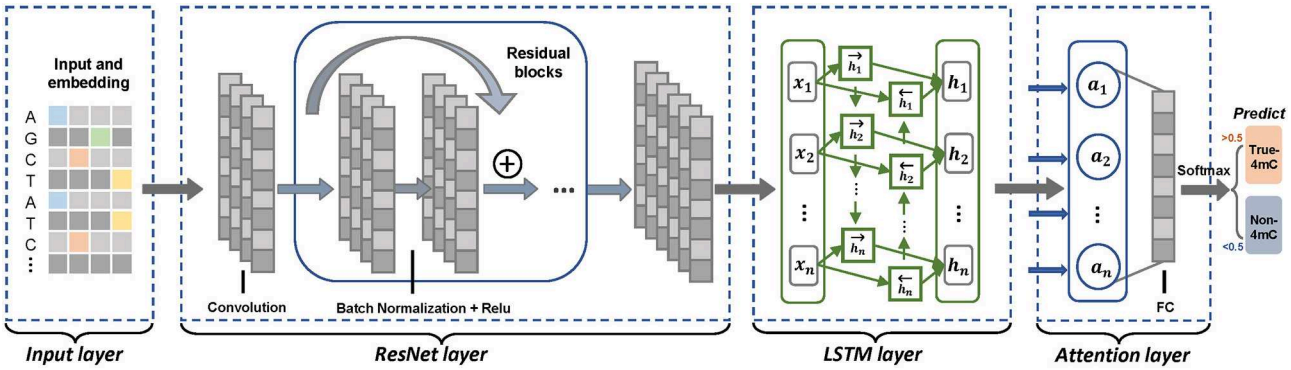


FIGURE 1 | The illustration of the deep neural network. It is a four-layer prediction system. Firstly, given sequences are fed to the input layer for feature representation using one-hot encoding, thus generating feature matrixes. Next, we feed the matrixes to the ResNet layer for extracting global features. After that, in LSTM layer, we use bidirectional LSTM model to gather feature information from two directions. In Attention layer can learn more relevant feature information. Ultimately, the features are connected with full connected layer and Softmax can make predictions. If the prediction score is higher than 0.5, the predicted sequences are 4mc sequences; otherwise, they are not.

$$\begin{pmatrix} i_t \\ f_t \\ C_t' \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ tanh \end{pmatrix} \left(\begin{pmatrix} W_x \\ W_h \\ W_c \end{pmatrix} [h_{t-1}, x_t, c_{t-1}] + \begin{pmatrix} b_i \\ b_f \\ b_g \end{pmatrix} \right)$$

$$\begin{aligned} C_t &= i_t C'_t + f_t C_{t-1} \\ o_t &= \sigma \left(W_x h_{t-1} + W_h x_t + W_c C_t + b_o \right) \\ h_t &= o_t \tanh \left(C_t \right) \end{aligned}$$

where i_t, f_t and o_t represent the input, forget and output gate, respectively; C'_t is an auxiliary value for calculating the cell memory C_t; t denotes the recurrent time step; W_x, W_h, W_c, and b are the corresponding weight values for each equation; and the current output of LSTM cell is h_t at time step t.

In consideration of bidirectional LSTM, the final LSTM network is composed of two LSTM networks with opposite directions. Hence, the i-th deoxynucleotide of the DNA sequence can be encoded as below:

$$h_i = [\overrightarrow{h_i} \bigoplus \overleftarrow{h_i}]$$

Attention Mechanism

Inspired by human attention, the attention mechanism is an idea for solving problems that focuses on the important factors while ignoring the unimportant. The attention mechanism can quickly filter out high-level information from noises, which has recently demonstrated great success in many relevant classification tasks. To take advantage of this, we applied the attention mechanism after the LSTM layer in the model to obtain the final distinctive feature representation. Let H be the output vectors $[h_1, h_2, \ldots, h_s]$ generated by LSTM layer, where s is the length of the DNA sequence. As shown in **Figure 1**, the following formulations are performed in the attention layer:

$$\begin{aligned} M &= \tanh \left(H \right) \\ \alpha &= softmax \left(W^T M \right) \\ r &= H \alpha^T \end{aligned}$$

where W^T is a transpose of the trained parameter vector W. Then the final representation of the attention layer can be encoded as below:

$$h^* = \tanh \left(r \right)$$

Softmax

The generated vectors h^* after the attention module are fed into a softmax layer for classification as input. The softmax score of class k will be calculated as follows:

$$\alpha_k = \frac{e^{h^*}}{\sum_{k=1}^{C} e^{h^*}}$$

where C denotes the total number of categories, and $C = 2$ when dealing with the binary classification tasks.

The softmax function maps and the output of neurons to numbers between (0–1) and normalizes the sum to 1. In other words, the output scores of each category can be converted into a relative probability by softmax. Therefore, the predicted label can be determined by comparing the predicted probability α_k for each class.

At last, we generated a multi-layered neural network integrating ResNet with a LSTM layer and an attention module, which incorporates the strengths behind ResNet, LSTM, and the attention mechanism. Through applying such a comprehensive network structure, feature extraction and learning are combined in an end-to-end manner, which can significantly improve the prediction performance.

Performance Indicators

In our experiment, we used the following four indicators to evaluate the predictive performance of our proposed model, including Accuracy (ACC), Sensitivity (SN), Specificity (SP), and Mathew's Correlation Coefficient (MCC). They are the four commonly used indicators for classifier performance evaluation in other Bioinformatics fields (Zhang et al., 2008, 2018a,b,c, 2019b,c,d; Wei et al., 2017b, 2019b; Zeng et al., 2017b, 2019c; Chen et al., 2018; Lu et al., 2018a,b; Fu et al., 2019; Gong et al., 2019; Jin et al., 2019; Liu and Li, 2019; Liu et al., 2019c,d; Manavalan et al., 2019a,b,c,d; Basith et al., 2020). Their calculation formulas are as follows:

$$\begin{cases} Sn = \frac{TP}{TP+FN} & 0 \le Sn \le 1 \\ Sp = \frac{TN}{TN+FP} & 0 \le Sp \le 1 \\ ACC = \frac{TP+TN}{TP+FP+TN+FN} & 0 \le ACC \le 1 \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN+FN) \times (TN+FP) \times (TP+FN) \times (TP+FP)}} & -1 \le MCC \le 1 \end{cases}$$

where TP (True Positive) represents the number of positive samples correctly predicted; TN (True Negative) represents the number of negative samples correctly predicted; FP (False Positive) represents the number of negative samples incorrectly predicted to be the positives; FN (False Negative) represents the number of positive samples incorrectly predicted to be the negatives.

Moreover, we also used the area under the ROC curve (AUC) is to quantitively measure the predictive performance of the model (Yang et al., 2018; Lv et al., 2019b; Niu et al., 2019). A higher AUC represents a better predictor (Hanley and McNeil, 1982; Liu et al., 2018; Feng et al., 2019; Lai et al., 2019).

RESULTS AND DISCUSSIONS

Comparison of the Proposed Method and Existing Predictors

To examine the predictive performance of our deep learning model, we compared several existing predictors with our model, including iDNA4mC (Chen et al., 2017), 4mCPred (Su et al., 2018), 4mcPred-SVM (Wei et al., 2018a), and 4mcPred-IFL (Wei et al., 2019a). It is worth noting that besides our predictor using deep learning, other compared predictors are all traditional machine learning algorithm -SVM and different handcrafted sequential features to train their respective models. For fair comparison, all the predictors are evaluated with 10-fold cross validation on the same dataset used in this study.

Table 2 lists the performances of the proposed method and four existing predictors. We can see that our proposed deep learning method achieves the highest performance in two out of three species (*C. elegans* and *A. thaliana*), with only one exception in *D. melanogaster*, in which our method is slightly worse than existing predictors. Specifically, for *C. elegans*, our predictor achieves 91.5%, 87.2%, 89.3%, and 0.787 in terms of SN, SP, ACC, and MCC, respectively. The overall performances (ACC and MCC) by our predictor are significantly better than the runner-up predictor—4mcPred-IFL (with the ACC of 88.0% and the MCC of 0.761). The more significant improvement is observed in *A. thaliana*, in which our predictor outperforms existing predictors in all metrics, leading by 5.7%, 2.2%, and 0.045 in terms of SN, ACC, and MCC, respectively. In addition, we found that our model remarkably improves the SN in all three species, demonstrating that our deep learning model can more accurately identify true 4mC sites. To better illustrate the difference between various models, we used Delong's test from the R package pROC to compare the ROC curves, confirming that the performance gain from fixed-length to full-length version is statistically significant ($p = 0.0005$). Generally, the comparative results demonstrate that our deep learning model is better than existing predictors using traditional machine learning algorithms in prediction of 4mC sites. More importantly, our deep learning model can automatically learn high-level feature representations to capture the characteristics of 4mC sites, rather than specify sequence-based features before model training as existing predictors did.

TABLE 2 | Performance comparison of the proposed Deep4mcPred and existing sequence-based predictors.

Species	Predictors	SN (%)	SP (%)	ACC (%)	MCC
C. elegans	iDNA4mc	79.0	77.0	78.0	0.560
	4mcPred	82.5	82.6	82.6	0.652
	4mcPred_SVM	82.4	80.7	81.5	0.631
	4mcPred_IFL	89.0	87.1	88.0	0.761
	Deep4mcPred	**91.5**	**87.2**	**89.3**	**0.787**
D. melanogaster	iDNA4mc	83.3	79.0	81.2	0.620
	4mcPred	82.4	82.1	82.2	0.646
	4mcPred_SVM	83.8	82.2	83.0	0.661
	4mcPred_IFL	86.5	88.0	87.3	0.745
	Deep4mcPred	**87.6**	**86.6**	**87.1**	**0.742**
A. thaliana	iDNA4mc	76.6	75.5	76.1	0.520
	4mcPred	75.5	78.0	76.8	0.536
	4mcPred_SVM	77.8	79.6	78.7	0.573
	4mcPred_IFL	80.3	84.0	82.2	0.644
	Deep4mcPred	**86.0**	**82.9**	**84.4**	**0.689**

The performances are evaluated with 10-fold cross validation. Note that the performances of the other methods are cited from existing studies, since the source codes of existing methods are not available. The value in bold indicates the optimal value of the indicator.

TABLE 3 | Performance comparison of the model using the attention mechanism and the model not using the attention mechanism.

Species	Models	SN (%)	SP (%)	ACC (%)	MCC
C. elegans	ResNet_LSTM_Attention	91.5	87.2	89.3	0.787
	ResNet_LSTM	90.9	87.2	89.0	0.781
D. melanogaster	ResNet_LSTM_Attention	87.6	86.6	87.1	0.742
	ResNet_LSTM	87.7	86.3	87.0	0.740
A. thaliana	ResNet_LSTM_Attention	86.0	82.9	84.4	0.689
	ResNet_LSTM	84.9	83.9	84.4	0.688

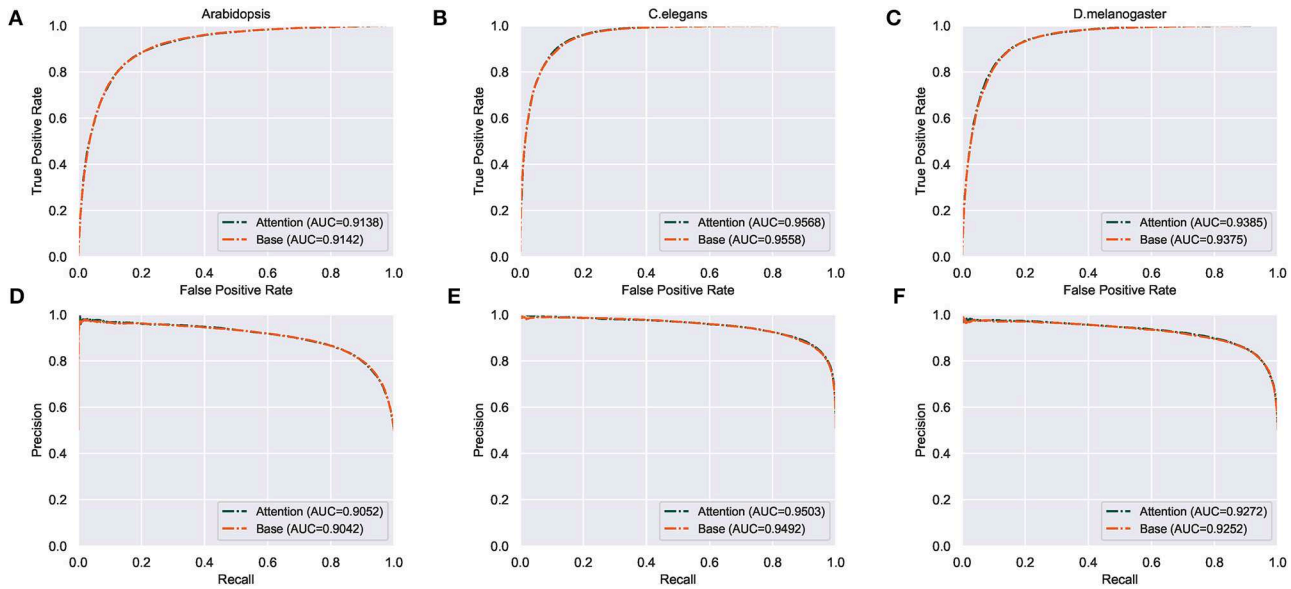


FIGURE 2 | Performance of the model using the attention mechanism and the model not using the attention mechanism. **(A–C)** represent the ROC curves of the two models in the three species. **(D–F)** represent the PR curves of the two models in the three species, respectively.

Performance Impact by Integrating Attention Mechanism

In this section, we evaluated whether or not the attention mechanism can improve the performance of 4mC site prediction. Subsequently, we compared the models taking into account attention mechanism and the model not taking into account attention mechanism for prediction. Both models were trained and evaluated with 10-fold cross validation on the dataset used in this study.

Results in **Table 3** show that training with the attention mechanism, the model achieves 89.3% in ACC and 0.787 in MCC for *C. elegans* dataset, achieves 87.1% in ACC and 0.742 in MCC for the *D. melanogaster* dataset, achieves 84.4% in ACC and 0.689 in MCC for the *A. thaliana* dataset, respectively. These results demonstrate that using the attention mechanism we can achieve good performances for 4mC sites prediction for different species. The comparison between the models using and not using the attention mechanism is shown in **Figure 2**. We can observe that the model using attention mechanism performs better than the model not using the attention mechanism in ROC and PR curves. The details of the performances for both models are listed in **Table 3**. Results show that using the attention mechanism, the deep learning model can achieve the average improvement of 0.1% roughly in three species as compared to the model not using the attention mechanism. This demonstrates that the attention mechanism indeed helps to capture discriminative feature representations.

CONCLUSIONS

In this study, we have proposed Deep4mcPred, a novel predictor for the prediction of DNA 4mC sites. Different from existing predictors using traditional machine learning algorithms (like SVM), Deep4mcPred is the first deep learning-based predictor, in which we integrate residual network and recurrent neural network–biLSTM to build a multi-layer deep learning predictive system. As compared to existing predictors, our proposed method has two advantages. First, our deep learning framework does not need to specify the features when training the predictive model. It can automatically learn the high-level features and capture the characteristic specificity of 4mC sites, benefiting to distinguish true 4mC sites from non-4mC sites. On the other hand, our deep learning method outperforms the traditional machine learning predictors in performance by benchmarking comparison. It demonstrates that the proposed Deep4mcPred is more effective in the DNA 4mC site prediction. Moreover, via experimental comparison, we found that attention mechanism introduced into the deep learning framework is useful to capture the critical features.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://server.malab.cn/Deep4mcPred.

AUTHOR CONTRIBUTIONS

RZ wrote the manuscript, designed experiments, and did the results analysis. ML provided the idea.

FUNDING

This work was supported by National Natural Science Foundation of China.

REFERENCES

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., et al. (2014). Deep speech 2: end-to-end speech recognition in english and mandarin. *International Conference on Machine Learning 2016* (New York, NY), 173–182.

Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020). Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med. Res. Rev*. 2020, 1–39. doi: 10.1002/med.21658

Chen, W., Feng, P., Liu, T., and Jin, D. (2018). Recent advances in machine learning methods for predicting heat shock proteins. *Curr Drug Metab*. 20, 224–228. doi: 10.2174/1389200219666181031105916

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Collobert, R., and Weston, J. (2008). "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning* (New York, NY: ACM), 160–167. doi: 10.1145/1390156.1390177

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* 35, 1469–1477. doi: 10.1093/bioinformatics/bty827

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*. 7:461. doi: 10.1038/nmeth.1459

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150. doi: 10.1093/bioinformatics/bts565

Fu, X., Ke, L., Cai, L., Chen, X., Ren, X., and Gao, M. (2019). Improved prediction of cell-penetrating peptides via effective orchestrating amino acid composition feature representation. *IEEE Access*. 7, 163547–163555. doi: 10.1109/ACCESS.2019.2952738

Gong, Y., Niu, Y., Zhang, W., and Li, X. (2019). A network embedding-based multiple information integration method for the MiRNA-disease association prediction. *BMC Bioinf*. 20:468. doi: 10.1186/s12859-019-3063-3

Hanley, J. A., and McNeil, B. J. J. R. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778. doi: 10.1109/CVPR.2016.90

Hong, Z., Zeng, X., Wei, L., and Liu, X. J. B. (2019). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 36, 1037–1043. doi: 10.1093/bioinformatics/btz694

Jin, B., Li, Y., and Robertson, K. D. (2011). DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer*. 2, 607–617. doi: 10.1177/1947601910393957

Jin, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L., and Su, R. (2019). DUNet: a deformable network for retinal vessel segmentation. *Knowledge-Based Syst*. 178, 149–162. doi: 10.1016/j.knosys.2019.04.025

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, K., Wang, D.-T, Fong, S., Liu, L.-S, Wong, K. K., and Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. *J. Med. Syst*. 42:139. doi: 10.1007/s10916-018-1003-9

Li, C.-C., and Liu, B. (2019). MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Brief. Bioinf*. bbz133. doi: 10.1093/bib/bbz133

Liang, S., Ma, A., Yang, S., Wang, Y., and Ma, Q. (2018). A review of matched-pairs feature selection methods for gene expression data analysis. *Comput. Struct. Biotechnol. J*. 16, 88–97. doi: 10.1016/j.csbj.2018.02.005

Liu, B. (2019). BioSeq-analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief. Bioinf*. 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019d). 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:e127. doi: 10.1093/nar/gkz740

Liu, B., Han, L., Liu, X., Wu, J., and Ma, Q. (2018). Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 99:1. doi: 10.1109/TCBB.2018.2816032

Liu, B., Li, C., and Yan, K. (2019b). DeepSVM-fold: protein fold recognition by combining Support Vector Machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinf*. bbz098. doi: 10.1093/bib/bbz098

Liu, B., and Li, K. (2019). iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids*. 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., Zou, Q., et al. (2019c). Computational methods for identifying the critical nodes in biological networks. *Brief. Bioinf*. bbz011. doi: 10.1093/bib/bbz011

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2017). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinf*. 14, 905–915. doi: 10.1109/TCBB.2016.2550432

Liu, Z.-Y., Xing, J.-F., Chen, W., Luan, M.-W., Xie, R., Huang, J., et al. (2019a). MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. *Horticul. Res*. 6:78. doi: 10.1038/s41438-019-0160-4

Lu, X., Li, X., Liu, P., Qian, X., Miao, Q., and Peng, S. (2018b). The integrative method based on the module-network for identifying driver genes in cancer subtypes. *Molecules* 23:183. doi: 10.3390/molecules23020183

Lu, X., Qian, X., Li, X., Miao, Q., and Peng, S. (2018a). DMCM: a data-adaptive mutation clustering method to identify cancer-related mutation clusters. *Bioinformatics* 35, 389–397. doi: 10.1093/bioinformatics/bty624

Luo, G.-Z., Blanco, M. A., Greer, E. L., He, C., and Shi, Y. (2015). DNA N 6-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol*. 16:705. doi: 10.1038/nrm4076

Lv, H., Zhang, Z. M., Li, S. H., Tan, J. X., Chen, W., and Lin, H. (2019b). Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform*. bbz048. doi: 10.1093/bib/bbz048

Lv, Z. B., Ao, C. Y., and Zou, Q. (2019a). Protein function prediction: from traditional classifier to deep learning. *Proteomics* 19:2. doi: 10.1002/pmic.201900119

Manavalan, B., Basith, S., Shin, T. H., Lee, D. Y., Wei, L., and Lee, G. (2019c). 4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. *Cells* 8:1332. doi: 10.3390/cells8111332

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019a). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi: 10.1093/bioinformatics/bty1047

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019b). 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

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019d). AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput. Struct. Biotechnol. J*. 17, 972–981. doi: 10.1016/j.csbj.2019.06.024

Nie, L. L., Deng, L., Fan, C., Zhan, W. H., and Tang, Y. J. (2018). Prediction of protein S-sulfenylation sites using a deep belief network. *Curr. Bioinform*. 13, 461–467. doi: 10.2174/1574893612666171122152208

Niu, S.-Y., Liu, B., Ma, Q., and Chou, W.-C. (2019). rSeqTU—a machine-learning based R package for prediction of bacterial transcription units. *Front. Genetics* 10:374. doi: 10.3389/fgene.2019.00374

Peng, L., Peng, M. M., Liao, B., Huang, G. H., Li, W. B., and Xie, D. F. (2018). The advances and challenges of deep learning application in biological big data processing. *Curr Bioinform*. 13, 352–359. doi: 10.2174/1574893612666170707095707

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. Bioinform.* 16, 1231–1239. doi: 10.1109/TCBB.2018.2858756

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* 34, 4196–4204. doi: 10.1093/bioinformatics/bty508

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Montreal, QC), 3104–3012.

Wang, Y., Zhang, S., Yang, L., Yang, S., Tian, Y., and Ma, Q. (2019). Measurement of conditional relatedness between genes using fully convolutional neural network. *Front. Genet.* 10:1009. doi: 10.3389/fgene.2019.01009

Wei, L., Chen, H., and Su, R. (2018b). 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., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2018a). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35, 1326–1333. doi: 10.1093/bioinformatics/bty824

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019a). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*. 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Wei, L., Tang, J., and Zou, Q. (2017a). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019b). Fast prediction of methylation sites using sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 16, 1264–1273. doi: 10.1109/TCBB.2017.2670558

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). 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. (2018c). ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Xiao, C.-L., Zhu, S., He, M., Chen, D., Zhang, Q., Chen, Y., et al. (2018). N6-methyladenine DNA modification in the human genome. *Molecular Cell*. 71, 306–18. e7. doi: 10.1016/j.molcel.2018.06.015

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens. *J.Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004

Yang, J., Ma, A., Hoppe, A. D., Wang, C., Li, Y., Zhang, C., et al. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Res*. 47, 7809–7824. doi: 10.1093/nar/gkz672

Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C., and Xie, Z. (2016). MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res*. 2016:gkw950. doi: 10.1093/nar/gkw950

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Magaz.* 13, 55–75. doi: 10.1109/MCI.2018.2840738

Yu, M., Ji, L., Neumann, D. A., Chung, D.-H, Groom, J., Westpheling, J., et al. (2015). Base-resolution detection of N 4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite-sequencing. *Nucleic Acids Res*. 43:e148. doi: 10.1093/nar/gkv738

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017a). Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Transact. Comput. Biol. Bioinform*. 14, 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017b). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol*. 13:e1005420. doi: 10.1371/journal.pcbi.1005420

Zeng, X., Wang, W., Chen, C., and Yen G. G. (2019c). A consensus community-based particle swarm optimization for dynamic community detection. *IEEE Trans. Cyber*. 99:1. doi: 10.1109/TCYB.2019.2938895

Zeng, X., Zhang, X., and Zou, Q. (2016). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinform*. 17, 193–203. doi: 10.1093/bib/bbv033

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2019b). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief. Bioinform*. bbz080. doi: 10.1093/bib/bbz080

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019a). deepDR: a network-based deep learning approach to *in silico* drug repositioning. *Bioinformatics* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zhang, J., Chen, Q., and Liu, B. (2019a). DeepDRBP-2L: a new genome annotation predictor for identifying DNA-binding proteins and RNA-binding proteins using convolutional neural network and long short-term memory. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 1–1. doi: 10.1109/TCBB.2019.2952338

Zhang, W., Chen, Y., Li, D., and Yue, X. (2018b). Manifold regularized matrix factorization for drug-drug interaction prediction. *J. Biomed. Inform*. 88, 90–97. doi: 10.1016/j.jbi.2018.11.005

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019c). SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions. *Inform. Sci*. 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019d). "A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics/IEEE*, ACM. doi: 10.1109/TCBB.2019.2931546

Zhang, W., Liu, J., Niu, Y. Q., Wang, L., and Hu, X. (2008). A Bayesian regression approach to the prediction of MHC-II binding affinity. *Comp. Methods Programs Biomed*. 92, 1–7. doi: 10.1016/j.cmpb.2008.05.002

Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., and Li, X. (2018c). Feature-derived graph regularized matrix factorization for predicting drug side effects. *Neurocomputing* 287, 154–162. doi: 10.1016/j.neucom.2018.01.085

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018a). SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput. Biol*. 14:e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, X., Zou, Q., Rodriguez-Paton, A., Zeng X. J. (2019b). Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 16, 283–91. doi: 10.1109/TCBB.2017.2776280

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

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 © 2020 Zeng and Liao. 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 April 2020
doi: 10.3389/fcell.2020.00213

SOD1 Promotes Cell Proliferation and Metastasis in Non-small Cell Lung Cancer via an miR-409-3p/SOD1/SETDB1 Epigenetic Regulatory Feedforward Loop

*Shilong Liu[1], Bin Li[2], Jianyu Xu[1], Songliu Hu[1], Ning Zhan[3], Hong Wang[4], Chunzi Gao[5], Jian Li[1] and Xiangying Xu[1,6]**

[1] Department of Thoracic Radiation Oncology, The Third Affiliated Hospital of Harbin Medical University, Harbin, China, [2] Department of Plastic Surgery, Nanfang Hospital of Southern Medical University, Guangzhou, China, [3] Department of Radiation Oncology, Xiang'an Hospital of Xiamen University, Xiamen, China, [4] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [5] Department of Oncology, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [6] Department of Radiation Oncology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Wencheng Shao,
Harvard Medical School, United States
Chao Qu,
United States Department of Veterans Affairs, United States
Takahiko Shimizu,
National Center for Geriatrics and Gerontology (NCGG), Japan
Shuyuan Zhang,
Stanford University, United States

***Correspondence:**
Xiangying Xu
xuxiangyingxxy@163.com

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

Received: 08 January 2020
Accepted: 11 March 2020
Published: 23 April 2020

Citation:
Liu S, Li B, Xu J, Hu S, Zhan N, Wang H, Gao C, Li J and Xu X (2020) SOD1 Promotes Cell Proliferation and Metastasis in Non-small Cell Lung Cancer via an miR-409-3p/SOD1/SETDB1 Epigenetic Regulatory Feedforward Loop. Front. Cell Dev. Biol. 8:213. doi: 10.3389/fcell.2020.00213

Superoxide dismutase 1(SOD1) is a major antioxidant with oncogenic effects in many human cancers. Although SOD1 is overexpressed in various cancers, the clinical significance and functions of SOD1 in non-small cell lung cancer (NSCLC), particularly the epigenetic regulation of SOD1 in NSCLC carcinogenesis and progression have been less well investigated. In this study, we found that SOD1 expression was upregulated in NSCLC cell lines and tissues. Further, elevated SOD1 expression could promote NSCLC cell proliferation, invasion and migration. While inhibition of SOD1 expression induced NSCLC G1-phase cell cycle arrest and promoted apoptosis. In addition, miR-409-3p could repress SOD1 expression and significantly counteract its oncogenic activities. Bioinformatics analysis indicated that SET domain bifurcated histone lysine methyltransferase1 (SETDB1) was involved in the epigenetic regulation of miR-409-3p and SOD1 expression and functions in NSCLC cells. Identification of this miR-409-3p/SOD1/SETDB1 epigenetic regulatory feedforward loop may provide new insights into further understanding of NSCLC tumorigenesis and progression. Additionally, our results incicate that SOD1 may be a potential new therapeutic target for NSCLC treatment.

Keywords: non-small cell lung cancer, SOD1, miR-409-3p, SETDB1, methylation, feedforward loop

INTRODUCTION

Due to advances in clinical and experimental oncology, particularly the successful reduction of smoking prevalence, the incidence rate of lung cancer has continued to decline in recent decades, Nevertheless, the prognosis for patients with non-small cell lung cancer (NSCLC) remains unfavorable. Lung cancer remains among the leading causes of cancer death with a 5-year survival

rate of only 18% across all ethnicities and disease stages (Siegel et al., 2018). Lung cancer, particularly NSCLC, has usually reached an advanced stage when first diagnosed. Hence, early diagnosis and treatment are vital for improving survival rates, and substantial efforts have been made to identify molecular markers that can predict patient prognosis (Zhou et al., 2016), however, additional new treatment strategies are still needed.

Superoxide dismutases (SODs) are enzymes required for the conversion of superoxide into oxygen and hydrogen peroxide. There are three SOD subtypes: SOD1, SOD2, and SOD3. Among these three SOD enzymes, SOD1 (Cu/Zn superoxide dismutase) is the major cytosolic form and provides 80% of total cellular SOD activity (Watanabe et al., 2014). Although SOD1 is primarily distributed in the cytoplasm, it is also found in the mitochondrial intermembrane space and the nucleus (Tsang et al., 2014). The function of SOD2 and SOD3 are similar to that of SOD1, but they are located in different cell compartments. In response to oxidative stress, SOD1 maintains low levels of superoxide to protected the cell from death in response to oxidative stress damage (Tsang et al., 2014). The antioxidant effect of SOD1 can also provide protection for cancer cells or other dysfunctional cells. Mutations of the *SOD1* gene have been linked to numerous human diseases and cancers, such as and Down syndrome and familial amyotrophic lateral sclerosis (ALS), Indeed 20% of ALS cases are associated with mutations in SOD1 (Brasil et al., 2019), Somwar et al. (2011) reported that SOD1 was overexpressed in lung adenocarcinomas when compared with the normal lung tissue, while Glasauer et al. (2014) found that inhibition of SOD1 by the small molecule ATN-224 induced NSCLC cell death.

SOD1 also acts as a metabolic focal point, integrating O_2, nutrients, and reactive oxygen species (ROS) to direct energy metabolism (Tsang et al., 2018). Deficiency of SOD1 decreased the lifespan and accelerated aging in SOD1(−/−) mouse model (Watanabe et al., 2014; Zhang et al., 2017). Furthermore, the SOD1 inhibitor, ATN-224, has been tested in phase 1 clinical trials in patients with solid tumors (Lowndes et al., 2008) and in phase 2 clinical trials for prostate cancer (Lin et al., 2013), however, there have been few reports on the clinical significance of SOD1 functions in lung cancer, particularly the mechanism underlying the role of SOD1 in progression and carcinogenesis.

MicroRNAs make up a class of small non-coding RNAs that regulate gene expression at the post-transcriptional level through binding to specific sequences through binding to specific in the 3′untranslated regions (3′UTRs) of target mRNAs, leading to transcript degradation or translational inhibition (Lu and Clark, 2012). Dysregulation of miRNAs is involved in numerous human biological and pathological processes, including cell proliferation, differentiation, development, apoptosis, and tumorigenesis (Wu et al., 2019). miR-409-3p, maps to chromosome 14q32.31, and has been shown significantly downregulated in lung adenocarcinoma tissues when compared with corresponding noncancerous tissues, and can inhibit growth, migration, and invasion, as well as inducing apoptosis in lung adenocarcinoma cells via inactivation of Akt signaling by targeting c-Met (Wan et al., 2014).

In our study, we found that SOD1 expression levels are significantly increased in NSCLC compared with normal lung tissues and cells using bioinformatic and laboratory experiments. Furthermore, high levels of SOD1 promoted lung cancer cell proliferation and metastasis, while miR-409-3p inhibited SOD1 activity through binding to its 3′ UTR. We also found that SET domain bifurcated histone lysine methyltransferase 1 (SETDB1) may contribute to the interaction between miR-409-3p and SOD1 by an epigenetic transcription factor.

MATERIALS AND METHODS

Clinical Tissue Samples and Cell Lines

Tissue specimens (n = 196) obtained from patients diagnosed with stage I–IIIb NSCLC who underwent surgery at The Third Affiliated Hospital of Harbin Medical University between March 2007 and December 2009 were used for immunohistochemical staining. Eighteen pairs of NSCLC tumor and adjacent normal tissue samples were collected during surgery between April and August 2016, immediately frozen in liquid nitrogen and stored at −80°C for further analysis. None of the patients underwent any therapy before surgery. Informed consent was obtained from all patients. The study was approved by the Ethics Committee of The Third Affiliated Hospital of Harbin Medical University.

Seven NSCLC cell lines [A549, PC-9, NCI-H1299, NCI-H460, NCI-H1650, NCI-H520 and human bronchial epithelial cells (16HBE)] were purchased from American Type Culture Collection (ATCC, Manassas, VA, United States). PC-9 and 16HBE cells were cultured in DMEM (GIBCO, Invitrogen, Carlsbad, CA, United States), other cells were cultured in RPMI 1640 basic medium (GIBCO, Invitrogen, Carlsbad, CA, United States), supplemented with 10% fetal bovine serum (FBS; Invitrogen, Carlsbad, CA, United States) and antibiotics (Invitrogen, Carlsbad, CA, United States) at 37°C in a 5% CO_2 atmosphere.

Immunohistochemistry

Tissue sections were deparaffnized in xylene and rehydrated in a series of graded alcohol solutions according to standard procedures. Antigen retrieval was performed by placing the slides in citrate buffer (0.01 M, pH 6.0) heated for 10 min and cooled down naturally to enhance immunoreactivity. The slides were immersed in hydrogen peroxide (3%) for 15 min at room temperature. the specimens were then incubated with the SOD1 primary antibody (1:100, rabbit polyclonal; wanleibio china) at 4°C overnight. After incubation with biotin-labeled secondary antibodies goat anti-rabbit IgG antibodies (1:200 Beyotime A0277 china), at room temperature for 30 min, the specimens were incubated in Horseradish Peroxidase-labeled Avidinat 37°C for 30 min. The slides were stained using DAB and counterstained using hematoxylin. Finally, each section was dehydrated by gradient alcohol and covered with a coverslip. The staining results were analyzed by two independent pathologists experienced in evaluating IHC, both of whom were blinded to the clinicopathological data. The staining results were scored according to the following criteria: the sum of intensity and proportion of positively stained cells. The staining intensity was classifed: 0 (no staining), 1 (light yellow/weak staining), 2 (yellow-brown/moderate staining),

or 3 (brown/strong staining). The proportion of positively stained tumor cells was scored as follows: 1 (positive cells 1–25%), 2 (positive cells 26–50%), 3 (positive cells 51–75%), or 4 (positive cells 75–100%). Thus, the final score ranged from 0 to 7, and the median value of score 3 was used to distinguish low versus high SOD1 expression. The specimen with a final score of 3 was classifed as having low expression. Otherwise, the specimen was classifed as having high expression. Any discrepancies between scores were reviewed by the two pathologists plus a senior pathologist until a consensus was reached.

RNA Extraction and Quantitative Reverse Transcription Polymerase Chain Reaction (qRT-PCR) Assays

Total RNA was extracted from tissue samples and cells using Trizol Reagent (Invitrogen, Carlsbad, CA, United States), cDNA was synthesized using the Transcriptor First Strand cDNA Synthesis Kit (Roche, Mannheim, Germany), and qRT-PCR performed using a Fast Start Universal SYBR Green Master (ROX) (Roche, Mannheim, Germany) on an ABI7500 system, according to the manufacturer's instructions. Results were normalized to GAPDH expression levels. Fold-change in gene expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method.

Cell Transfection

Cells were transfected using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, United States), following the manufacturer's instructions. Three *SOD1*-specific si-RNA oligonucleotides (si-SOD1-1: 5′-TTC GAG CAG AAG GAA AGT AAT GGA CCA-3′,si-SOD1-2: 5′-GCA GAG GGA GAA TGC TTA GCA-3′, and si-SOD1-3: 5′-GGCCTGCATGGATTCCATG-3′) and scrambled negative control (FAM-F: 5′-GUC ACA CGG GAA GAG AGU UAA AGA CUA -3′ R: 5′-GGA UAU GGG AAG AGC GUA GUU AAU-3′) were designed and purchased from Invitrogen (Carlsbad, CA, United States). The *SOD1*sequence, 5′-TCCCTTGGATGTAGTCTGAGGACTCCATT-3′ was synthesized and cloned into the pLVX-puro vector. Lentiviral particles were constructed and packaged by GeneChem Co., Ltd. (Shanghai, China). miR-409-3p mimic (5′-GAAUGUUGCUCGGUGAACCCCU-3′) and a negative control (5′-ACTACTGAGTGACAGTAGA-3′) were also obtained from GeneChem Co., Ltd. (Shanghai, China). Sequences of other primers are provided in **Table 1**.

Cell Proliferation Assay

Cell viability was assessed using a Cell Counting Kit-8 (CCK-8; Dojindo, Kumamoto, Japan), according to the manufacturer's protocol. Briefly, cells were plated in 96-well plates at a density of 2×10^3 cells per well, and incubated for 24, 48, 72, or 96 h, and treated with 10 μl/well of CCK-8 solution at the indicated time points after 2 h of incubation, and cell proliferation curves were plotted based on absorbance values (OD = 450 nm) at each time point. The assay was conducted using five replicate wells per sample and three parallel experiments were performed.

TABLE 1 | Oligonucleotides used in this study.

Gene	Orientation	5′- 3′ primer sequence
miR-409-3p inhibitor	Forward	GAATGTTGCTCGGTGA
	Reverse	GTGCAGGGTCCGAGGT
U6	Forward	GCTTCGGCAGCACATATACT
	Reverse	GTGCAGGGTCCGAGGTATTC
GAPDH	Forward	GCACCGTCAAGGCTGAGAC
	Reverse	TGGTGAAGACGCCAGTGGA
SOD1	Forward	CTGAAGGCCTGCATGGATTC
	Reverse	CCAAGTCTCCAACATGCCTCT

Flow Cytometry Assay

An AnnexinV-PE Apoptosis Detection Kit (BD Biosciences,San Jose, CA,United States) was used to examine apoptosis, according to the manufacturer's instructions. Briefly, cells were washed twice in cold phosphate-buffered saline (PBS) and then harvested and resuspended in 1 × binding buffer. Next, 100 μl of suspended cells (1×10^5 cells) was transferred to a 5 ml culture tube, and 5 μl of annexin V-PE and 5 μl of (propidium iodide) PI were added. Cells were gently vortexed and incubated for 15 min at room temperature in the dark, followed by addition of 400 μl 1 × binding buffer to each tube. For cell cycle analysis, the CycleTEST^{TM} PLUS DNA Reagent Kit (BD Biosciences. Cat No. 340242) was used. Briefly, cells were washed with ice-cold PBS and fixed with 75% ethanol overnight at −20°C. After fixation, cells were washed and resuspended twice in PBS and then incubated with PI and RNase for 30 min at room temperature. A Canto II flow cytometer (BD Biosciences, San Jose, CA, United States) was used to evaluate levels of apoptosis and the cell cycle in each sample, following the manufacturer's instructions.

Cell Invasion and Migration Assay

Cells were plated in serum-free medium in the upper chamber of transwell plates (8-μm pore size, Corning, Tewksbury, MA, United States). For the invasion assay, transwell membranes were precoated with 45 μg matrigel (Sigma Aldrich, United States) to form a matrix barrier, while matrigel was not used for migration assay. Standard medium supplemented with 15% FBS was added to the bottom chambers. After incubation at 37°C for 12 h (migration) or 24 h (invasion), cells that had migrated to the lower membranes surfaces were fixed in 90% ethanol for 30 min, stained with 0.5% crystal violet, and photographed under microscope. Cell numbers in five random fields in each well were counted under a light microscope (magnification, × 200) and the number of invasion or migration cell normalized to the total cell number, normalized invasion cell or migration cell number = actual invasion cell or migration cell number/each cell growth rate. Each experiment was performed in triplicate.

Luciferase Reporter Assay

A *SOD1* 3′-UTR sequence containing an miR-409-3p binding site was cloned and inserted into the pmiR-RB-luciferase reporter vector (Promega, United States), to generate the *SOD1* 3′-UTR-WT construct. Mutation of the *SOD1* miR-409-3p binding site was conducted by site-directed mutagenesis using the Quick

Change Lightning kit (Stratagene, La Jolla, CA, United States) to generate the SOD1 3′-UTR-MUT construct. For the experiments: cells were plated in 24-well plates, incubated overnight, and then H460 cells co-transfected with SOD1 3′-UTR-WT, SOD1 3′-UTR-MUT, miR-409-3p mimic, or miR-409-3p inhibitor. Luciferase activity was tested using the Dual-Luciferase Reporter Assay System (Promega, United States). Renilla activity was used to normalized luciferase activity values for each sample.

Western Blot Analysis

Protein samples were extracted using RIPA lysis buffer containing 1% phenylmethanesulfonyl fluoride. Protein concentrations in lysates were measured using the protein BCA assay kit (Waileibio WAL004,China) and aliquots containing 40 μg protein separated by 10% sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), and transferred to polyvinylidene fluoride membranes. Then, membranes were incubated with SOD1 antibody (1:500) overnight at 4°C, followed by incubation with horseradish peroxidase-labeled goat anti-rabbit IgG antibody (1:1000). Bound antibodies were detected using an ECL Western Blotting Detection system and SOD1 levels quantified by densitometric analysis of the protein bands relative to total protein loaded, using Gel-Pro-Analyzer software. β-actin was used as the loading control. All antibodies were purchased from Wanlei Bio (Shenyang, China).

Statistical Analysis

All data are presented as mean ± standard deviation (SD) from three independent experiments. Statistical analyses were performed using GraphPad Prism version 6.0 (GraphPad Software Inc., La Jolla, CA, United States), R software, or SPSS 21.0 for Windows (SPSS, Chicago, IL, United States). Clinicopathological data were analyzed using the chi-square test and survival curves were assessed by Kaplan-Meier analysis. The Cox proportional hazards model was used to identify independent prognostic factors for overall survival (OS) and disease-ress survival (DFS). Quantitative variables were analyzed using the Student's *t*-test, while the chi-square or Fisher's exact tests were used to compare qualitative variables. *P*-values < 0.05 were considered statistically significant.

RESULTS

Expression of SOD1 Is Elevated in NSCLC Tissues and Cell Lines

First, we used an algorithm to extract expression data for *SOD1* in NSCLC from The Cancer Genome Atlas (TCGA) database. *SOD1* expression was significantly higher in lung adenocarcinoma (ADC) and lung squamous cell carcinoma (SCC) than in normal tissues (**Figure 1A**). Further data analysis demonstrated that *SOD1* expression also differed according to TNM stage (**Figure 1B**). In addition, we found that mass pan-cancer *SOD1* expression levels differed from those in normal tissues (**Figure 1C**), also the *SOD1* expression levels varied in different normal tissues (**Supplementary Figure S1**).

Next, we used qRT-PCR, western blotting, and immunohistochemical staining to verify our findings regarding SOD1 expression in NSCLC tissue samples. Eighteen pairs of cancer and corresponding adjacent normal tissue samples were used for qRT-PCR and western blot assays. The results showed that both mRNA and protein expression levels of SOD1 expression in cancer tissues were significantly higher than those in adjacent normal tissues (**Figures 1D,E**). Furthermore, immunohistochemical staining results show a significantly darker brown staining SOD1 in cancerous than adjacent tissues; with SOD1 expression was upregulated in 62.2% (122/196) of tumor tissues but only in 12.2% (22/196) of normal tissue samples ($p < 0.05$). These results confirm that SOD1 expression is significantly increased in tumors relative to adjacent normal tissues (**Figure 1F** and **Supplementary Figure S2**).

Next, we used qRT-PCR and western blotting to verify the differential expression of SOD1 in human NSCLC cell lines and a normal human bronchial epithelial cell line. The results confirmed that SOD1 was expressed at significantly higher levels in NSCLC than normal bronchial epithelial cells at both the mRNA and protein levels (**Figure 1G** and **Supplementary Figure S3**).

Association of SOD1 With Clinicopathological Characteristics in Patients With NSCLC

Immunohistochemical analysis confirmed that high expression of SOD1 was associated with larger tumor size ($p = 0.011$), lymph node metastasis ($p < 0.001$), and advanced pTNM stage ($p < 0.001$) in NSCLC. SOD1 was highly expressed in 62.2% (122/196) of NSCLC tissue samples, including 35.2% of ADC and 64.8% of SCC specimens (**Table 2**). No significant correlation was observed between SOD1 expression and age, sex, differentiation or smoking status.

SOD1 Expression Can Predict Survival Prognosis in Patients With NSCLC

Next, we used univariate and multivariate COX regression analysis to determine whether SOD1 expression is related to survival in patients with NSCLC (**Table 3**). Univariate analysis indicated that advanced TNM stage, larger tumor size, lymph node metastasis, no adjuvant chemotherapy, and high SOD1 expression levels were factors associated with inferior overall survival (OS) with multivariate analysis, TNM stage (HR, 2.421; 95% CI, 1.221–4.801; $P = 0.001$), and SOD1 expression level (HR, 1.858; 95% CI, 1.164–2.966; $P = 0.009$) were identified as independent prognostic factors for OS. In addition, univariate analysis of disease-free survival (DFS) of patients with NSCLC showed that advanced TNM stage, large tumor, lymph node metastasis, no adjuvant chemotherapy, no adjuvant radiotherapy, and high SOD1 expression predicted poorer DFS, while TNM stage (HR, 2.055; 95% CI, 1.100–3.840; $p = 0.024$), adjuvant radiation therapy (HR, 0.497; 95% CI, 0.263–0.940; $P = 0.032$), and SOD1 expression (HR, 1.605; 95% CI, 1.056–2.440; $P = 0.027$) were independent prognostic factors for DFS on multivariate analysis.

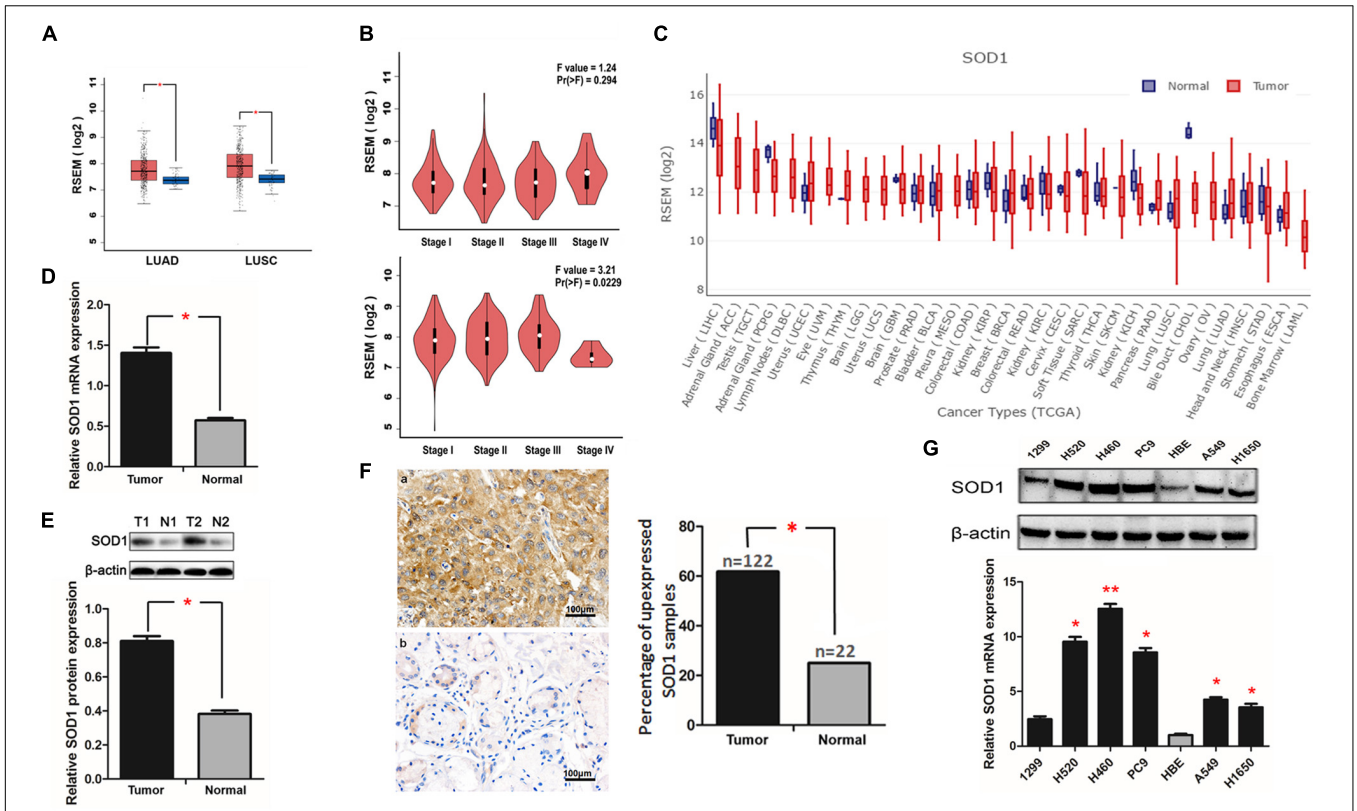


FIGURE 1 | Superoxide dismutase 1 (SOD1) expression is elevated in non-small cell lung cancer (NSCLC) tissues and cell lines. **(A,B)** SOD1 was overexpressed in both ADC and SCC tumors relative to normal tissues, and its expression also differed according to TNM stage as recorded in TCGA database. **(C)** Pan-cancer analysis of SOD1 expression. **(D,E)** Representative qRT-PCR and western blot analysis of SOD1 expression in 18 paired NSCLC clinical tissue specimens. Histogram of pooled data from NSCLC (T, $n = 18$) and adjacent normal lung (N, $n = 18$) tissue samples. Data are expressed as means ± standard deviation (SD). $^{*}P < 0.05$. **(F)** Representative immunohistochemistry images of NSCLC (T) and adjacent normal lung (N) tissue samples. Significantly darker brown staining of SOD1 protein was detected in cancer tissues **(a)** than in adjacent normal tissues **(b)**. Histogram of pooled data from NSCLC (tumor, $n = 122$) and normal lung (normal, $n = 22$) tissue. The percentage of NSCLC tissues with high SOD1 expression was significantly greater than that of normal lung tissues; $^{*}P < 0.05$. **(G)** qRT-PCR and western blot analysis of SOD1 expression in NSCLC cell lines and a normal bronchial epithelial cell line. The histogramillustrates of overexpression of SOD1 mRNA in NSCLC tumor tissues relative to normal lung tissues; $^{*}P < 0.05$, $^{**}P < 0.01$.

Kaplan-Meier analysis revealed that high levels of SOD1 expression was correlated with poor prognosis of both OS ($\chi 2 = 20.37$; $P < 0.001$, **Figure 2A-a**) and DFS ($\chi 2 = 20.37$; $p < 0.001$, **Figure 2A-b**) in our patient population. Further stratified analysis showed that SOD1 levels were significantly negatively correlated with OS ($\chi 2 = 6.827$; $P = 0.009$, **Figure 2B-a**) and DFS ($\chi 2 = 4.467$; $P = 0.035$, **Figure 2C-a**) in ADC as well as with OS ($\chi 2 = 6.182$; $P = 0.013$, **Figure 2B-b**) and DFS ($\chi 2 = 5.062$; $P = 0.024$, **Figure 2C-b**) in SCC.

To further verify the impact of SOD1 levels on prognosis in patients with NSCLC, we used the GEPIA analysis tool (Tang et al., 2017) to evaluate the expression levels of SOD1 and NSCLC survival progression in large database, and we found that SOD1 expression was significantly negatively correlated with OS in patients with ADC ($P = 0.014$, **Figure 2D-a**), but positively correlated with SOD1 expression the OS in SCC patients ($P = 0.032$, **Figure 2D-b**). This result is likely to be related to the significantly reduced expression of SOD1 in stage IV SCC. Nevertheless, No significant association was found between SOD1 levels and DFS in patients with ADC or SCC ($P = 0.34$, **Figure 2E-a** and $P = 0.55$, **Figure 2E-b**, respectively), or between SOD1 levels and OS and DFS in patients with NSCLC overall using this dataset ($P = 0.86$, **Figure 2F-a** and $P = 0.65$, **Figure 2F-b**, respectively).

SOD1 Can Promote the Proliferation, Invasion, and Migration of NSCLC Cells

Based on our findings of elevated SOD1 expression in NSCLC cell lines, we selected the highest (H460) and lowest (H1299) expressing cell lines for transfection with si-SOD1 and pLVX-puro-SOD1, respectively. The knockdown efficiency of si-SOD1-1 in H460 cells was most significant test at both the mRNA and protein levels (**Figure 3A** and **Supplementary Figure S4**), while SOD1 expression was significantly upregulated in H1299 cells after pLVX-puro-SOD1 transfection (**Figure 3B**); subsequently, we used these two cell lines to perform loss- and gain-of-function experiments, respectively.

Proliferative capacity is essential for the growth and development of malignant tumors; therefore, we used the

TABLE 2 | Association between SOD1 expression and clinicopathological characteristics patients with NSCLC.

Variable	All patients	SOD1 expression		*P*
		High (%)	Low (%)	
	(*n* = 196)	(*n* = 122)	(*n* = 74)	
Smoking				
Never	80	53 (43)	27 (36)	0.337
Ever	116	69 (57)	47 (64)	
Gender				
Male	129	79 (65)	50 (68)	0.687
Female	67	43 (35)	24 (32)	
Age (years)				
<60	89	54 (44)	35 (47)	0.679
≥60	107	68 (56)	39 (53)	
Differentiation				
Well	20	11(9)	9 (12)	0.846
Moderate	75	46 (37)	29 (39)	
Poor	101	65 (44)	36 (49)	
Histological cell type				
Adenocarcinoma	74	43 (35)	31 (42)	0.352
Squamous cell carcinoma	122	79 (65)	43 (58)	
PStage				
I	93	45 (37)	48 (65)	< 0.001*
II	40	25 (20)	15 (20)	
III	63	52 (43)	11 (15)	
pT classification				
T1	34	14 (11)	20 (27)	0.011*
T2	129	82 (67)	47 (64)	
T3/4	33	27 (22)	7 (9)	
Lymph node metastasis				< 0.001*
N0	114	56 (46)	58 (78)	
N1	26	17 (14)	9 (12)	
N2	56	49 (40)	7 (10)	
Adjuvant chemotherapy				
Yes	121	76 (62)	45 (61)	0.836
No	75	46 (38)	29 (39)	
Adjuvant radiotherapy				
Yes	89	59 (48)	30 (41)	0.286
No	107	63 (52)	44 (59)	

*Abbreviations: NSCLC, non-small cell lung cancer; pTNM stage, tumor node metastasis (pathological stage); pT, pathological T stage; n, number of patients. Ever, smoking at any time from the beginning of life. P-value, the difference of clinicopathological characteristics between the SOD1 high expression group and low expression group. *P < 0.05 was considered statistically significant.*

CCK-8 assay to test the effect of SOD1 on NSCLC cell proliferation. Compared with controls and normals, knockdown of SOD1 significantly reduced proliferation ability of H460 cells (**Figure 3C**). In contrast, cell proliferation was significantly enhanced in H1299 cells over-expressing SOD1 (**Figure 3D**).

Migration and invasion are vital for tumor progression and metastasis. Here, we used transwell assays to assess cell invasive and migratory ability. According the cell proliferation rate can affect the cell number went through the trans-well, we use the respective cell proliferation rates to normalized the number of migrate and invaded cells, and use this to evaluate the ability of migration and invasion. Knockdown of SOD1 significantly repressed the invasion and migration of H460 cells, compared with controls and normals (**Figure 3E**). Conversely, overexpression of SOD1 markedly enhanced the invasive and migratory of H1299 cells (**Figure 3F**).

SOD1 Reduces Apoptosis and Promotes Cell Cycle Progression of NSCLC Cells

To further explore the mechanism underlying how SOD1 enhances NSCLC cell proliferation, we examined the effect of SOD1 on apoptosis and the cell cycle of NSCLC cells using flow cytometry. Knockdown of SOD1 expression led to significantly increases of both early and late apoptotic H460 cells, relative to untreated control cells (**Figure 3G**). In contrast, overexpression of SOD1 in H1299 cells led to a significantly decreased apoptosis rate (**Figure 3H**). Regarding the cell cycle, flow cytometry revealed that inhibition of SOD1 result in G1-phase arrest, We observed a significant increase in the number of cells in G1-phase and a concomitant significant decrease in those in S-phases, relative to control group H460 cells (**Figure 3I**). Conversely, overexpression of SOD1 in H1299 cells promoted cell cycle progression, reducing the proportion of cells in G1-phase and increasing the proportion of those in S phases (**Figure 3J**).

SOD1 Expression Can Be Regulated by miR-409-3p

MicroRNAs are critical for tumor occurrence and progression and can act as oncogenes or tumor suppressor genes. Here, we identified 49 microRNAs related to SOD1 in starBase v3.0 (Li et al., 2014), and used the TAM database to functionally annotate them. Twelve microRNAs related to NSCLC were identified. Finally, only miR-409-3p was predicted to bind to the *SOD1* 3′ UTR (**Figures 4A,B**). Furthermore, starBase v3.0 data analysis indicated that expression of *SOD1* and miR-409-3p was negatively correlated in NSCLC, particularly in SCC ($P = 0.0127$, **Figure 4C**). The data indicated that miR-409-3p might inhibit *SOD1* expression. So we speculate that *SOD1* may be a target mRNA of miR-409-3p.

miR-409-3p Inhibits SOD1 Expression Through Direct Interaction

To verify the bioinformatics data indicating that *SOD1* is a target of miR-409-3p, we designed additional experiments. First, we evaluated 18 pairs of cancerous and corresponding normal adjacent tissue samples from patients with NSCLC. Total RNA was extracted for qRT-PCR analysis and the results showed that miR-409-3p was significantly decreased in NSCLC relative to adjacent normal tissues (**Figure 4D**). In addition, we compared the expression of miR-409-3p in NSCLC and normal human bronchial epithelial cell lines. The results confirmed that miR-409-3p is expressed at significantly lower levels in NSCLC cells (**Figure 4E**). These findings are consistent with a previous report demonstrating that miR-409-3p is significantly downregulated in NSCLC compared with corresponding noncancerous tissues

TABLE 3 | Univariate and multivariate analyses of overall survival and disease-free survival.

	OS			DFS		
	Univariate analysis	Multivariate analysis		Univariate analysis	Multivariate analysis	
Variable	***P***	**HR (95% CI)**	***P***	***P***	**HR (95% CI)**	***P***
Age						
<60						
≥60	0.245	–	–	0.453	–	–
Gender						
Female						
Male	0.106	–	–	0.368	–	–
Smoking	0.126	–	–	0.103		
Never						
Ever						
Differentiation						
Good						
Moderate						
Poor	0.309	–	–	0.181	–	–
Histological cell type						
Squamous cell carcinoma						
Adenocarcinoma	0.249	–	–	0.062	–	–
pT stage						
I						
II						
III	0.152	–	–	0.002*	1.334 (0.958 to 1.859)	0.088
pTNM stage						
I						
II						
III	< 0.001*	2.421 (1.221 to 4.801)	0.011*	< 0.001*	2.055 (1.100 to 3.840)	0.024*
Lymph node metastasis						
Present						
Absent	< 0.001*	0.928 (0.561 to 1.534)	0.771	0.001*	0.881(0.547 to1.419)	0.602
Adjuvant chemotherapy						
No						
Yes	0.003*	0.982 (0.496 to 1.946)	0.959	0.018*	1.154(0.657 to2.028)	0.618
Adjuvant radiotherapy						
No						
Yes	0.165	–	–	0.008*	0.497(0.263 to 0.940)	0.032*
SOD1 expression						
Low	< 0.001*	1.858 (1.164 to 2.966)	0.009*	0.002*	1.605 (1.056 to 2.440)	0.027*
High						

*Abbreviations: NSCLC, non-small cell lung cancer; pTNM stage, tumor node metastasis (pathological stage); pT, pathological T stage; n, number of patients; OS, overall survival; DFS, disease-free survival; HR, hazard ratio; CI, confidence interval. pTNM stage, lymph node metastasis, and SOD1 expression were selected for inclusion in multivariate analysisfor OS. Histological cell type, pTNM stage, lymph node metastasis, and SOD1 expression were selected for inclusion in multivariate analysis for DFS. *P < 0.05 was considered statistically significant.*

(Qu et al., 2018). Furthermore, qRT-PCR revealed that miR-409-3p expression in H460 cells transfected with si-SOD1-1 was significantly higher than that in untreated controls (**Figure 4F**), whereas miR-409-3p expression was significantly lower in H1299 cells transfected with pLVX-puro-SOD1 than that in controls (**Figure 4F**).

Next, we transfected miR-409-3p mimics into H460 cells and miR-409-3p inhibitors into H1299 cells. Subsequent qRT-PCR analysis showed that *SOD1* expression was significantly reduced in H460 cells, while it was significantly increased in H1299 cells (**Supplementary Figure S5**). Western blotting confirmed that SOD1 protein expression levels were also negatively regulated by miR-409-3p expression (**Figure 4G**). these result confirm that the expression of SOD1 and miR-409-3p was negatively correlated in NSCLC.

We predicted that miR-409-3p could partially bind to the *SOD1* 3′ UTR based on starBase v3.0 data nanlysis (**Figure 4H**). the results of luciferase reporter assays showed that luciferase

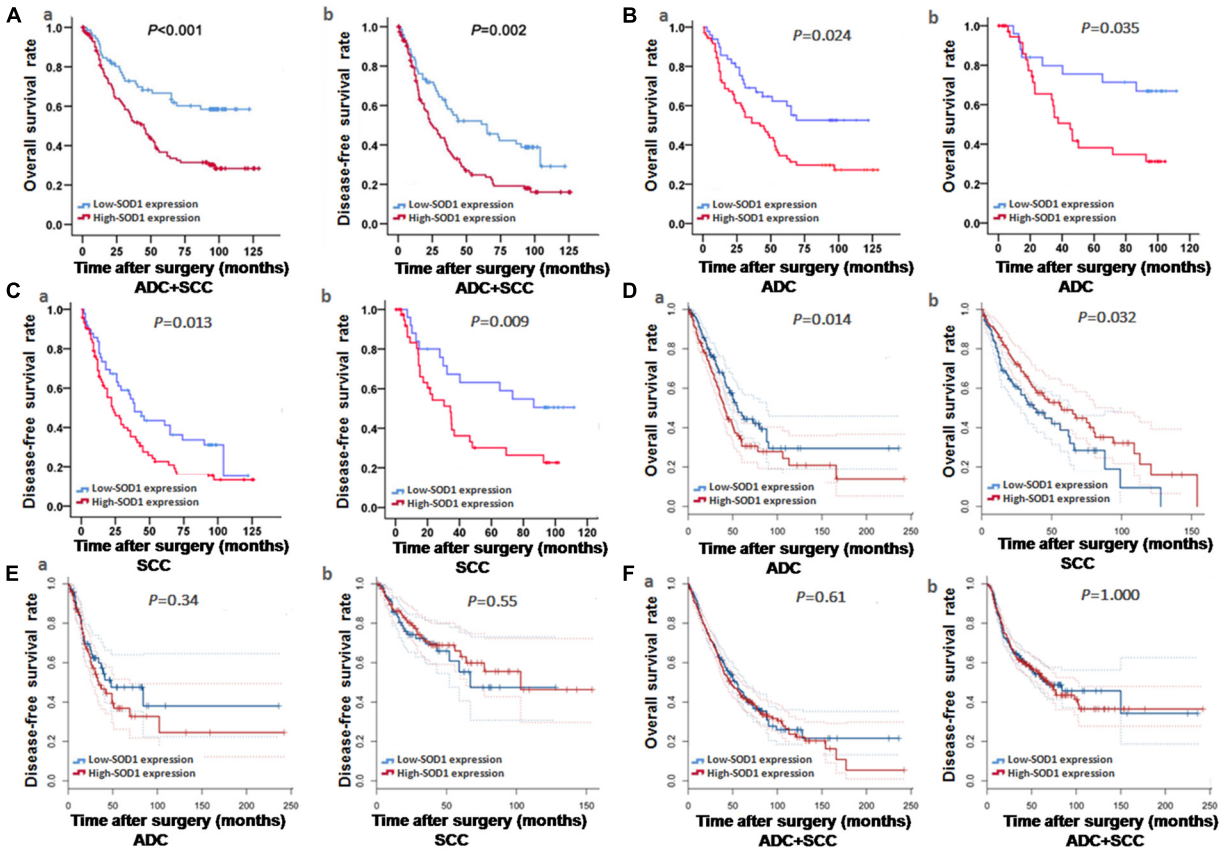


FIGURE 2 | Kaplan-Meier curve analysis of survival rates in patients with adenocarcinoma (ADC) and squamous cell carcinoma (SCC), based on experimental and bioinformatics data. **(A–C)** Kaplan-Meier curve analysis of Overall survival (OS) and disease-free survival (DFS) based on our exprimental data. OS and DFS of patients with ADC or SCC exhibiting high SOD1 protein expression were significantly poorer to those exhibiting low SOD1 protein levels. These results were consistent with subgroup analysis of OS and DFS in ADC and SCC. **(D–F)** Kaplan-Meier curve analysis of OS and DFS based on bioinformatics data. OS was significantly reduced in patients with high than in those with low SOD1 protein expression in ADC ($P = 0.014$ **D-a**), but the OS in SCC patients was positively correlated with SOD1 expression ($P = 0.032$, **D-b**); In addition, no significant differences between high and low SOD1 expression levels in ADC **(E-a)** and SCC **(E-b)**. **(F)**. No significant differences were also observed in total OS **(F-a)** and DFS **(F-b)** for NSCLC (including ADC and SCC). $P < 0.05$ was considered significant.

activity was significantly reduced in H460 cells co-transfected with SOD1 3′-UTR-WT and miR-409-3p mimic (**Figure 4I**). In contrast, luciferase activity was significantly increased by treatment with miR-409-3p inhibitor (**Figure 4J**); however, luciferase activity did not vary significantly in H460 cells which co-transfected with SOD1 3′-UTR-MUT and miR-409-3p mimic or miR-409-3p inhibitor (**Figures 4I,J**).

SOD1-Medoated Promotion of Cell Proliferation and Inhibition of Apoptosis Was Partly Modulated via miR-409-3p

To further investigate whether SOD1 promotes cell proliferation through miR-409-3p, we co-transfected miR-409-3p mimics and pLVX-puro-SOD1 into H1299 cells and compared cell proliferation and apoptosis with other transfected controls. The CCK-8 assay results revealed that transfection of the miR-409-3p mimic into H1299 cells reduced the proliferative capacity of cells, while co-transfected miR-409-3p mimics and pLVX-puro-SOD1 into H1299 cells led to a modest increase in proliferative capacity cpmpared with that in controls, This indicatied that miR-409-3p could significantly suppress the effects of SOD1 in promoting H1299 cell proliferation (**Figure 4K**). Flow cytometry demonstrated that overexpression of miR-409-3p promoted H1299 cell apoptosis. Furthermore, repression of apoptosis in response to SOD1 overexpression was relieved when cells were transfected with miR-409-3p mimic (**Figure 4L**), However, overexpression of miR-409-3p did not significantly reverse the effect of SOD1 in promoting cell cycle progression (**Figure 4M**).

SETDB1 May Contribute to the Relationship Between miR-409-3p and SOD1 in NSCLC

Finally, since bioinformatics analyses and our experimental results show discrepancies in the relationship between SOD1 and miR-409-3p (especially the correlation between SOD1 and miR-409-3p was not significant in ADC), we used another method, GSCA Lite software (Liu et al., 2018) to analyze the association between SOD1 and miR-409-3p and determine whether other

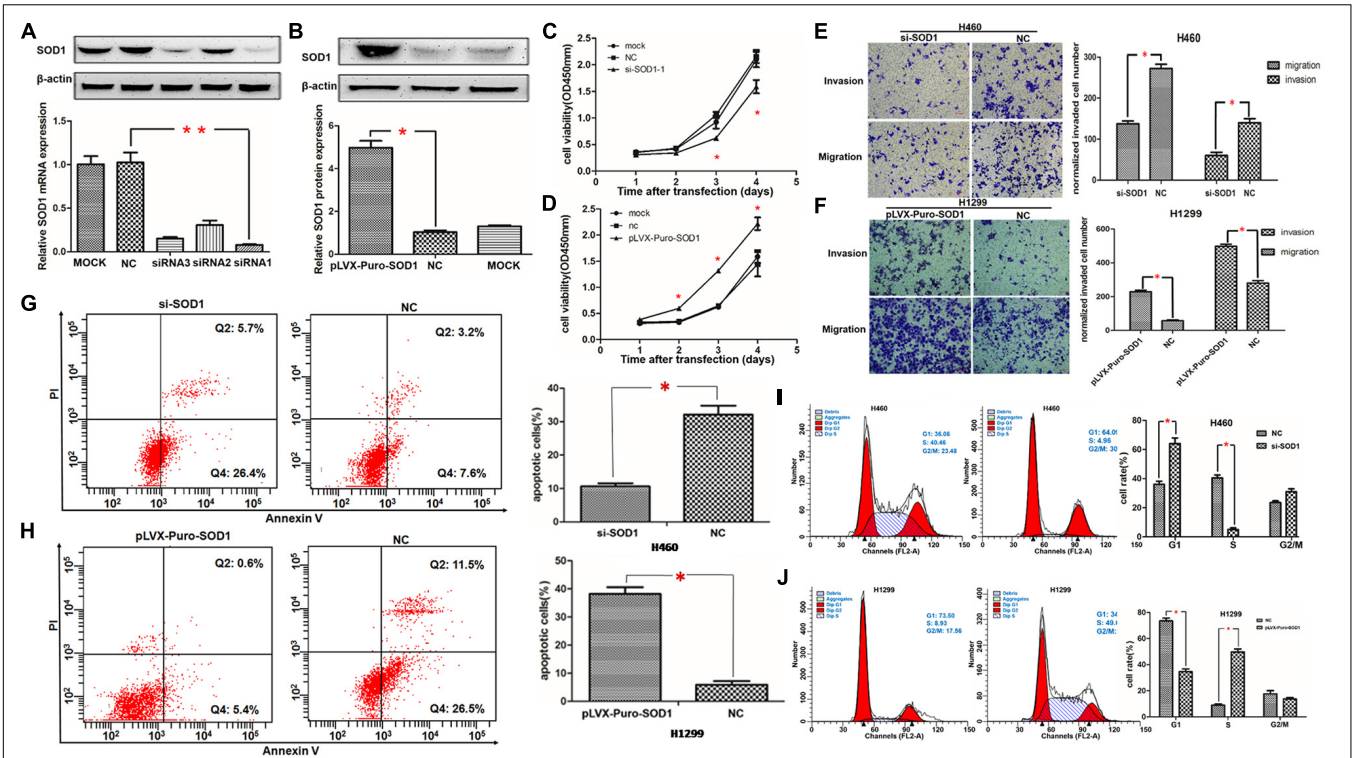


FIGURE 3 | Regulation of NSCLC cell proliferation, invasion, metastasis, apoptosis, and the cell cycle by SOD1. **(A,B)** SOD1 mRNA and protein expression levels were analyzed following transfection of SOD1 siRNA or pLVX-puro-SOD1 into H460 and H1299 cells. Representative western blot results are shown. **(C,D)** Proliferation of H460 and H1299 cells analyzed by CCK-8 assay after inhibition or overexpression of SOD1. **(E,F)** Evaluation of H460 and H1299 cell migration and invasion by transwell assay after inhibited or overexpressed of SOD1. **(G–J)** Measurement of apoptosis and the cell cycle by flow cytometryin cells with upregulated or downregulated SOD1 expression. Experiments were conducted using triplicate samples and each experiment was conducted three times; $^{*}P < 0.05$, $^{**}P < 0.01$.

factors are involved in the regulation of SOD1 expression by miR-409-3p. As shown in **Figure 5A**, 90 transcription factors related to lung cancer were identified by this analysis, and 12 were highly co-expressed with SOD1, of which only SETDB1 could be regulated by miR-409-3p. Further analysis indicated that SETDB1 was significantly repressed when miR-409-3p present (Pearson = −0.1, p = 0.01, **Figure 5B**) and positively correlated with SOD1 (Pearson = 0.1, p = 0.16, **Figure 5B**). Moreover, SOD1 function was regulated by both miR-409-3p and SETDB1 (**Figure 5C**), while SOD1 expression was inversely proportional with the degree of methylation in both ADC and SCC (**Figure 5D**). Overall, we hypothesize that SETDB1 may contribute to the regulatory relationship between miR-409-3p and SOD1 in NSCLC cells via a feedforward loop (**Figure 5E**).

DISCUSSION

The SOD1 gene maps to chromosome 21q22.11 and produces a 32 kDa homopolymer antioxidant enzyme. Numerous cancers express high level of SOD1, including lung adenocarcinoma, breast cancer, and leukemia, and its elevated expression is associated with poor survival (Huang et al., 2000; Somwar et al., 2011; Papa et al., 2014). In the present study, we confirmed that expression of SOD1 is markedly elevated in NSCLC by informatics analyses. Subsequently, we used qRT-PCR, western blot, and immunohistochemistry to verify these results in both cell lines and tissues from NSCLC patients. Our results were in accordance with the bioinformatics findings as well as those of previous research, implying that SOD1 has an oncogenic role in NSCLC.

Our prognostic analysis of 196 patients with NSCLC revealed that OS and DFS were significantly worse in the group with high SOD1 expression group than those in the group with low SOD1 expression. Subgroup analysis showed that this trend remained statistically significant in both ADC and SCC, while Cox regression analysis demonstrated that SOD1 is an independent prognostic factor for OS and DFS, further supporting an oncogenic role for SOD1 in NSCLC. Nevertheless, Cox regression analysis also demonstrated that TNM stage was an independent prognostic factor for OS and DFS in patients with NSCLC, and Kaplan-Meier analysis demonstrated that SOD1 expression was significantly negatively associated with OS and DFS, however, the bioinformatics analysis indicated that SOD1 levels did not increase with TNM stage. In particular, SOD1 levels were significantly reduced in stage IV SCC, this is contrary to our experimental conclusion and bioinformatics data in ADC, because we are unable to obtain the original data of patients with stage IV SCC in the database. Thus, we speculate this part of the patients may have heterogeneity

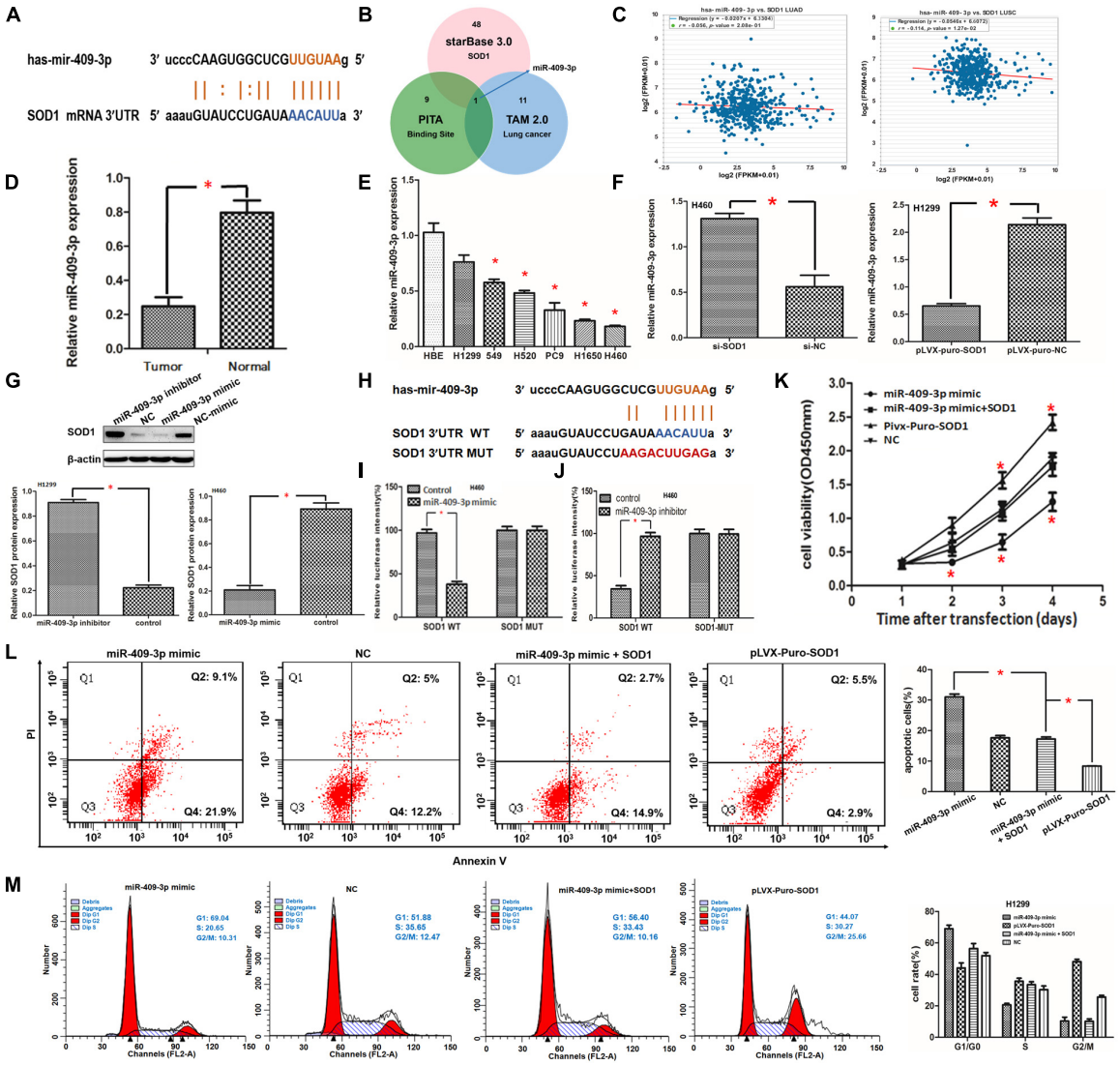


FIGURE 4 | (A,B) Superoxide dismutase 1 (*SOD1*) mRNA may be a target of miR-409-3p. Binding sites for miR-409-3p in the SOD1 promoter predicted using the bioinformatics website (starBasev3.0 including TAM and PITA) are shown. **(C)** Relationship between the expression of miR-409-3p and *SOD1* level in starBasev3.0. **(D,E)** qRT-PCR showing that miR-409-3p expression was inhibited in NSCLC tissue and cells. **(F)** Relative expression profiles of miR-409-3p in NSCLC cells with or without SOD1 inhibition. **(G)** Representative western blot assay and histogram demonstrating relative SOD1 expression in NSCLC cells transfected with miR-409-3p inhibitor or miR-409-3p mimic. **(H)** Wild-type (WT) or mutant miR-409-3p binding sequences in the SOD1 promoter were constructed using starBase v3.0. **(I,J)** Dual-luciferase reporter gene assay to determine luciferase activity from wild-type (WT) or mutant SOD1 promoter sequences 48 h after transfection. **(K,L)**The effects of SOD1 on proliferation and apoptosis of NSCLC cells was rescued by overexpression of miR-409-3p; however, the effect of SOD1 on cell cycle phase distribution was not significantly affected by miR-409-3p mimic **(M)**. Each experiment was conducted in triplicate; $^*P < 0.05$.

or data differences, which may lead to the negative effect on the prognosis. SOD1 probably be more meaningful to ADC. This requires further research on a large amount of samples. In addition, SOD1 was also differentially expressed in pan-cancer and normal tissues, indicating that it is tissue-specific and may play different or even opposite roles in different types of cancer. Although SOD1 shows highly expressed in many cancers and functions as an oncogene, it is still need to be furtherly studied whether SOD1 acts as the similar role in pancreatic cancer.

High SOD1 expression can promote tumor cell proliferation through multiple mechanisms. Glasauer et al. (2014) reported that SOD1 inhibition reduced antioxidant protein activity, leading to increased intracellular H_2O_2, which induced p38 MAPK-mediated A549 cell death, while inhibition of SOD1 in normal bronchial epithelial cells had no effect. Another group found that SOD1 inhibition increased reactive oxygen species (ROS) levels, resulting in increased DNA double strand breaks and leading to selective killing of RAD54B-deficient colorectal cancer cells (Sajesh et al., 2013). Li et al. designed

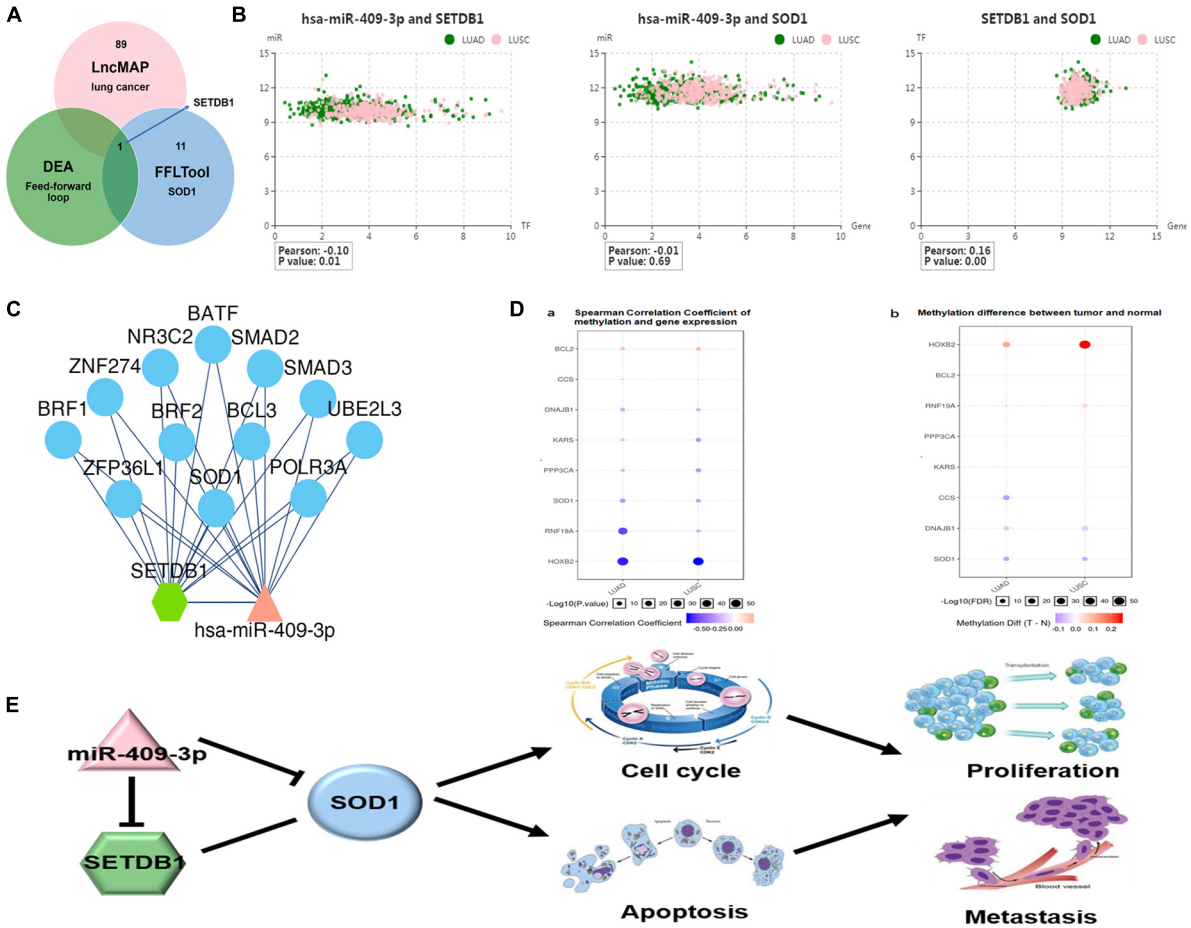


FIGURE 5 | (A) SET domain bifurcated histone lysine methyltransferase1 (SETDB1), a transcription factor identified using the GSCALite website, influences the relationship between miR-409-3p and SOD1 in NSCLC. **(B)** Relationship among SOD1, miR-409-3p, and SETDB1. **(C)** The regulatory networks of miR-409-3p and SETDB1 both include SOD1. **((D-a)** Spearman correlation coefficient analyses of methylation and gene expression in ADC and SCC. **(D-b)** Differences in methylation between tumors and normal tissue. **(E)** Proposed feedforward loop by which SOD1 exert its functions in NSCLC cells.

a small molecule inhibitor, which selectively represses SOD1, and promotes cancer cell apoptosis via regulation of the ROS signaling network (Li et al., 2019a). Chang et al. found that up-regulation of SOD2 could increase ROS levels and decrease SOD1 expression, leading to cyclin D1 up-regulation and G1-phase cell cycle arrest. Eventually, activated caspase-3 and subsequent apoptosis in H1299 cells (Chang et al., 2018). SOD1 deficiency was shown to leads to increased oxidative stress, which can induce DNA hypomethylation in prostate tissue (Bhusari et al., 2010). Therefore, we speculate that SOD1 expression in cancer cells promotes growth and inhibits apoptosis, primarily via its antioxidantive capacity (Somwar et al., 2011; Reddi and Culotta, 2013; Glasauer et al., 2014; Wu et al., 2019). Of course, Other regulatory mechanisms are also likely to be involved; for example, SOD1 induces DNA damage and triggers the apoptotic response by activating p53 (Barbosa et al., 2010). Our results demonstrate that down-regulation of SOD1 can significantly inhibit the proliferation, invasion, and migration of NSCLC, whereas SOD1 over-expression had opposite effects. Our flow cytometry analysis verifies that alterations in SOD1 expression can promote tumor cell proliferation and metastasis by influencing the cell cycle and apoptosis, consistent with previous research results (Glasauer et al., 2014; Chang et al., 2018).

Substantial progress has been made in determining the function of SOD1 in modulating tumor cell proliferation and metastasis, however, little is known about how and which transcription factors (including miRNAs or long non-coding RNAs) could regulate *SOD1* gene transcription or translation (Bao et al., 2019). Afonso et al. (2006) identified TNF-α as a transcription factor that inhibits SOD1 transcription, translation, and promoter activity through binding to a proximal promoter sequence 157 bp upstream of the *SOD1* transcription initiation site. Moreover, AUF-1 was shown to bind to the *SOD1* 3′ UTR and primarily promotes SOD1 protein translation, leading to increased SOD1 protein expression (Zhang et al., 2015). Other mechanisms also regulate SOD1, including oxidative stress and phosphorylation. Chang et al. (2018) reported that up-regulation of SOD2 can increase ROS levels and decrease SOD1 expression

in NSCLC H1299 cells. While, Tsang et al. (2018) found that the mechanistic target of rapamycin complex 1 (mTORC1) regulates SOD1 activity through reversible phosphorylation at S39 in yeast and T40 in humans in response to nutrient availability. Here, we used bioinformatics to analyze whether any NSCLC-related microRNAs could bind to the *SOD1* 3′ UTR, the result identified miR-409-3p as being associated with both NSCLC and SOD1, and able to bind to the 3′ UTR of *SOD1*, and showed that *SOD1* mRNA is a target of miR-409-3p. Based on the binding sequence, we designed a luciferase reporter assay and the results confirmed that miR-409-3p can bind directly to the *SOD1* 3′ UTR.

There is good evidence that miR-409-3p functions as a tumor suppressor in several cancers, for example, it is frequently downregulated in colorectal cancer, and acts as a tumor inhibitor by directly targeting the 3′ UTR of *NLK* (Liu et al., 2015). The expression of miR-409-3p is reduced in papillary thyroid carcinoma and may negatively regulate cell proliferation and cell cycle progression through repression of its target gene, cyclin D2, in these tumors (Zhao et al., 2018). Further, miR-409-3p expression is also decreased in tongue squamous cell carcinoma, and suppresses the proliferation, invasion, and migration of the tongue squamous cell carcinoma cells by targeting *RDX* (Chen and Dai, 2018). In the present study, we demonstrated that miR-409-3p is significantly downregulated in NSCLC cells and tissues, consistent with previous reports (Wan et al., 2014), In addition, we showed that miR-409-3p can negatively regulate SOD1 mRNA and protein levels in NSCLC cells. To further explore whether miR-409-3p can negatively regulate SOD1 functions in NSCLC cells, we designed a rescue experiment. We cotransfected miR-409-3p mimic and pLVX-puro-SOD1 into H1299 cells, and showed that overexpression of miR-409-3p could reverse the effects of SOD1 in promoting cell proliferation and repressing apoptosis. This consequence implied miR-409-3p can inhibit SOD1 activity and expressions in NSCLC. However, miR-409-3p did not inhibit the effect of SOD1 on cell cycle progression. We speculate that the effect of SOD1 on NSCLC cell cycle progression was not significantly affected by miR-409-3p, or that other regulatory factors may be involved in this regulatory process. However, this needs further experimental confirmation.

Bioinformatics analysis showed that miR-409-3p and *SOD1* expression levels were negatively correlated; but, the degree of correlation was modest, and was not significant in ADC. We believe that there may be other regulatory factors (such as transcription factors or lncRNAs) (Zhou et al., 2016) that coordinately regulate SOD1 expression with miR-409-3p. For example, increasing evidence has highlighted lncRNAs may play a critical role in tumorigenesis and prognosis of NSCLC (Sun et al., 2014; Zhou et al., 2015), and we also found that lncRNA-CBR3-AS1 exert oncogenic functions in NSCLC by targeting SOD1 (data not yet published) and this will be discussed in other articles. In this study, we assume a transcription factor could be affected by miR-409-3p or SOD1 via epigenetic mechanism, constituting a feedforward loop to regulate SOD1 function in NSCLC. Based on this hypothesis, we used GSCALite software to identify 90 lung cancer-related transcription factors related, among which only SETDB1 was associated with both SOD1 and miR-409-3p. We also found that expression of SETDB1 and SOD1 were highly correlated, and that SETDB1 levels were significantly negatively correlated with miR-409-3p. In addition, we found SOD1 expression can repress methylation in NSCLC, this consistent with findings that SOD1 knockdown induces oxidative stress and DNA methylation loss in prostate cancer (Bhusari et al., 2010), The underlying mechanism is likely that inhibition of reactive oxygen species (ROS) leads to an increase in SETDB1, as Park et al. (2019) reported that piperlongumine (Huang et al.) can downregulate SETDB1 to selectively kill breast cancer cells via accumulation of ROS, while the ROS inhibitor N-acetyl cysteine could recover the decrease in SETDB1 expression induced by PL. Consequently, we believe that SOD1 may maintain SETDB1 overexpression by down-regulating ROS levels. Our bioinformatics analyses also indicated that both SETDB1 and miR-409-3p are both related to SOD1 function, and SOD1 may associate hypomethylation of NSCLC. This suggests that miR-409-3p, SETDB1, and SOD1 form a feedforward loop to regulate cell cycle progression and apoptosis in NSCLC.

SETDB1 is an H3K9-specific histone methyltransferase that has been described as a repressive transcription factor involved in methylation and silencing of various genes (Fei et al., 2015; Riviere et al., 2016; Karanth et al., 2017). SETDB1 levels are greatly increased in numerous cancers, including lung cancer (Lafuente-Sanchis et al., 2016; Chen B. et al., 2018), and is an important epigenetic regulator involved in control of histone methylation in tumorigenesis, dysregulation of histone methylation, and aberrant miRNA profiling, contributing to tumorigenesis and progression (Chen Y. et al., 2018; Michalak et al., 2019). Moreover, DNA methylation can directly influence miRNA biogenesis (Glaich et al., 2019). Numerous microRNAs can be epigenetically silenced by DNA methylation of their promotor regions. For example, down-regulation of miR-145-5p can be attributed to methylation of the miR-145 promoter in brain metastasis of lung cancer (Donzelli et al., 2015), while Wang et al. found that H3K27me3 can bind to the miR-145 core promoter region to co-regulate *LASP1* expression in colorectal cancer cells (Wang et al., 2016). Histone methylation also mediates decreased expression of miR-449a via SUZ12 in NSCLC (You et al., 2015). Methylation of the promoter region inhibits the expression of miR-520c-3p, which functions as a novel tumor suppressor in lung adenocarcinoma (Li et al., 2019b). Furthermore, SETDB1 expression can also be regulated by microRNAs. Ectopic expression of miR-29 family molecules significantly decreases SETDB1 expression at both mRNA and protein levels by binding to the *SETDB1* 3′ UTR in NSCLC cells (Wong et al., 2016). Moreover, SETDB1 knockdown may suppress breast cancer progression, at least partly through miR-381-3p-related regulation, as SETDB1 is a verified target of miR-381-3p (Wu et al., 2018). In the present study, we speculate that SETDB1 may either promote the methylation of the miR-409-3p promoter region, or that miR-409-3p may bind to the *SETDB1* 3′ UTR, resulting in mutual regulation; however, this requires further experimental verification. Our research still has limitations, the interaction between SETDB1 and miR-409-3p, SETDB1 and SOD1 needs further study.

In summary, we have shown that SOD1 is overexpressed in NSCLC cell lines and tissues and that elevated SOD1 expression

can stimulate cell proliferation and metastasis, likely through promotion of cell cycle progression and suppression of apoptosis. Furthermore, we also shown that miR-409-3p can bind to the *SOD1* 3′ UTR to repressing SOD1 expression and reversing its oncogenic effects. Meanwhile, our results suggest that SETDB1 may contribute to the regulation of miR-409-3p and SOD1 expression and functions in NSCLC cells, and SOD1 functions is likely exert via a miR-409-3p/ SETDB1/SOD1 feedforward loop. Hence, our findings provide a new insight into regulatory networks in NSCLC.

DATA AVAILABILITY STATEMENT

The datasets used and analyzed in this study are all included in the article.

ETHICS STATEMENT

This study was approved by the Ethics Committee of The Third Affiliated Hospital of Harbin Medical University, Harbin, China. All patients gave their written consent before inclusion in the study.

AUTHOR CONTRIBUTIONS

SL and XX carried out the study and wrote the manuscript. SL, BL, JX, and SH helped most of the experiments. NZ, HW, and CG contributed to verify the results and revise the manuscript. HW and JL contributed to statistical analysis. All authors read the manuscript and approved the final version.

FUNDING

This study was funded by the scientific research program of Health and Family Planning Commission of Heilongjiang Province, China (number: 2017-12) and the Hai Yan Fund of The Third Affiliated Hospital of Harbin Medical University, Harbin, China (number: JJZD2016-06).

ACKNOWLEDGMENTS

We wish to thank all participants and investigators for their participation of the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.00213/full#supplementary-material

FIGURE S1 | SOD1 expression in normal tissues. SOD1 is expressed varied in different tissues and is Tissue-specific.

FIGURE S2 | Representative images of SOD1 immunohistochemical staining in NSCLC tissues. **(a)** Low expression in SCC histotype. **(b)** High expression in SCC histotype. **(c)** Low expression in ADC histotype. **(d)** High expression in ADC histotype. (×400 original magnification).

FIGURE S3 | Western blot analysis of SOD1 expression in NSCLC cell lines and a normal bronchial epithelial cell line. The histogramillustrates of overexpression of SOD1 protein in NSCLC tumor tissues relative to normal lung tissues. $^{*}P < 0.05$, $^{**}P < 0.01$.

FIGURE S4 | The histogramillustrates of expression of SOD1 protein levels were analyzed following transfection of SOD1 siRNA into H460 cells. $^{*}P < 0.05$.

FIGURE S5 | The histogramillustrates of relative SOD1 mRNA expression in NSCLC cells transfected with miR-409-3p inhibitor or miR-409-3p mimic. $^{*}P < 0.05$.

REFERENCES

Afonso, V., Santos, G., Collin, P., Khatib, A. M., Mitrovic, D. R., Lomri, N., et al. (2006). Tumor necrosis factor-alpha down-regulates human Cu/Zn superoxide dismutase 1 promoter via JNK/AP-1 signaling pathway. *Free Radic. Biol. Med.* 5, 709–721. doi: 10.1016/j.freeradbiomed.2006.05.014

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief Bioinform.* doi: 10.1093/bib/bbz118 [Online ahead of print],

Barbosa, L. F., Cerqueira, F. M., Macedo, A. F., Garcia, C. C., Angeli, J. P., Schumacher, R. I., et al. (2010). Increased SOD1 association with chromatin, DNA damage, p53 activation, and apoptosis in a cellular model of SOD1-linked ALS. *Biochim. Biophys. Acta* 1802, 462–471. doi: 10.1016/j.bbadis.2010.01.011

Bhusari, S. S., Dobosy, J. R., Fu, V., Almassi, N., Oberley, T., and Jarrard, D. F. (2010). Superoxide dismutase 1 knockdown induces oxidative stress and DNA methylation loss in the prostate. *Epigenetics* 5, 402–409. doi: 10.4161/epi.5.5.11853

Brasil, A. A., de Carvalho, M. D. C., and Gerhardt, E. (2019). Characterization of the activity, aggregation, and toxicity of heterodimers of WT and ALS-associated mutant Sod1. *Proc. Natl. Acad. Sci. U.S.A.* 116, 25991–26000. doi: 10.1073/pnas.1902483116

Chang, Y. C., Fong, Y., Tsai, E. M., Chang, Y. G., Chou, H. L., Wu, C. Y., et al. (2018). Exogenous C(8)-ceramide induces apoptosis by overproduction of ROS and the switch of superoxide dismutases SOD1 to SOD2 in human lung cancer cells. *Int. J. Mol. Sci.* 19:3010. doi: 10.3390/ijms19103010

Chen, H., and Dai, J. (2018). miR-409-3p suppresses the proliferation, invasion and migration of tongue squamous cell carcinoma via targeting RDX. *Oncol. Lett.* 16, 543–551. doi: 10.3892/ol.2018.8687

Chen, B., Wang, J., Wang, J., Wang, H., Gu, X., Tang, L., et al. (2018). A regulatory circuitry comprising TP53, miR-29 family, and SETDB1 in non-small cell lung cancer. *Biosci. Rep.* 38, doi: 10.1042/bsr20180678

Chen, Y., Liu, X., Li, Y., Quan, C., Zheng, L., and Huang, K. (2018). Lung cancer therapy targeting histone methylation: opportunities and challenges. *Comput. Struct. Biotechnol. J.* 16, 211–223. doi: 10.1016/j.csbj.2018.06.001

Donzelli, S., Mori, F., Bellissimo, T., Sacconi, A., Casini, B., Frixa, T., et al. (2015). Epigenetic silencing of miR-145-5p contributes to brain metastasis. *Oncotarget* 6, 35183–35201. doi: 10.18632/oncotarget.5930

Fei, Q., Shang, K., Zhang, J., Chuai, S., Kong, D., Zhou, T., et al. (2015). Histone methyltransferase SETDB1 regulates liver cancer cell growth through methylation of p53. *Nat. Commun.* 6:8651. doi: 10.1038/ncomms9651

Glaich, O., Parikh, S., Bell, R. E., Mekahel, K., Donyo, M., Leader, Y., et al. (2019). DNA methylation directs microRNA biogenesis in mammalian cells. *Nat. Commun.* 10, 5657.

Glasauer, A., Sena, L. A., Diebold, L. P., Mazar, A. P., and Chandel, N. S. (2014). Targeting SOD1 reduces experimental non-small-cell lung cancer. *J. Clin. Invest.* 124, 117–128. doi: 10.1172/jci71714

Huang, P., Feng, L., Oldham, E. A., Keating, M. J., and Plunkett, W. (2000). Superoxide dismutase as a target for the selective killing of cancer cells. *Nature* 407, 390–395. doi: 10.1038/35030140

Karanth, A. V., Maniswami, R. R., Prashanth, S., Govindaraj, H., Padmavathy, R., Jegatheesan, S. K., et al. (2017). Emerging role of SETDB1 as a therapeutic target. *Expert. Opin. Ther. Targets* 21, 319–331. doi: 10.1080/14728222.2017.1279604

Lafuente-Sanchis, A., Zuniga, A., Galbis, J. M., Cremades, A., Estors, M., Martinez-Hernandez, N. J., et al. (2016). Prognostic value of ERCC1, RRM1, BRCA1 and SETDB1 in early stage of non-small cell lung cancer. *Clin. Transl. Oncol.* 18, 798–804. doi: 10.1007/s12094-015-1440-6

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, X., Chen, Y., Zhao, J., Shi, J., Wang, M., Qiu, S., et al. (2019a). The specific inhibition of SOD1 selectively promotes apoptosis of cancer cells via regulation of the ROS signaling network. *Oxid Med. Cell Longev.* 2019:9706792. doi: 10.1155/2019/9706792

Li, X., Fu, Q., Li, H., Zhu, L., Chen, W., Ruan, T., et al. (2019b). MicroRNA-520c-3p functions as a novel tumor suppressor in lung adenocarcinoma. *Febs. J.* 286, 2737–2752. doi: 10.1111/febs.14835

Lin, J., Zahurak, M., Beer, T. M., Ryan, C. J., Wilding, G., Mathew, P., et al. (2013). A non-comparative randomized phase II study of 2 doses of ATN-224, a copper/zinc superoxide dismutase inhibitor, in patients with biochemically recurrent hormone-naive prostate cancer. *Urol. Oncol.* 31, 581–588. doi: 10.1016/j.urolonc.2011.04.009

Liu, C. J., Hu, F. F., Xia, M. X., Han, L., Zhang, Q., and Guo, A. Y. (2018). GSCALite: a web server for gene set cancer analysis. *Bioinformatics* 34, 3771–3772. doi: 10.1093/bioinformatics/bty411

Liu, M., Xu, A., Yuan, X., Zhang, Q., Fang, T., Wang, W., et al. (2015). Downregulation of microRNA-409-3p promotes aggressiveness and metastasis in colorectal cancer: an indication for personalized medicine. *J. Transl. Med.* 13:195. doi: 10.1186/s12967-015-0533-x

Lowndes, S. A., Adams, A., Timms, A., Fisher, N., Smythe, J., Watt, S. M., et al. (2008). P hase I study of copper-binding agent ATN-224 in patients with advanced solid tumors. *Clin. Cancer Res.* 14, 7526–7534. doi: 10.1158/1078-0432.CCR-08-0315

Lu, J., and Clark, A. G. (2012). Impact of microRNA regulation on variation in human gene expression. *Genome Res.* 22, 1243–1254. doi: 10.1101/gr.132514.111

Michalak, E. M., Burr, M. L., and Bannister, A. J. (2019). The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat. Rev. Mol. Cell Biol.* 20, 573–589. doi: 10.1038/s41580-019-0143-1

Papa, L., Hahn, M., Marsh, E. L., Evans, B. S., and Germain, D. (2014). SOD2 to SOD1 switch in breast cancer. *J. Biol. Chem.* 289, 5412–5416. doi: 10.1074/jbc.C113.526475

Park, J. A., Na, H. H., Jin, H. O., and Kim, K. C. (2019). Increased expression of fosb through reactive oxygen species accumulation functions as pro-apoptotic protein in Piperlongumine Treated MCF7 breast cancer cells. *Mol. Cells* 42, 884–892. doi: 10.14348/molcells.2019.0088

Qu, R., Chen, X., and Zhang, C. (2018). LncRNA ZEB1-AS1/miR-409-3p/ZEB1 feedback loop is involved in the progression of non-small cell lung cancer. *Biochem. Biophys. Res. Commun.* 507, 450–456. doi: 10.1016/j.bbrc.2018.11.059

Reddi, A. R., and Culotta, V. C. (2013). SOD1 integrates signals from oxygen and glucose to repress respiration. *Cell* 152, 224–235. doi: 10.1016/j.cell.2012.11.046

Riviere, L., Gerossier, L., Hantz, O., and Neuveut, C. (2016). [Hepatitis B virus and chromatin remodeling: HBx counteracts SETDB1/HP1/H3K9me3 transcriptional silencing]. *Med. Sci.* 32, 455–458. doi: 10.1051/medsci/20163205010

Sajesh, B. V., Bailey, M., Lichtensztejn, Z., Hieter, P., and McManus, K. J. (2013). Synthetic lethal targeting of superoxide dismutase 1 selectively kills RAD54B-deficient colorectal cancer cells. *Genetics* 195, 757–767. doi: 10.1534/genetics.113.156836

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA Cancer J. Clin.* 68, 7–30. doi: 10.3322/caac.21442

Somwar, R., Erdjument-Bromage, H., Larsson, E., Shum, D., Lockwood, W. W., Yang, G., et al. (2011). Superoxide dismutase 1 (SOD1) is a target for a small molecule identified in a screen for inhibitors of the growth of lung adenocarcinoma cell lines. *Proc. Natl. Acad. Sci. U.S.A.* 108, 16375–16380. doi: 10.1073/pnas.1113554108

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, 2074–2081. doi: 10.1039/c3mb70608g

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, W98–W102. doi: 10.1093/nar/gkx247

Tsang, C. K., Liu, Y., Thomas, J., Zhang, Y., and Zheng, X. F. (2014). Superoxide dismutase 1 acts as a nuclear transcription factor to regulate oxidative stress resistance. *Nat. Commun.* 5:3446. doi: 10.1038/ncomms4446

Tsang, C. K., Chen, M., Cheng, X., Qi, Y., Chen, Y., Das, I., et al. (2018). SOD1 phosphorylation by mTORC1 couples nutrient sensing and redox regulation. *Mol. Cell* 70, 502.e8–515e8. doi: 10.1016/j.molcel.2018.03.029

Wan, L., Zhu, L., Xu, J., Lu, B., Yang, Y., Liu, F., et al. (2014). MicroRNA-409-3p functions as a tumor suppressor in human lung adenocarcinoma by targeting c-Met. *Cell Physiol. Biochem.* 34, 1273–1290. doi: 10.1159/000366337

Wang, W., Ji, G., Xiao, X., Chen, X., Qin, W. W., Yang, F., et al. (2016). Epigenetically regulated miR-145 suppresses colon cancer invasion and metastasis by targeting LASP1. *Oncotarget* 7, 68674–68687. doi: 10.18632/oncotarget.11919

Watanabe, K., Shibuya, S., Ozawa, Y., Nojiri, H., Izuo, N., Yokote, K., et al. (2014). Superoxide dismutase 1 loss disturbs intracellular redox signaling, resulting in global age-related pathological changes. *BioMed Res. Int.* 2014, 140165–140165. doi: 10.1155/2014/140165

Wong, C. M., Wei, L., Law, C. T., Ho, D. W., Tsang, F. H., Au, S. L., et al. (2016). Up-regulation of histone methyltransferase SETDB1 by multiple mechanisms in hepatocellular carcinoma promotes cancer metastasis. *Hepatology* 63, 474–487. doi: 10.1002/hep.28304

Wu, K. L., Tsai, Y. M., Lien, C. T., Kuo, P. L., and Hung, A. J. (2019). The roles of microRNA in lung cancer. *Int. J. Mol. Sci.* 20:1611. doi: 10.3390/ijms20071611

Wu, M., Fan, B., Guo, Q., Li, Y., Chen, R., Lv, N., et al. (2018). Knockdown of SETDB1 inhibits breast cancer progression by miR-381-3p-related regulation. *Biol. Res.* 51:39.

You, J., Zhang, Y., Li, Y., Fang, N., Liu, B., Zu, L., et al. (2015). MiR-449a suppresses cell invasion by inhibiting MAP2K1 in non-small cell lung cancer. *Am. J. Cancer Res.* 5, 2730–2744.

Zhang, S., Xue, J., Zheng, J., Wang, S., Zhou, J., Jiao, Y., et al. (2015). The superoxide dismutase 1 3'UTR maintains high expression of the SOD1 gene in cancer cells: the involvement of the RNA-binding protein AUF-1. *Free Radic. Biol. Med.* . 85, 33–44. doi: 10.1016/j.freeradbiomed.2015.04.012

Zhang, Y., Unnikrishnan, A., Deepa, S. S., Liu, Y., Li, Y., Ikeno, Y., et al. (2017). A new role for oxidative stress in aging: the accelerated aging phenotype in Sod1(-/-)(-) mice is correlated to increased cellular senescence. *Redox Biol.* 11, 30–37. doi: 10.1016/j.redox.2016.10.014

Zhao, Z., Yang, F., Liu, Y., Fu, K., and Jing, S. (2018). MicroRNA-409-3p suppresses cell proliferation and cell cycle progression by targeting cyclin D2 in papillary thyroid carcinoma. *Oncol, Lett.* 16, 5237–5242. doi: 10.3892/ol.2018.9246

Zhou, M., Guo, M., He, D., Wang, X., Cui, Y., Yang, H., et al. (2015). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J. Transl. Med.* 13:231. doi: 10.1186/s12967-015-0556-3

Zhou, M., Xu, W., Yue, X., Zhao, H., Wang, Z., Shi, H., et al. (2016). Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma. *Oncotarget* 7, 29720–29738. doi: 10.18632/oncotarget.8825

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 © 2020 Liu, Li, Xu, Hu, Zhan, Wang, Gao, Li 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 Genetics

ORIGINAL RESEARCH
published: 13 May 2020
doi: 10.3389/fgene.2020.00427

Bioinformatics Analysis Reveals Biomarkers With Cancer Stem Cell Characteristics in Lung Squamous Cell Carcinoma

*Yi Liao†, Hua Xiao†, Mengqing Cheng and Xianming Fan**

Department of Respiratory and Critical Care Medicine II, The Affiliated Hospital of Southwest Medical University, Luzhou, China

Background: Tumor stem cells play important roles in the survival, proliferation, metastasis and recurrence of tumors. We aimed to identify new prognostic biomarkers for lung squamous cell carcinoma (LUSC) based on the cancer stem cell theory.

Methods: RNA-seq data and relevant clinical information were downloaded from The Cancer Genome Atlas (TCGA) database. Weighted gene coexpression network analysis (WGCNA) was applied to identify significant modules and hub genes, and prognostic signatures were constructed with the prognostic hub genes.

Results: LUSC patients in the TCGA database have higher mRNA expression-based stemness index (mRNAsi) in tumor tissue than in adjacent normal tissue. In addition, some clinical features and outcomes were highly correlated with the mRNAsi. WGCNA revealed that the pink and yellow modules were the most significant modules related to the mRNAsi; the top 10 hub genes in the pink module were enriched mostly in epidermal development, the secretory granule membrane, receptor regulator activity and the cytokine-cytokine receptor interaction. The protein–protein interaction (PPI) network revealed that the top 10 hub genes were significantly correlated with each other at the transcriptional level. In addition, the top 10 hub genes were all highly expressed in LUSC, and some were differentially expressed in different TNM stages. Regarding the survival analysis, the nomogram of a prognostic signature with three hub genes showed high predictive value.

Conclusion: mRNAsi-related hub genes could be a potential biomarker of LUSC.

Keywords: lung squamous cell carcinoma, cancer cell stemness, prognosis, WGCNA, TCGA

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Firoz Ahmed,
Jeddah University, Saudi Arabia
Alfred Grant Schissler,
University of Nevada, Reno, United States

***Correspondence:**
Xianming Fan
fxm129120@sina.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *17 January 2020*
Accepted: *06 April 2020*
Published: *13 May 2020*

Citation:
Liao Y, Xiao H, Cheng M and Fan X (2020) Bioinformatics Analysis Reveals Biomarkers With Cancer Stem Cell Characteristics in Lung Squamous Cell Carcinoma. Front. Genet. 11:427. doi: 10.3389/fgene.2020.00427

INTRODUCTION

Lung cancer ranks first in morbidity (11.6%) and mortality (18.4%) according to the latest worldwide survey of 20 regions from five continents and is the leading male cancer in both developed and developing countries (Bray et al., 2018) among all cancers. In February 2018, the latest national cancer statistics released by China's National Cancer Center revealed the same grim situation: lung cancer is still the most common malignant tumor in China in terms of morbidity

and mortality. Lung cancer is also the leading cause of death from malignant tumors in all regions of China (Chen et al., 2018). According to different histopathological characteristics, lung cancer is divided into non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), and lung squamous cell carcinoma (LUSC) is one of most common subtypes of NSCLC (Barlesi et al., 2016).

Research on the pathogenesis and pathological mechanism of NSCLC is still controversial; some studies (Friedmann-Morvinski and Verma, 2014; Leon et al., 2016; Shibue and Weinberg, 2017) have shown that tumor stem cells play important roles in tumor survival, proliferation, metastasis and recurrence. For example, a recent study used bioinformatics methods found that *FOXM1* and *MYBL2*, which are involved in the process of cell proliferation, can be used as potential biomarkers and therapeutic targets of NSCLC (Ahmed, 2019). This theory provides a new direction and idea for us to understand the origin and nature of the tumor and clinical treatment. In essence, tumor stem cells maintain the viability of tumor cell populations through self-renewal and infinite proliferation. The movement and migration capacities of tumor stem cells make the metastasis of tumor cells possible. Tumor stem cells can remain dormant for a long time and express a variety of resistant molecules but are not sensitive to the external physical and chemical factors that kill tumor cells. Therefore, tumor stem cells provide a new direction and visual perspective for us to re-understand the origin and nature of tumors, as well as clinical tumor therapy.

In addition, based on the theory of tumor stem cells, some scholars have introduced a new concept — stemness indices (Malta et al., 2018). The expression profile and methylation data of different tumor samples were collected from The Cancer Genome Atlas (TCGA) and other public databases. An innovative one-class logistic regression machine learning algorithm (OCLR) (Sokolov et al., 2016) was used to extract transcriptomic and epigenetic feature sets derived from non-transformed pluripotent stem cells and their differentiated progeny. Two independent stemness indices, the mRNA expression-based stemness index (mRNAsi) and the epigenetically regulated-mRNAsi (EREG-mRNAsi), were calculated. Among them, the index range was 0–1; the closer the value was to 1, the stronger the stem cell characteristics of tumor cells.

With the continuous development of high-throughput sequencing technology, it is convenient to explore the occurrence and development of tumors at the genetic level and to identify possible therapeutic targets (Goldfeder et al., 2017). Traditional methods use differential expression detection to identify potential biomarkers but may miss useful genes. Therefore, weighted gene coexpression network analysis (WGCNA) (Langfelder and Horvath, 2008) was adopted in the current study. The WGCNA was based on two hypotheses: (1) genes with similar expression patterns may share common regulatory networks and/or functional correlations or be involved in the same pathway; and (2) the gene network conforms to scale-free distribution. Based on these two hypotheses, the gene network can be divided into different modules according to expression similarity, and hub genes can be identified.

The purpose of this study was to obtain modules that are closely related to stem cell characteristics and to further identify the hub genes located in the regulatory center with the help of high-throughput sequencing data from a public database and the WGCNA method. We also determined whether these genes have a clear effect on prognosis.

MATERIALS AND METHODS

Data Processing

The flow diagram of our study was shown in **Supplementary Figure S1**. Level 3 RNA-seq data (HTSEQ-FPKM-UQ) and clinical information were downloaded from the TCGA website[1]. Ensembl IDs were converted to gene names via the Ensembl database[2], and log2 processing of the data was performed. If a gene had multiple expression values, they were averaged. Each sample from the TCGA and its corresponding mRNAsi and EREG-mRNAsi data were downloaded from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5902191/. Any samples that were missing the stemness index or clinical information were excluded.

Correlation of the mRNAsi and Clinical Characteristics

Differences in the mRNAsi between normal and LUSC tissues were compared using the unpaired *t*-test. Relapsed and non-relapsed patients who did not receive adjuvant therapy were also compared based on their mRNAsi. One-way ANOVA was used to compare significant differences in the mRNAsi between groups of variables. GraphPad Prism version 7 (64 bit) was used to perform the above analysis. To compare differences in prognosis, two indicators were evaluated: overall survival (OS) and progression-free survival (PFS). OS was defined as the time between the diagnosis of a tumor and death from any cause. PFS was defined as the time between the diagnosis of a tumor and the time to progression (in any form) or death from any cause. X-tile software (version 3.6.1) (Camp et al., 2004) was used to determine the best cut-off value in the survival data. The working principle of this software is to group different values as truncation values for the statistical test, and the result with the smallest p value is considered the best truncation value. Kaplan–Meier analysis was performed, and the *p*-value for two groups was calculated by the log-rank test with the survminer package in R software (v 3.6.0). $P < 0.05$ was considered a significant difference.

Differentially Expressed Genes (DEGs)

The limma package (Ritchie et al., 2015) was used to identify DEGs between LUSC and normal tissues. The inclusion criteria for DEGs were log2-fold change (FC) > 1 and adjusted $P < 0.05$.

[1]https://cancergenome.nih.gov/

[2]http://asia.ensembl.org/index.html

Weighted Gene Coexpression Network Analysis

Construction of a Coexpression Network

We used the DEGs obtained in the previous step to construct coexpression networks with the WGCNA package (Langfelder and Horvath, 2008) in R software (v3.6.0). The *goodSamplesGenes* function was used to determine whether the sample data were complete. It was also used to perform sample clustering to identify and remove outliers. Pearson correlation coefficients between each group of genes were also calculated, and their absolute values were used to construct the gene expression similarity matrix according to the following formula: $a_{ij} = |\mathrm{cor}(x_i, x_j)|^{\beta}$, where x_i and x_j represent nodes i and j of the network, respectively. A β value was selected to build the proximity matrix so that gene distribution conformed to a scale-free network based on connectivity. The adjacency matrix and topological overlap matrix (TOM) were constructed after obtaining the β value. The TOM obtained was then clustered by dissimilarity between genes, and the trees were then cut into different modules by the dynamic shear method (the minimum number of genes in the module was 50). Some modules were combined according to the correlation coefficient.

Identification of Significant Modules

We selected the hierarchical clustering module that was the most closely related to the mRNAsi and EREG-mRNAsi for further analysis. Genetic significance (GS), module significance (MS) and the module eigengene (ME) were also calculated. GS was defined as the level of correlation between gene expression and the mRNAsi and EREG-mRNAsi. The calculation method used was the log10 transformation of the p value in linear regression. It represents the relevance of each gene in the module to characteristics. MS was defined as the average of significance of all genes in the module. We merged similar modules using a cut-off value 0.55, and then the modules that had the largest MS were considered the most sample trait-related modules. ME was defined as the first principal component obtained by principal component analysis of the gene expression matrix of each module. Among all the modules, the module with the highest MS was considered to be related to the mRNAsi and EREG-mRNAsi and was selected for further research.

Identification of Hub Genes

The GS and module membership (MM, correlation between the module's own genes and gene expression profiles) of each gene were calculated after defining significant modules. The stronger the correlation was between the genes and significant modules, the stronger their relation to the stemness indices. Therefore, the inclusion criteria for a hub gene were set as follows: MM > 0.6 and cor. gene GS > 0.4.

Functional Enrichment Analysis

The clusterProfiler package (Yu et al., 2012) was used to perform functional enrichment for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses of the selected module. GO analysis consists of three terms: biological process (BP), cellular component (CC), and molecular function (MF). An adjusted $P < 0.05$ was used as the threshold.

Relationships and Interactions Among Hub Genes

STRING (version 11.0)[3] is an online database that can be used to study and visualize the network of interactions among proteins. Coexpression relationships among the hub genes were calculated based on gene expression levels to determine their strength at the transcriptional level. The Pearson correlation between genes was calculated using the R corrplot package[4], and the correlation matrix was visualized.

Validation of Hub Genes

GEPIA[5] is an online site that allows differential expression profiling, pathological staging, and patient survival analysis of tumors and normal tissues in the TCGA database. GEPIA data are derived not only from the TCGA database but also from the sequencing data of normal tissues in the Genotype-Tissue Expression (GTEx) project[6] (Tang et al., 2017), which compensates for the shortage of normal tissue samples in the TCGA database. Therefore, this database was used to verify whether the expression of the hub gene was higher in tumor tissue than in normal tissue and whether the hub gene was differentially expressed in different TNM stages.

Survival Analysis

Establishment of the Prognostic Signature

The relationship between each hub gene's expression level and OS was assessed by univariate Cox regression analysis, and hub genes with $P < 0.05$ were entered into the multivariate Cox regression process using the Akaike information criterion (AIC) (Yamaoka et al., 1978). A risk score formula was created using the corresponding data obtained through multivariate Cox proportional hazards regression analyses with the hub genes whose *p*-value was <0.05. In Equation 1, n denotes the number of prognostic hub genes, Gi represents the expression value of the ith hub genes, and weight i denotes the coefficient of each significant hub gene. Patients were divided into high-risk (>median risk score) and low-risk (<median risk score) groups according to the median risk score. The Kaplan–Meier method was used to estimate the survival outcomes of the high- and low-risk patients, and differences in OS were evaluated with the log-rank test. We generated a time-dependent receiver operating characteristic (ROC) curve to verify the accuracy of this signature. The classic ROC curve analysis method assumes that individual events and outcomes are fixed over time, but in practice, both disease status and outcomes change over time. Moreover, the traditional ROC curve cannot be used to analyze survival data; therefore, we adopted a time-dependent

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

[4]https://cran.r-project.org/web/packages/corrplot/index.html

[5]http://gepia.cancer-pku.cn/

[6]https://www.gtexportal.org/

ROC curve (Kamarudin et al., 2017). $P < 0.05$ was considered statistically significant.

$$Score = \sum_{i=1}^{n} G_i * weight_i \quad (1)$$

Construction and Assessment of the Nomogram

To lessen the influence of confounding factors on the relationship between gene expression and prognosis as much as possible, univariate and multivariate Cox regression analyses were performed to assess differences in clinical characteristics and risk scores. $P < 0.05$ was considered statistically significant. The nomograms of the 1-, 3-, and 5-year survival rates were constructed using the rms package in R software to visualize the prediction results[7]. The predictive ability of the nomogram was evaluated by the AUC of the ROC curve and the calibration curves for 1, 3, and 5 years.

RESULTS

Correlation of the mRNAsi and Clinical Characteristics in Patients With LUSC

Ninety-two samples without enough clinical information and 12 samples without mRNAsi information from the TCGA database were excluded. As shown in **Figure 1A**, there was a significant difference between the mRNAsi of LUSC and normal tissues. The mRNAsi of tumor tissues was higher than that of normal tissues. Significant differences in T stage (**Figure 1D**) and N stage (**Figure 1E**) were also observed in addition to AJCC stage (**Figure 1F**). However, there was no significant difference in the mRNAsi based on the treatment effect (**Figure 1B**) or whether the patient had relapsed (**Figure 1C**). LUSC patients with a high mRNAsi had significantly worse OS and PFS rates than those with a low mRNAsi (**Figures 1G,I**).

Screening of DEGs

There was a significant difference between the mRNAsi in normal tissues and LUSC tissues; therefore, we aimed to identify DEGs based on the comparison between the two. After log2 processing of the data, we found a total of 6122 DEGs, including 3427 upregulated genes and 2695 downregulated genes. The volcano map is shown in **Figure 1H**.

WGCNA: Identification of the Most Significant Modules and Genes

The results of WGCNA was shown in **Supplementary Table S1**. All DEGs were included in the coexpression network after excluding 48 outlier samples (**Figure 2A**). β = 3 met the soft-threshold parameter of the construction requirements for scale-free distribution, and the curve reached $R^2 = 0.925$. MEDissThres was set as 0.55 to merge the similar modules, and 13 modules were ultimately obtained (**Figure 2B**). After the modules were evaluated for their associations with the traits of LUSC and the patient's mRNAsi and EREG-mRNAsi, the pink ($R^2 = 0.68$, $P = 6e{-}60$) module was found to be positively correlated with the mRNAsi of LUSC patients (**Figure 2C**), while the yellow module ($R^2 = -0.76$, $P = 5e{-}58$) was found to be negatively correlated with the mRNAsi of LUSC patients (**Figure 2C**). In addition, the genes in the pink (cor = 0.71, $P < 1e{-}200$) and yellow (cor = 0.74, $P < 1e{-}200$) modules were characterized by high GS and MM based on an intramodular analysis (**Figures 2D,E**). Therefore, we chose the pink module as the most significant module for subsequent research because it showed the highest positive correlation. Based on the threshold for key genes (MM > 0.6 and cor GS > 0.4), we ultimately obtained 10 hub genes.

Functional Enrichment Analysis

For the two modules that were most closely related to the mRNAsi, GO and KEGG pathway enrichment analyses were performed, and the top 5 enriched results are presented in **Figure 3**. The pink module, which exhibited the strongest positive correlation with the mRNAsi, is highly enriched in epidermal development, the secretory granule membrane, receptor regulator activity and the cytokine–cytokine receptor interaction, while the yellow module, which exhibited the strongest negative correlation with the mRNAsi, is highly enriched in organelle fission, the chromosomal region, ion-gated channel activity and the cell cycle.

Protein-Protein Interaction (PPI) Network and Hub Gene Validation

The PPI network, consisting of the top 10 hub genes, was constructed using the STRING database. In total, 10 nodes and 33 edges were included in this PPI network (**Figure 4A**), with an average node degree of 6.6 and strong correlations. The 10 hub genes were also significantly correlated with each other at the transcriptional level (**Figure 4B**). The expression levels of the top 10 hub genes were higher in tumor tissue than in normal tissue (**Figure 5**). However, only 5 hub genes were differentially expressed in different TNM stages (**Figure 6**): *CDCA5, CENPA, NCAPH, SPAG5,* and *TIMELESS*.

Survival Analysis

Establishment of the Prognostic Signature

The prognostic signature consists of three hub genes (**Figure 7A**) that were incorporated into the multivariate Cox proportional hazards regression analysis to obtain the coefficients of the three hub genes that were used in Equation 1 to calculate the risk scores (**Table 1**). The risk score was calculated as follows: (−0.630* expression level of BUB1B) + (−0.652* expression level of CENPA) + (2.163* expression level of NCAPH) (**Table 1**). The prognosis of high-risk patients was significantly worse than that of low-risk patients (**Figure 7B**). The AUC values of 1-, 3-, and 5-year OS were 0.680, 0.704, and 0.674, respectively (**Figure 7C**).

Construction and Assessment of the Nomogram

The nomogram included the independent risk factors identified by the univariate and multivariate Cox regression analyses (**Table 2**): race (Caucasian vs. non-Caucasian), AJCC stage (stages I-II vs. stages III-IV), pharmacotherapy (no vs. yes) and risk

[7]http://CRAN.Rproject.org/packagerms

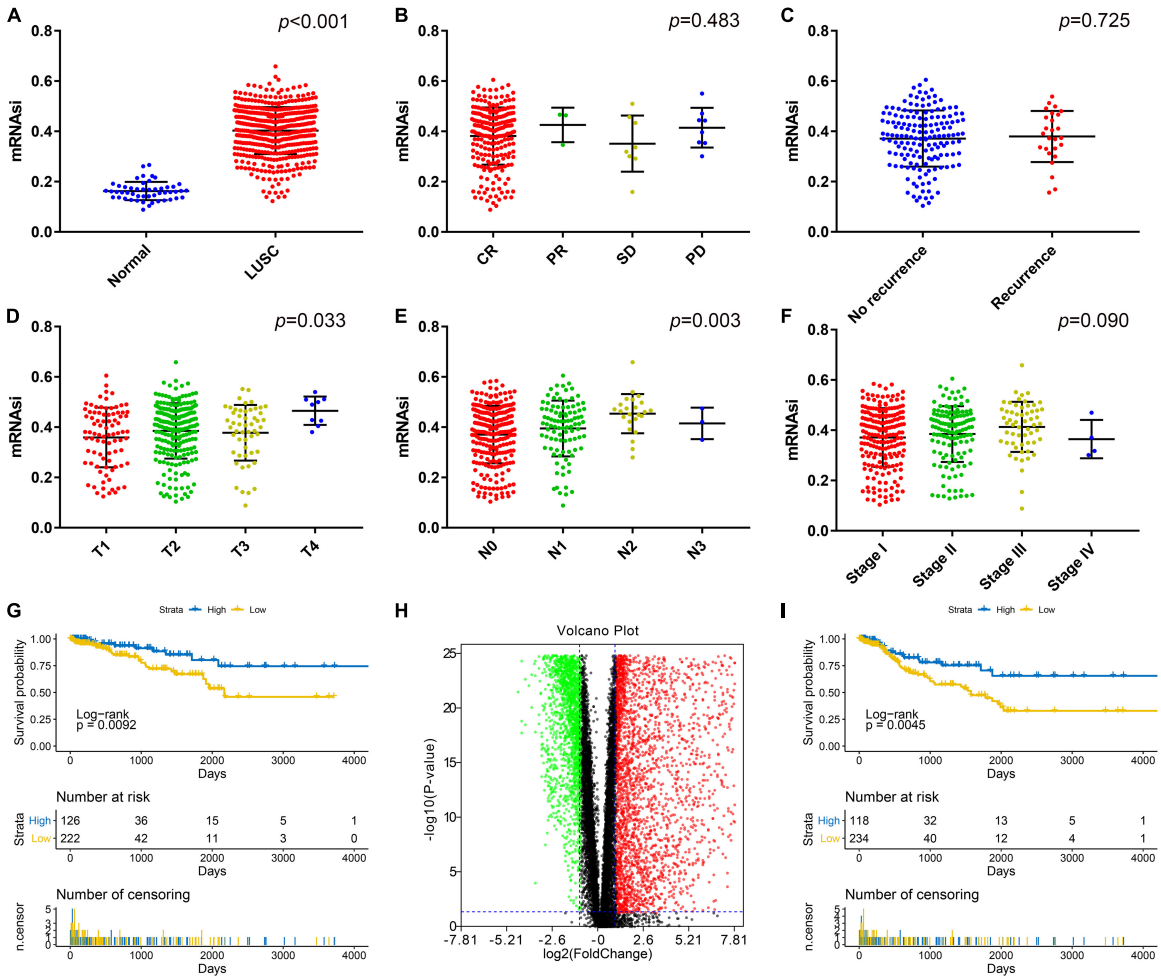


FIGURE 1 | (A) Differences in the mRNAsi between normal (50 samples) and LUSC (487 samples) tissues. **(B)** Differences in the mRNAsi based on the treatment effect: CR (234 samples), PR (3 samples), SD (8 samples), and PD (8 samples). **(C)** Differences in the mRNAsi between LUAD patients without (159 samples) and with (26 samples) recurrence after primary treatment without adjuvant therapy. **(D)** Comparison of the mRNAsi between four different T stages (T1, 86 samples; T2, 221 samples; T3, 52 samples; and T4, 9 samples). **(E)** Comparison of the mRNAsi between four different N stages (N0, 254 samples; N1, 99 samples; N2, 24 samples; and N3, 3 samples). **(F)** Comparison of the mRNAsi between four different AJCC stages (stage I; 201 samples, stage II; 132 samples, stage III; 58 samples, and stage IV; 4 samples). **(G)** Kaplan–Meier survival curves show that the low mRNAsi group had a better OS rate than the high mRNAsi group. **(H)** Volcano map of DEGs: green indicates downregulated genes, and red indicates upregulated genes. **(I)** Kaplan–Meier curves show that the low mRNAsi group had a better PFS rate than the high mRNAsi group. LUSC, lung squamous cell carcinoma; mRNAsi, mRNA expression-based stemness index; AJCC, American Joint Committee on Cancer; OS, overall survival; PFS, progression-free survival. CR, complete response, PR, partial response, SD, stable disease, PD, progressive disease.

score (low vs. high). The nomogram was constructed with these risk factors (**Figure 8A**). The AUC values for 1-, 3-, and 5-year OS were 0.754, 0.876 and 0.836 (**Figure 8B**), respectively. The calibration curve also demonstrated good capacity for the nomogram to predict 1- (**Figure 8C**), 3- (**Figure 8D**), and 5- (**Figure 8E**) year OS. The code for R software in this study can be obtained from **Supplementary Data Sheet S1**.

DISCUSSION

LUSC is associated with extremely high mortality and morbidity, but its pathogenesis is still unclear. However, an increasing number of studies have found that cancer stem cells (CSCs) play an important role in the development and drug resistance of NSCLC (Herreros-Pomares et al., 2019; Huang et al., 2019; Tahmasebi et al., 2019). In this study, we identified the significance of the mRNAsi in the clinical characteristics of patients with LUSC with the help of data from the TCGA database and the mRNAsi corresponding to each sample. Moreover, hub genes related to the mRNAsi were obtained by the WGCNA method, and the association between the change in hub gene expression and clinical features was verified by external data from the GEPIA database. The results also indicated that all hub genes are highly expressed in tumor tissues, and some hub genes are of great significance as the disease progresses. Finally, after adjusting for the effects of confounding factors, we obtained a prognostic signature

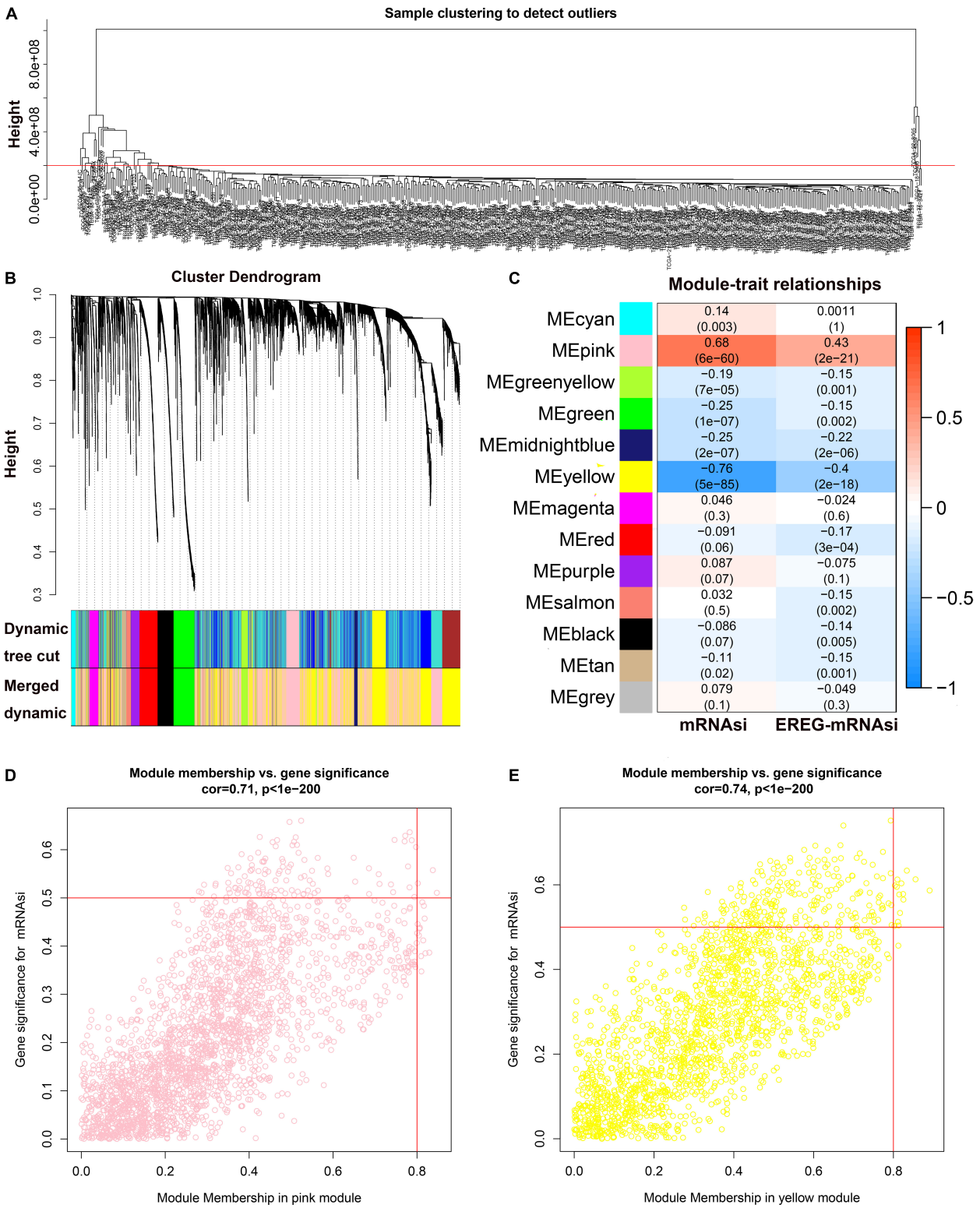


FIGURE 2 | WGCNA of LUSC. **(A)** Clustering of samples and removal of outliers. **(B)** Cluster dendrogram of genes in LUSC patients. Each branch in the figure represents one gene, and each color represents one coexpression module. **(C)** Correlation between the gene module and clinical characteristics, including the mRNAsi and EREG-mRNAsi. The correlation coefficient in each cell represents the correlation between the gene module and clinical characteristics and decreases in size, from red to blue. **(D)** Scatter diagram for MM vs. GS for the mRNAsi in the pink module. **(E)** Scatter diagram for MM vs. GS for the mRNAsi in the yellow module. LUSC, lung squamous cell carcinoma; mRNAsi, mRNA expression-based stemness index; EREG, epigenetically regulated.

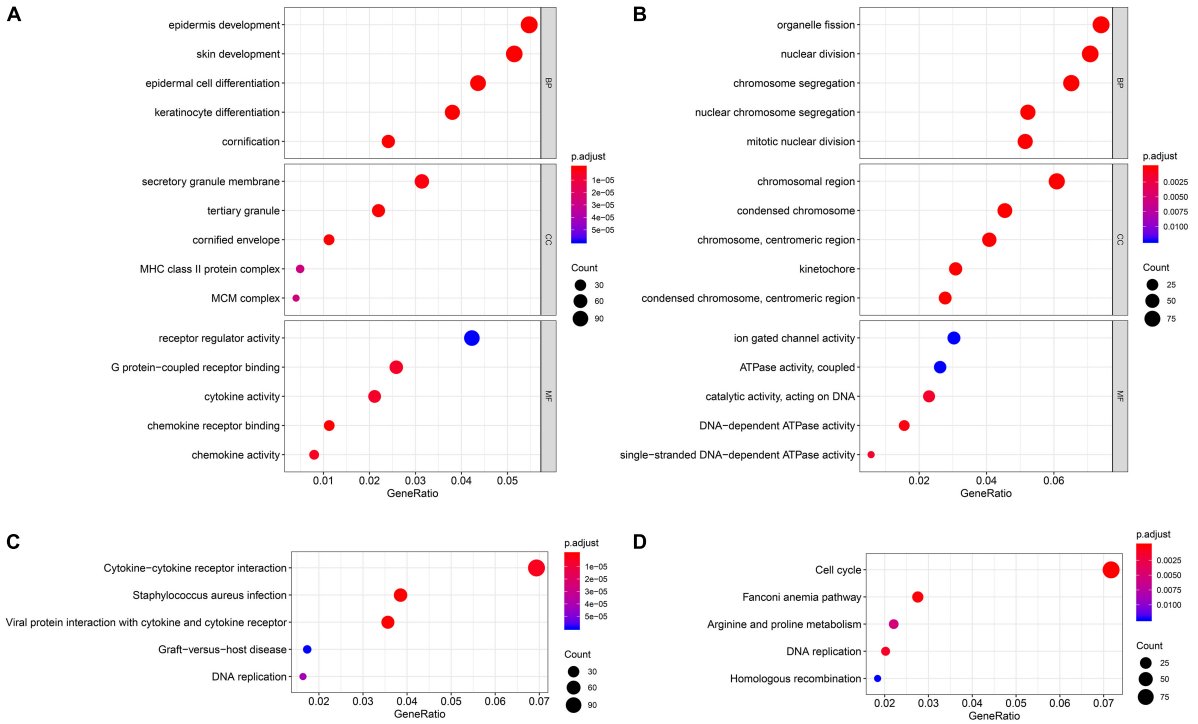


FIGURE 3 | Enrichment analyses of the significant module. **(A)** GO enrichment analysis of the pink module. **(B)** GO enrichment analysis of the yellow module. **(C)** KEGG pathway enrichment analysis of the pink module. **(D)** KEGG pathway enrichment analysis of the yellow module. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

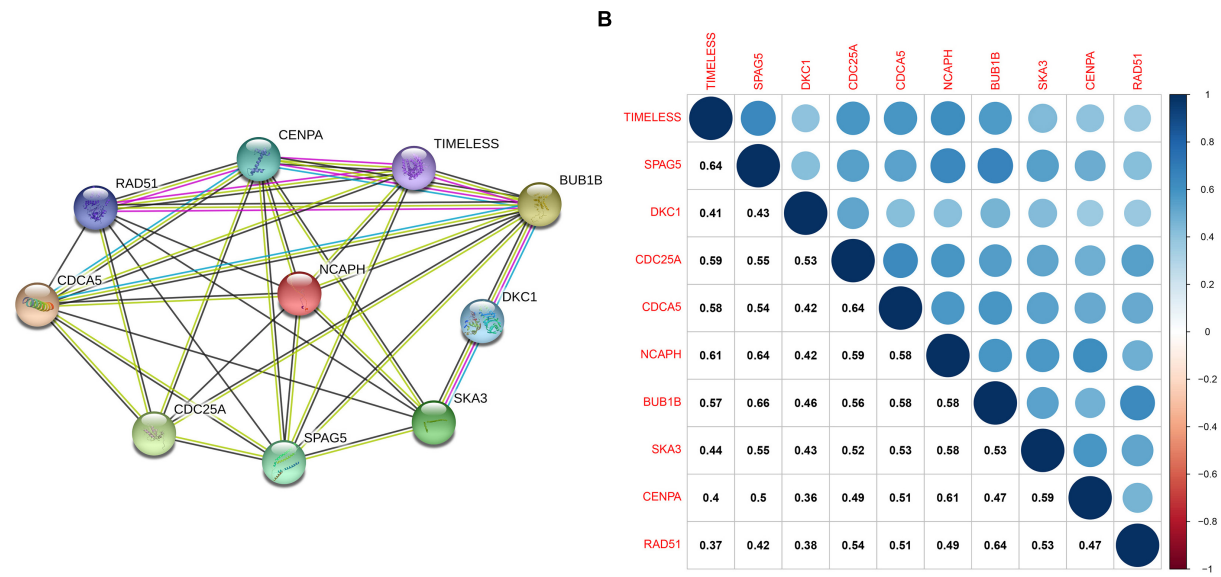


FIGURE 4 | (A) Protein–protein interactions between hub genes. The thickness of the solid line represents the strength of the relationship. **(B)** Correlation between the hub genes.

containing three prognostic genes with great predictive capacity. Two hub genes were not only highly expressed in LUSC but also associated with TNM stage and prognosis: *CENPA* and *NCAPH*.

CENPA (centromere protein A) is the key determinant of centromere identity (Fujita et al., 2015). The centromere is the basis of centromere formation and the key to chromosome separation during mitosis. *CENPA* controls the epigenetic

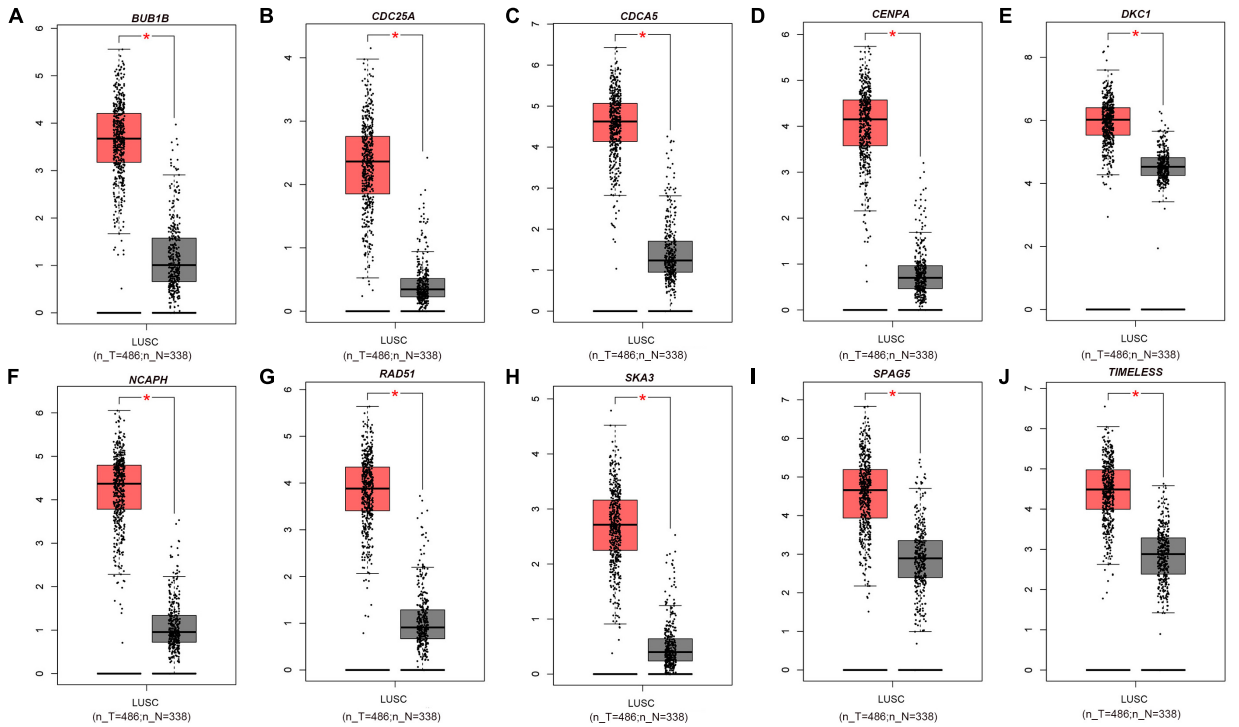


FIGURE 5 | Box plot of the top 10 hub genes whose expression levels were verified with data from the GEPIA database: **(A)** *BUB1B*, **(B)** *CDC25A*, **(C)** *CDCA5*, **(D)** *CENPA*, **(E)** *DKC1*, **(F)** *NCAPH*, **(G)** *RAD51*, **(H)** *SKA3*, **(I)** *SPAG5*, and **(J)** *TIMELESS*. Red * indicates $P < 0.05$.

identity of the centromere, which is essential for recruiting centromere elements to connect chromosomes to the mitotic spindle during mitosis (Rošic and Erhardt, 2016). Behnan et al. (2016) showed that the inhibition of *CENPA* expression in glioblastoma cells can reduce sphere-forming capacity, proliferation, and cell viability. *In vitro* experiments by Cheng et al. (2019) in mice confirmed that elevated *FOXM1* expression enhanced *CENPA* and *CENPB* expression, which promoted cell cycle progression and cell proliferation, thereby promoting LUSC cell growth. Regarding prognosis, lung adenocarcinoma (LUAD) patients with high *CENPA* expression experience poor OS based on data integrated from six different GEO chips by Liu et al. (2018).

NCAPH (non-SMC condensin I complex subunit H), whose expression is significantly high in both LUAD and LUSC (Ma et al., 2019), belongs to the protein superfamily defined as kleisins (Neuwald and Hirano, 2000). *NCAPH* plays an important role in the separation of the cell's sister chromatids and the maintenance of the mitotic chromosomal structure, and studies have demonstrated that its biallelic mutation can lead to a significant reduction in brain size in mice (Martin et al., 2016). Gene knockout experiments in mice by Sun et al. (2019) confirmed that *NCAPH* promotes the proliferation, migration and invasion of liver cancer *in vivo* and *in vitro*. The same experimental method also indicated that *NCAPH* is a potential candidate for radiation tolerance (Wang X.C. et al., 2019). High *NCAPH* expression also suggests a poor prognosis in prostate (Cui et al., 2019) and rectal (Yin et al., 2017) cancer patients.

BUB1B (BUB1 mitotic checkpoint serine/threonine kinase B) encodes a kinase involved in spindle checkpoint function and mitosis and plays an important role in the development of many types of cancer (Baker et al., 2013). Interestingly, *BUB1B* is also highly expressed in SCLC. For patients with SCLC, the higher the expression level of *BUB1B* is, the worse the prognosis (Liao et al., 2019). *BUB1B* has also been found to regulate the development of stem cells; for example, in experiments with embryonic stem cells, Su et al. (2019) found that knocking down *BUB1B* can lead to DNA damage and other forms of genomic instability, activate p53 and eventually lead to embryonic stem cell differentiation and possibly cancer.

CDC25A (cell division cycle 25A) and *CDCA5* (cell division cycle-associated 5) are important regulators of cell mitosis. *CDC25A* is a "switch" protein that controls G1/S and G2/M checkpoints and plays an important role in maintaining the stability of DNA replication and the integrity of the cell division cycle (Russell and Nurse, 1986). *CDCA5* is located on human chromosome 11q13.1, and its function is to ensure accurate separation of the sister chromosome during anaphase mitosis (Schmitz et al., 2007). *CDC25A* is overexpressed in many tumor cells and is associated with the malignancy and prognosis of several cancers, such as human glioma (Yamashita et al., 2010), retinoblastoma (Singh et al., 2015), breast cancer (Brunetto et al., 2013), and liver cancer (Lu et al., 2016). *CDCA5* is highly expressed in NSCLC (Chang et al., 2015) and liver cancer (Shen et al., 2018) and is associated with a poor prognosis. The genes in the CDC family

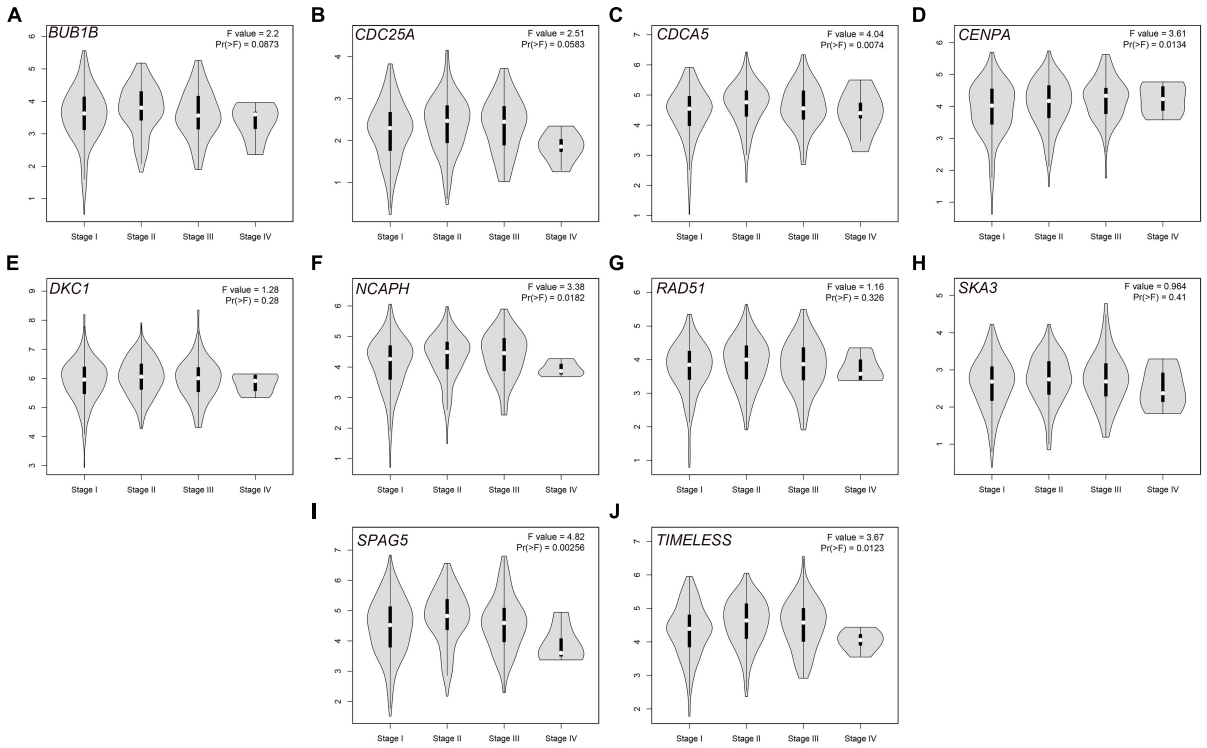


FIGURE 6 | Violin plots of the expression levels of the selected hub genes in different stages via the GEPIA database: **(A)** *BUB1B*, **(B)** *CDC25A*, **(C)** *CDCA5*, **(D)** *CENPA*, **(E)** *DKC1*, **(F)** *NCAPH*, **(G)** *RAD51*, **(H)** *SKA3*, **(I)** *SPAG5*, and **(J)** *TIMELESS*.

are key in driving the cell cycle, as well as promoting the formation of centrosomes and mitotic spindles. *CDCA5* is also a biomarker of malignant glioma of neural stem cell origin (Hembram et al., 2019).

DKC1 (dyskerin pseudouridine synthase 1) is highly conserved and widely expressed and may play additional roles in nucleocytoplasmic shuttling, the DNA damage response, and cell adhesion. The *DKC1* gene encodes a protein responsible for the stability of the telomerase whole enzyme complex. The mutation of *DKC1* has a strong influence on telomere repair and hematopoietic development. Induced pluripotent stem cells extracted from the fibroblasts of patients with X-linked keratosis disorder demonstrated that defective *DKC1* count not extend telomeres (Donaires et al., 2019). A study by Nersisyan et al. (2019) found that *DKC1* may also be involved in the activation of telomere maintenance mechanisms that lead to cancer. *DKC1* downregulation can also inhibit the growth of glioma cells by altering the expression of cell cycle-related molecules, causing cells to arrest in G1 phase. *In vitro* experiments also confirmed that glioma cells with *DKC1* knockout showed low activity (Miao et al., 2019). *DKC1* is highly expressed in clear cell renal cell carcinoma. *DKC1* knockout inhibits tumor proliferation, migration, and invasion by regulating the NF-κB/MMP-2 signaling pathway *in vitro* (Zhang M. et al., 2018).

RAD51 (RAD51 recombinase) is a key player in homologous recombination because it is closely related to and binds to DNA and exhibits ATPase activity following stimulation (Zhang et al., 2009). Chiu et al. (2019) found that the overexpression of *RAD51* promotes the viability of esophageal cancer cells, while its inhibition weakens the viability of esophageal cancer cells through cell cycle entry, migration/invasion and epithelial-mesenchymal transformation. Another study also found a mutation in the stem cell marker *RAD51* upon exon sequencing in patients with esophageal squamous cell carcinoma (ESCC) and revealed that *RAD51* is related to the drug sensitivity of ESCC (Golyan et al., 2020). High *RAD51* expression enhances cancer progression through the p38/Akt/Snail signaling pathway (Chiu et al., 2019). *RAD51* may lead to increased drug resistance in triple-negative breast cancer patients (Zhao et al., 2019) and is also associated with the radiosensitivity of some tumors, such as nasopharyngeal cancer (Zhang Z. et al., 2018) and prostate cancer (Maranto et al., 2018).

Three other genes also merit further study. *SKA3* (spindle and kinetochore-associated complex subunit 3) is an important subunit of the spindle and centromere-related protein complex and plays a regulatory role in cell proliferation and apoptosis (Zhang et al., 2017). A previous study showed that the incidence of age-related neurodegenerative diseases is consistent with a dramatic decline in the number and function of adult neural stem cells, while *SKA3* is closely associated with age-related central nervous system diseases. *SKA3* overexpression promotes the growth and migration of cervical cancer cells by activating the PI3K-AKT signaling pathway and promoting cell cycle progression (Hu

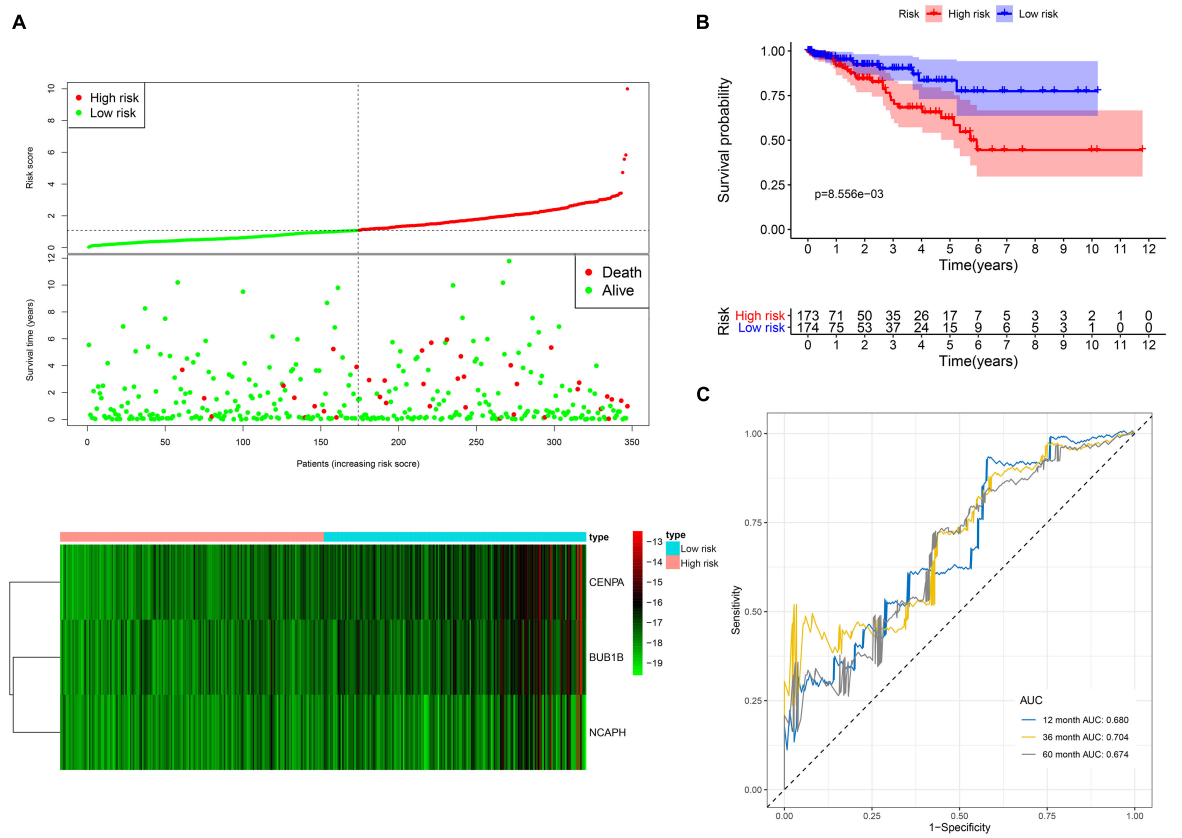


FIGURE 7 | The three-hub mRNA signature used to predict OS in LUSC patients. **(A)** Distribution, patient survival status and heatmap of the three-hub mRNA expression profiles. **(B)** Kaplan–Meier survival estimates of OS in LUSC patients according to the three-hub mRNA signature. **(C)** ROC analysis for the prediction of 1-, 3-, and 5-year OS as the defining point of the three-hub mRNA signature. LUSC, lung squamous cell carcinoma; ROC, receiver operating characteristic; OS, overall survival.

TABLE 1 | Multivariable Cox regression analysis of the three-hub gene signature.

Gene	Univariate analysis			Multivariate analysis			Coefficient
	HR	95%CI	*P*-value	HR	95%CI	*P*-value	
BUB1B	0.642	0.425–0.970	0.035	0.543	0.233–0.764	0.008	−0.630
CDC25A	0.757	0.521–0.972	0.046	1.028	0.484–2.184	0.940	–
CDCA5	0.797	0.495–1.238	0.350				
CENPA	0.665	0.449–0.985	0.042	0.498	0.244–0.952	0.046	−0.652
DKC1	0.607	0.352–1.048	0.073				–
NCAPH	1.313	1.172–2.213	0.006	2.286	1.172–4.213	0.023	2.163
RAD51	0.719	0.454–1.138	0.159				–
SKA3	0.781	0.512–1.191	0.025	1.346	0.693–2.617	0.380	–
SPAG5	0.685	0.474–0.990	0.044	0.604	0.313–1.169	0.134	–
TIMELESS	1.049	1.020–2.418	0.029	1.049	0.455–2.418	0.904	–

et al., 2018). *SPAG5* (sperm-associated antigen 5) binds to microtubules and centromeres of the spindle during mitosis or meiosis as a spindle-binding protein (Cheng et al., 2007). A previous study showed that *SPAG5* inhibits apoptosis by activating mammalian target of rapamycin 1 (mTORC1) (Thedieck et al., 2013). *SPAG5* has also been linked to the development and metastasis of LUAD (Wang T. et al., 2019), stomach cancer (Liu et al., 2019), and breast cancer (Jiang et al., 2019).

TIMELESS (timeless circadian regulator) is one of the core genes involved in the biological rhythm and is abnormally expressed in liver cancer, breast cancer and lung cancer (Truong et al., 2016; Shostak, 2017). In recent years, it has been found that the expression of *TIMELESS* in tumors is not completely

TABLE 2 | Univariable and multivariable Cox regression analyses of clinical characteristics.

Variable	Univariate analysis			Multivariate analysis		
	HR	95%CI	*P*-value	HR	95%CI	*P*-value
Age	1.060	1.011–1.111	0.015	1.036	0.990–1.083	0.128
Sex (Female/Male)	0.878	0.442–1.745	0.711	–		
Race (Non-Caucasian/Caucasian)	0.274	0.107–0.703	0.007	0.231	0.086–0.609	0.003
T stage (T1-T2/T3-T4)	1.449	0.603–3.485	0.407	–		
N stage (N0-N1/N2-N3)	1.577	0.612–4.068	0.346	–		
M stage (M0/M1)	1.674	0.843–3.271	0.670	–		
AJCC stage (I-II/III-IV)	1.588	1.079–3.423	0.027	3.628	1.539–8.549	0.006
Pharmacotherapy (No/Yes)	0.095	0.013–0.696	0.020	0.072	0.009–0.552	0.011
Radiotherapy (No/Yes)	0.737	0.226–2.400	0.612	–		
Risk score (Low/High)	6.616	2.877–15.212	<0.001	4.517	1.956–10.434	<0.001

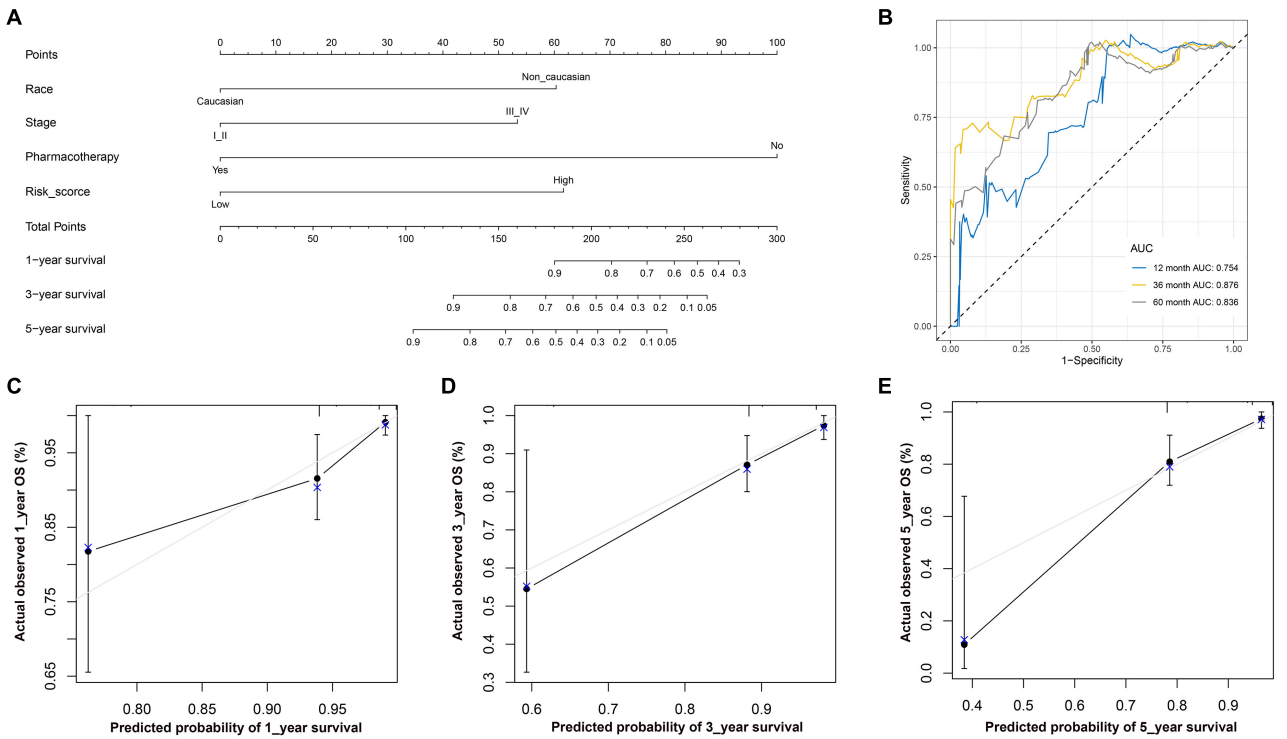


FIGURE 8 | (A) Nomogram used to predict 1-, 3-, and 5-year OS. **(B)** ROC curve based on the nomogram for 1-, 3-, and 5-year OS probability. **(B)** ROC analysis of 1-, 3-, and 5-year OS as the defining points of the nomogram. **(C)** Calibration curves for predicting 1–year OS. The nomogram–predicted probability of survival is plotted on the *x*–axis; actual survival is plotted on the *y*–axis. **(D)** Calibration curves for predicting 3–year OS. **(E)** Calibration curves for predicting 5–year OS. According to the scores of the corresponding variables in the nomogram, namely, the points at the top of the chart, the 1-, 3-, and 5-year survival rates of patients can be predicted according to the total scores of all corresponding individual scores. Theoretically, the standard curve is a straight line with a slope of 1 through the origin of the coordinate axis. If the predicted calibration curve is closer to the standard curve, the better the prediction capacity of the nomogram is. LUSC, lung squamous cell carcinoma; ROC, receiver operating characteristic; OS, overall survival.

uniform: it is weakly expressed in renal cell carcinoma, ductal cell carcinoma of the pancreas and other malignant tumors and is closely related to poor patient survival. However, it is highly expressed in colon cancer and significantly correlated with lymphatic metastasis, satellite metastasis and TNM stage (Top et al., 2016; Ozturk et al., 2017). *TIMELESS* plays a crucial role in the self-renewal process of breast CSCs and interacts with Sp1/c-jun to induce miR-5188 expression by promoting c-jun-mediated transcription, thus promoting breast cancer progression (Zou et al., 2020).

Our research also has the following limitations. First, we used data from a public database to confirm our findings and did not perform further experiments to confirm the expression of related genes or the molecular mechanisms and pathways involved.

Second, since our study examined data from a public database and online tools, the quality of these data may not be guaranteed. Finally, the data we studied were obtained almost entirely from the United States and are not representative of patients worldwide. Therefore, further well-designed biological studies with large sample sizes are needed to confirm our findings.

CONCLUSION

CDC25A, DKC1, CDCA5, BUB1B, SKA3, TIMELESS, NCAPH, SPAG5, CENPA, and *RAD51* may have a strong influence on LUSC stem cell maintenance. These hub genes may serve as control targets for LUSC CSCs, and further study of these genes may lead to new anticancer therapies.

DATA AVAILABILITY STATEMENT

RNA-seq and clinical information for each sample were downloaded from the TCGA website (https://cancergenome.nih.gov/).

AUTHOR CONTRIBUTIONS

YL conceived and designed the study, acquired and analyzed the data, and wrote the manuscript. HX and MC contributed to data analysis and manuscript drafting. XF designed the study and revised the manuscript.

ACKNOWLEDGMENTS

We would like to thank American Journal Experts for their assistance with language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00427/full#supplementary-material

FIGURE S1 | A flow chart of the research conducted.

TABLE S1 | The results of WGCNA.

DATA SHEET S1 | The code of R software.

REFERENCES

Ahmed, F. (2019). Integrated network analysis reveals FOXM1 and MYBL2 as key regulators of cell proliferation in non-small cell lung cancer. *Front. Oncol.* 9:1011. doi: 10.3389/fonc.2019.01011

Baker, D. J., Dawlaty, M. M., Wijshake, T., Jeganathan, K. B., Malureanu, L., Van Ree, J. H., et al. (2013). Increased expression of BubR1 protects against aneuploidy and cancer and extends healthy lifespan. *Nat. Cell Biol.* 15, 96–102. doi: 10.1038/ncb2643

Barlesi, F., Mazieres, J., Merlio, J. P., Debieuvre, D., Mosser, J., Lena, H., et al. (2016). Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French cooperative thoracic intergroup (IFCT). *Lancet* 387, 1415–1426. doi: 10.1016/S0140-6736(16)00004-0

Behnan, J., Grieg, Z., Joel, M., Ramsness, I., and Stangeland, B. (2016). Gene knockdown of CENPA reduces sphere forming ability and stemness of glioblastoma initiating cells. *Neuroepigenetics* 7, 6–18. doi: 10.1016/j.nepig.2016.08.002

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

Brunetto, E., Ferrara, A. M., Rampoldi, F., Talarico, A., Cin, E. D., Grassini, G., et al. (2013). CDC25A protein stability represents a previously unrecognized target of HER2 signaling in human breast cancer: implication for a potential clinical relevance in trastuzumab treatment. *Neoplasia* 15, 579–590. doi: 10.1593/neo.122054

Camp, R. L., Dolled-Filhart, M., and Rimm, D. L. (2004). X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin. Cancer Res.* 10, 7252–7259. doi: 10.1158/1078-0432.ccr-04-0713

Chang, I. W., Lin, V. C., He, H. L., Hsu, C. T., Li, C. C., Wu, W. J., et al. (2015). CDCA5 overexpression is an indicator of poor prognosis in patients with urothelial carcinomas of the upper urinary tract and urinary bladder. *Am. J. Transl. Res.* 7, 710–722.

Chen, W., Sun, K., Zheng, R., Zeng, H., Zhang, S., Xia, C., et al. (2018). Cancer incidence and mortality in China, 2014. *Chin. J. Cancer Res.* 30, 1–12.

Cheng, T. S., Hsiao, Y. L., Lin, C. C., Hsu, C. M., Chang, M. S., Lee, C. I., et al. (2007). hNinein is required for targeting spindle-associated protein astrin to the centrosome during the S and G2 phases. *Exp. Cell Res.* 313, 1710–1721. doi: 10.1016/j.yexcr.2007.02.023

Cheng, Z., Yu, C., Cui, S., Wang, H., Jin, H., Wang, C., et al. (2019). circTP63 functions as a ceRNA to promote lung squamous cell carcinoma progression by upregulating FOXM1. *Nat. Commun.* 10:3200. doi: 10.1038/s41467-019-11162-4

Chiu, W. C., Fang, P. T., Lee, Y. C., Wang, Y. Y., Su, Y. H., Hu, S. C., et al. (2019). DNA repair protein Rad51 induces tumor growth and metastasis in esophageal squamous cell carcinoma via a p38/Akt-dependent pathway. *Ann. Surg. Oncol.* doi: 10.1245/s10434-019-08043-x [Epub ahead of print]

Cui, F., Hu, J., Xu, Z., Tan, J., and Tang, H. (2019). Overexpression of NCAPH is upregulated and predicts a poor prognosis in prostate cancer. *Oncol. Lett.* 17, 5768–5776. doi: 10.3892/ol.2019.10260

Donaires, F. S., Alves-Paiva, R. M., Gutierrez-Rodrigues, F., da Silva, F. B., Tellechea, M. F., Moreira, L. F., et al. (2019). Telomere dynamics and hematopoietic differentiation of human DKC1-mutant induced pluripotent stem cells. *Stem Cell Res.* 40:101540. doi: 10.1016/j.scr.2019.101540

Friedmann-Morvinski, D., and Verma, I. M. (2014). Dedifferentiation and reprogramming: origins of cancer stem cells. *EMBO Rep.* 15, 244–253. doi: 10.1002/embr.201338254

Fujita, R., Otake, K., Arimura, Y., Horikoshi, N., Miya, Y., Shiga, T., et al. (2015). Stable complex formation of CENP-B with the CENP-A nucleosome. *Nucleic Acids Res.* 43, 4909–4922. doi: 10.1093/nar/gkv405

Goldfeder, R. L., Wall, D. P., Khoury, M. J., Ioannidis, J. P. A., and Ashley, E. A. (2017). Human genome sequencing at the population scale: a primer on high-throughput DNA sequencing and analysis. *Am. J. Epidemiol.* 186, 1000–1009. doi: 10.1093/aje/kww224

Golyan, F. F., Moghaddassian, M., Forghanifard, M. M., Talebi, S., Farshchian, M., Mahmoudian, R. A., et al. (2020). Whole exome sequencing reveals a novel damaging mutation in human fibroblast activation protein in a family with esophageal squamous cell carcinoma. *J Gastrointest Cancer* 51, 179–188. doi: 10.1007/s12029-019-00224-x

Hembram, K. C., Chatterjee, S., Sethy, C., Nayak, D., Pradhan, R., Molla, S., et al. (2019). Comparative and mechanistic study on the anticancer activity of quinacrine-based silver and gold hybrid nanoparticles in head and neck

cancer. *Mol. Pharm.* 16, 3011–3023. doi: 10.1021/acs.molpharmaceut.9b00242

Herreros-Pomares, A., De-Maya-Girones, J. D., Calabuig-Fariñas, S., Lucas, R., Martínez, A., Pardo-Sánchez, J. M., et al. (2019). Lung tumorspheres reveal cancer stem cell-like properties and a score with prognostic impact in resected non-small-cell lung cancer. *Cell Death Dis.* 10:660. doi: 10.1038/s41419-019-1898-1

Hu, R., Wang, M. Q., Niu, W. B., Wang, Y. J., Liu, Y. Y., Liu, L. Y., et al. (2018). SKA3 promotes cell proliferation and migration in cervical cancer by activating the PI3K/Akt signaling pathway. *Cancer Cell Int.* 18, 183. doi: 10.1186/s12935-018-0670-4

Huang, T. H., Wu, A. T. H., Cheng, T. S., Lin, K. T., Lai, C. J., Hsieh, H. W., et al. (2019). In silico identification of thiostrepton as an inhibitor of cancer stem cell growth and an enhancer for chemotherapy in non-small-cell lung cancer. *J. Cell. Mol. Med.* 23, 8184–8195. doi: 10.1111/jcmm.14689

Jiang, J., Wang, J., He, X., Ma, W., Sun, L., Zhou, Q., et al. (2019). High expression of SPAG5 sustains the malignant growth and invasion of breast cancer cells through the activation of Wnt/beta-catenin signalling. *Clin. Exp. Pharmacol. Physiol.* 46, 597–606. doi: 10.1111/1440-1681.13082

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

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

Leon, G., MacDonagh, L., Finn, S. P., Cuffe, S., and Barr, M. P. (2016). Cancer stem cells in drug resistant lung cancer: targeting cell surface markers and signaling pathways. *Pharmacol. Ther.* 158, 71–90. doi: 10.1016/j.pharmthera.2015.12.001

Liao, Y., Yin, G., Wang, X., Zhong, P., Fan, X., and Huang, C. (2019). Identification of candidate genes associated with the pathogenesis of small cell lung cancer via integrated bioinformatics analysis. *Oncol. Lett.* 18, 3723–3733. doi: 10.3892/ol.2019.10685

Liu, G., Liu, S., Cao, G., Luo, W., Li, P., Wang, S., et al. (2019). SPAG5 contributes to the progression of gastric cancer by upregulation of survivin depend on activating the wnt/beta-catenin pathway. *Exp. Cell Res.* 379, 83–91. doi: 10.1016/j.yexcr.2019.03.024

Liu, W. T., Wang, Y., Zhang, J., Ye, F., Huang, X. H., Li, B., et al. (2018). A novel strategy of integrated microarray analysis identifies CENPA, CDK1 and CDC20 as a cluster of diagnostic biomarkers in lung adenocarcinoma. *Cancer Lett.* 425, 43–53. doi: 10.1016/j.canlet.2018.03.043

Lu, X., Sun, W., Tang, Y., Zhu, L., Li, Y., Ou, C., et al. (2016). Identification of key genes in hepatocellular carcinoma and validation of the candidate gene, cdc25a, using gene set enrichment analysis, meta-analysis and cross-species comparison. *Mol. Med. Rep.* 13, 1172–1178. doi: 10.3892/mmr.2015.4646

Ma, Q., Xu, Y., Liao, H., Cai, Y., Xu, L., Xiao, D., et al. (2019). Identification and validation of key genes associated with non-small-cell lung cancer. *J. Cell. Physiol.* 234, 22742–22752. doi: 10.1002/jcp.28839

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.e315. doi: 10.1016/j.cell.2018.03.034

Maranto, C., Udhane, V., Hoang, D. T., Gu, L., Alexeev, V., Malas, K., et al. (2018). STAT5A/B blockade sensitizes prostate cancer to radiation through inhibition of RAD51 and DNA repair. *Clin. Cancer Res.* 24, 1917–1931. doi: 10.1158/1078-0432.CCR-17-2768

Martin, C. A., Murray, J. E., Carroll, P., Leitch, A., Mackenzie, K. J., Halachev, M., et al. (2016). Mutations in genes encoding condensin complex proteins cause microcephaly through decatenation failure at mitosis. *Genes Dev.* 30, 2158–2172. doi: 10.1101/gad.286351.116

Miao, F. A., Chu, K., Chen, H. R., Zhang, M., Shi, P. C., Bai, J., et al. (2019). Increased DKC1 expression in glioma and its significance in tumor cell proliferation, migration and invasion. *Invest. New Drugs* 37, 1177–1186. doi: 10.1007/s10637-019-00748-w

Nersisyan, L., Hopp, L., Loeffler-Wirth, H., Galle, J., Loeffler, M., Arakelyan, A., et al. (2019). Telomere length maintenance and its transcriptional regulation in lynch syndrome and sporadic colorectal carcinoma. *Front. Oncol.* 9:1172. doi: 10.3389/fonc.2019.01172

Neuwald, A. F., and Hirano, T. (2000). HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. *Genome Res.* 10, 1445–1452. doi: 10.1101/gr.147400

Ozturk, N., Ozturk, D., Kavakli, I. H., and Okyar, A. (2017). Molecular aspects of circadian pharmacology and relevance for cancer chronotherapy. *Int. J. Mol. Sci.* 18:E2168. doi: 10.3390/ijms18102168

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

Rošic, S., and Erhardt, S. (2016). No longer a nuisance: long non-coding RNAs join CENP-A in epigenetic centromere regulation. *Cell Mol. Life. Sci.* 73, 1387–1398. doi: 10.1007/s00018-015-2124-7

Russell, P., and Nurse, P. (1986). cdc25+ functions as an inducer in the mitotic control of fission yeast. *Cell* 45, 145–153. doi: 10.1016/0092-8674(86)90546-5

Schmitz, J., Watrin, E., Lenart, P., Mechtler, K., and Peters, J. M. (2007). Sororin is required for stable binding of cohesin to chromatin and for sister chromatid cohesion in interphase. *Curr. Biol.* 17, 630–636. doi: 10.1016/j.cub.2007.02.029

Shen, Z., Yu, X., Zheng, Y., Lai, X., Li, J., Hong, Y., et al. (2018). CDCA5 regulates proliferation in hepatocellular carcinoma and has potential as a negative prognostic marker. *OncoTargets Ther.* 11, 891–901. doi: 10.2147/OTT.S154754

Shibue, T., and Weinberg, R. A. (2017). EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nat. Rev. Clin. Oncol.* 14, 611–629. doi: 10.1038/nrclinonc.2017.44

Shostak, A. (2017). Circadian clock, cell division, and cancer: from molecules to organism. *Int. J. Mol. Sci.* 18:E873. doi: 10.3390/ijms18040873

Singh, L., Pushker, N., Sen, S., Singh, M. K., Bakhshi, S., Chawla, B., et al. (2015). Expression of CDC25A and CDC25B phosphatase proteins in human retinoblastoma and its correlation with clinicopathological parameters. *Br. J. Ophthalmol.* 99, 457–463. doi: 10.1136/bjophthalmol-2014-305830

Sokolov, A., Paull, E. O., and Stuart, J. M. (2016). One-class detection of cell states in tumor subtypes. *Pac. Symp. Biocomput.* 21, 405–416.

Su, J., Zhu, D., Huo, Z., Gingold, J. A., Ang, Y. S., Tu, J., et al. (2019). Genomic integrity safeguards self-renewal in embryonic stem cells. *Cell Rep.* 28:1400-1409.e1404. doi: 10.1016/j.celrep.2019.07.011

Sun, C., Huang, S., Wang, H., Xie, R., Zhang, L., Zhou, Q., et al. (2019). Non-SMC condensin I complex subunit H enhances proliferation, migration, and invasion of hepatocellular carcinoma. *Mol. Carcinog.* 58, 2266–2275. doi: 10.1002/mc.23114

Tahmasebi, S., Amiri, M., and Sonenberg, N. (2019). Translational control in stem cells. *Front. Genet.* 9:709. doi: 10.3389/fgene.2018.00709

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, W98–W102. doi: 10.1093/nar/gkx247

Thedieck, K., Holzwarth, B., Prentzell, M. T., Boehlke, C., Klasener, K., Ruf, S., et al. (2013). Inhibition of mTORC1 by astrin and stress granules prevents apoptosis in cancer cells. *Cell* 154, 859–874. doi: 10.1016/j.cell.2013.07.031

Top, D., Harms, E., Syed, S., Adams, E. L., and Saez, L. (2016). GSK-3 and CK2 kinases converge on timeless to regulate the master clock. *Cell Rep.* 16, 357–367. doi: 10.1016/j.celrep.2016.06.005

Truong, K. K., Lam, M. T., Grandner, M. A., Sassoon, C. S., and Malhotra, A. (2016). Timing matters: circadian rhythm in sepsis, obstructive lung disease, obstructive sleep apnea, and cancer. *Ann. Am. Thorac. Soc.* 13, 1144–1154. doi: 10.1513/AnnalsATS.201602-125FR

Wang, T., Li, K., Song, H., Xu, D., Liao, Y., Jing, B., et al. (2019). p53 suppression is essential for oncogenic SPAG5 upregulation in lung adenocarcinoma. *Biochem. Biophys. Res. Commun.* 513, 319–325. doi: 10.1016/j.bbrc.2019.03.198

Wang, X. C., Yue, X., Zhang, R. X., Liu, T. Y., Pan, Z. Z., Yang, M. J., et al. (2019). Genome-wide RNAi screening identifies RFC4 as a factor that mediates radioresistance in colorectal cancer by facilitating nonhomologous end joining repair. *Clin. Cancer Res.* 25, 4567–4579. doi: 10.1158/1078-0432.CCR-18-3735

Yamaoka, K., Nakagawa, T., and Uno, T. (1978). Application of akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations. *J. Pharmacokinet. Biopharm.* 6, 165–175. doi: 10.1007/bf01117450

Yamashita, Y., Kasugai, I., Sato, M., Tanuma, N., Sato, I., Nomura, M., et al. (2010). CDC25A mRNA levels significantly correlate with Ki-67 expression in human glioma samples. *J. Neurooncol.* 100, 43–49. doi: 10.1007/s11060-010-0147-3

Yin, L., Jiang, L. P., Shen, Q. S., Xiong, Q. X., Zhuo, X., Zhang, L. L., et al. (2017). NCAPH plays important roles in human colon cancer. *Cell Death Dis.* 8:e2680. doi: 10.1038/cddis.2017.88

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, M., Pan, Y., Jiang, R., Hou, P., Shan, H., Chen, F., et al. (2018). DKC1 serves as a potential prognostic biomarker for human clear cell renal cell carcinoma and promotes its proliferation, migration and invasion via the NFkappaB pathway. *Oncol. Rep.* 40, 968–978. doi: 10.3892/or.2018.6484

Zhang, Z., Huo, H., Liao, K., Wang, Z., Gong, Z., Li, Y., et al. (2018). RPA1 downregulation enhances nasopharyngeal cancer radiosensitivity via blocking RAD51 to the DNA damage site. *Exp. Cell Res.* 371, 330–341. doi: 10.1016/j.yexcr.2018.08.025

Zhang, Q., Sivakumar, S., Chen, Y., Gao, H., Yang, L., Yuan, Z., et al. (2017). Ska3 phosphorylated by Cdk1 binds Ndc80 and recruits Ska to kinetochores to promote mitotic progression. *Curr. Biol.* 27:1477-1484.e1474. doi: 10.1016/j.cub.2017.03.060

Zhang, X. P., Galkin, V. E., Yu, X., Egelman, E. H., and Heyer, W. D. (2009). Loop 2 in saccharomyces cerevisiae Rad51 protein regulates filament formation and ATPase activity. *Nucleic Acids Res.* 37, 158–171. doi: 10.1093/nar/gkn914

Zhao, L., Si, C. S., Yu, Y., Lu, J. W., and Zhuang, Y. (2019). Depletion of DNA damage binding protein 2 sensitizes triple-negative breast cancer cells to poly ADP-ribose polymerase inhibition by destabilizing Rad51. *Cancer Sci.* 110, 3543–3552. doi: 10.1111/cas.14201

Zou, Y., Lin, X., Bu, J., Lin, Z., Chen, Y., Qiu, Y., et al. (2020). Timeless-Stimulated miR-5188-FOXO1/β-Catenin-c-Jun feedback loop promotes stemness via ubiquitination of β-catenin in breast cancer. *Mol. Ther.* 28, 313–327. doi: 10.1016/j.ymthe.2019.08.015

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 © 2020 Liao, Xiao, Cheng and Fan. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 03 July 2020
doi: 10.3389/fbioe.2020.00553

Identification of a Multi–Long Noncoding RNA Signature for the Diagnosis of Type 1 Diabetes Mellitus

Guannan Geng[1†], *Zicheng Zhang*[2†] *and Liang Cheng*[2*]

[1] *Department of Endocrinology, The First Affiliated Hospital of Harbin Medical University, Harbin, China,* [2] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*

OPEN ACCESS

Edited by:
Jianzhong Su,
Wenzhou Medical University, China

Reviewed by:
Lei Deng,
Central South University, China
Hui Ding,
University of Electronic Science and Technology of China, China

***Correspondence:**
Liang Cheng
liangcheng@hrbmu.edu.cn

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *07 March 2020*
Accepted: *07 May 2020*
Published: *03 July 2020*

Citation:
Geng G, Zhang Z and Cheng L (2020) Identification of a Multi–Long Noncoding RNA Signature for the Diagnosis of Type 1 Diabetes Mellitus. Front. Bioeng. Biotechnol. 8:553. doi: 10.3389/fbioe.2020.00553

Due to the increasing prevalence of type 1 diabetes mellitus (T1DM) and its complications, there is an urgent need to identify novel methods for predicting the occurrence and understanding the pathogenetic mechanisms of the disease. Accumulated data have demonstrated the potential of long noncoding RNAs (lncRNAs), as biomarkers in establishing diagnosis and predicting prognosis of numerous diseases. Yet, little is known about the expression patterns and regulatory roles of lncRNAs in the pathogenesis of T1DM and whether they can be used as diagnostic biomarkers for the disease. To further explore these questions, in the present study, we conducted a comparative analysis of the expression patterns of lncRNAs between 20 T1DM patients and 42 health controls by retrospectively analyzing a published microarray data set. Our results indicate that, compared with healthy controls, diabetic patients had altered levels of lncRNAs. Then, we used three time cross-validation strategy and support vector machine to propose a specific 26-lncRNA signature (termed 26LncSigT1DM). This 26LncSigT1DM signature can be used to effectively distinguish between healthy and diabetic individuals (area under the curve = 0.825) of a validation cohort. After the 26LncSigT1DM was prospectively validated, we used Pearson correlation to identify 915 mRNAs, whose expression levels were positively correlated with those of the 26 lncRNAs. According to their Gene Ontology annotations, these mRNAs participate in processes including cellular response to stimulus, cell communication, multicellular organismal process, and cell motility. Kyoto Encyclopedia of Genes and Genomes analysis demonstrated that the genes encoding the 915 mRNAs may be associated with the NOD-like receptor signaling pathway, transforming growth factor β signaling pathway, and mineral absorption, suggesting that the deregulation of these lncRNAs may mediate inflammatory abnormalities and immune dysfunctions, which jointly promote the pathogenesis of T1DM. Thus, our study identifies a novel diagnostic tool and may shed more light on the molecular mechanisms underlying the pathogenesis of T1DM.

Keywords: long noncoding RNAs, biomarkers, type 1 diabetes mellitus, diagnosis, signature

INTRODUCTION

As one of the most notorious autoimmune disorders, type 1 diabetes mellitus (T1DM) is a chronic childhood-onset disease caused by selective destruction of pancreatic islet beta cells (Petersmann et al., 2018). Pathogenic factors of T1DM include epigenetic, environmental, and genetic factors (Bluestone et al., 2010; Groop and Pociot, 2014). Additionally, it has been demonstrated that both immune dysfunctions and islet beta cell defects contribute to the pathogenesis of T1DM (Lennon et al., 2009; Bluestone et al., 2010). During the onset of the disease β-cell function in T1DM patients may have already been completely destroyed (Polychronakos and Li, 2011). Therefore, it is necessary to diagnose T1DM early in its development. However, as onsets of T2DM and other types of diabetes occurring earlier, traditional methods for diagnosing diabetes are no longer satisfactory (Zou et al., 2018). Therefore, there is an urgent need to understand the causal factors and pathogenetic mechanisms of T1DM for more effective diagnosis and treatment.

Long noncoding RNAs (lncRNAs) are a group of RNAs that are longer than 200 nucleotides and do not encode proteins (Ponting et al., 2009; Sun et al., 2013; Zhang et al., 2017; Su et al., 2018; Yao et al., 2018). Recent advancement in large-scale genomic analysis has greatly enriched our knowledge on this type of RNA. Increasing evidence suggests that lncRNAs are involved in a wide range of cellular biological processes (Huang et al., 2012; Zhang et al., 2017; Guo et al., 2018), including cell proliferation (Jin et al., 2017; Luo et al., 2017), differentiation (Alvarez-Dominguez et al., 2015), and apoptosis (Lu et al., 2016). Specifically, βlinc1 (formerly referred to as HI-LNC15), a lncRNA uniquely expressed in the islet, has been reported to facilitate the proper specification and maintain the normal function of islet β cells (Arnes et al., 2016). Interestingly, in βlinc1-deficient mature islet β cells, the expression of GLIS3, a causative gene for both T1DM and T2DM, was downregulated (Moran et al., 2012). MEG3 is another lncRNA implicated in diabetes. In mouse islet β cells, MEG3 has been demonstrated to play important roles in insulin synthesis and secretion (You et al., 2016).

Recently, researchers have shown that lncRNAs control the differentiation and function of innate and adaptive immune cells to coordinate different immune functions (Atianand and Fitzgerald, 2014; Chen et al., 2017). In a previous study, it was confirmed that lincRNA LincR-Ccr2-5′AS plays an important role in regulating the expression of T_H2-specific genes and is essential for migration of T_H2 cells (Hu et al., 2013). Furthermore, LincRNACox2 and lncRNA THRIL are two lncRNAs that are crucial for inflammatory activation because they can regulate the TLR signaling pathways (Carpenter et al., 2013). Taken together, all the aforementioned studies suggest that lncRNAs can regulate both the activation of the innate immune system and islet β cell function, the defects of which contribute significantly to the pathogenesis of T1DM. More importantly, accumulating evidence has shown that dysregulated lncRNA expression is associated with the development of T1DM (Motterle et al., 2015; Mirza et al., 2017), suggesting that lncRNAs could be used as biomarkers to assess the risk of T1DM. However, neither the expression patterns of lncRNAs in T1DM patients nor their potential as T1DM biomarkers has been thoroughly investigated.

To provide more insights into the expression patterns of lncRNAs in T1DM patients and evaluate their potential as T1DM biomarkers, in this study, we comparatively analyzed lncRNA expression levels in 42 healthy individuals and 20 T1DM patients based on a published microarray data set and identified a group of differentially expressed lncRNAs. We also demonstrated that these lncRNAs may represent a multi–long noncoding RNA signature (namely 26LncSigT1DM) that can be used to effectively distinguish between healthy and diabetic individuals and identify T1DM susceptible individuals. After the 26LncSigT1DM signature was prospectively validated, we identified 915 mRNAs whose expression levels are positively correlated with those of the 26LncSigT1DM lncRNAs. Functional analysis of these mRNAs indicates that they are involved in a variety of biological processes, including cellular function and communication, and that the genes encoding these mRNAs are associated with pathways that can potentially mediate inflammatory abnormalities and immune dysfunctions. Our study provides a platform for developing 26LncSigT1DM into a diagnostic tool for T1DM and for future research into the molecular mechanisms underlying the pathogenesis of T1DM.

MATERIALS AND METHODS

Participant Information

We included two cohorts of individuals in this study. One cohort (62 individuals, accession number GSE35713) was from Hara's study (Levy et al., 2012), and the data are from peripheral blood mononuclear cells in the plasma samples of patients with new-onset T1DM. The other (22 individuals, GSE55100) was documented in Yang et al. (2015). Patients without survival time or events were excluded.

Data Acquisition and lncRNA Expression Analysis

We obtained the raw microarray data (.CEL format) deposited in the Gene Expression Omnibus database from the individuals mentioned above. To ensure uniformity, we used the Robust Multichip Average algorithm to preprocess the data (Irizarry et al., 2003). To scale probe expression intensity, the data set was quantile-normalized and log2-transformed after background correction, and then it was independently standardized by Z score transformation (Cheadle et al., 2003).

Gene expression profiles of the individuals had been previously analyzed by an Affymetrix Human Genome U133 Plus 2.0 array (HG-U133 Plus_2.0). We visited the Affymetrix website (http://www.affymetrix.com) to obtain the probe sequences used in the array.

By repurposing the Affymetrix array probes, lncRNA expression profiles of the 62 individuals in cohort GSE35713 were determined as described in previous studies (Du et al., 2013; Zhou et al., 2015). Briefly, we mapped the probes to the human genome (GRCh38) using the SeqMap tool (Jiang and Wong,

2008) and used GENCODE Release 21 to determine lncRNA-encoding genes. If a probe corresponds to numerous lncRNAs, it will be directly abandoned. If an lncRNA is targeted by multiple probes, its expression value was defined as the average value of the expression levels determined by all the corresponding probes. Using this method, we were able to obtain the expression profiles of 1,326 lncRNAs.

To identify differentially expressed lncRNAs (DELs), we compared the lncRNA expression patterns between healthy and T1DM individuals using two-tailed *t*-tests. Bonferroni statistical tests were then carried out, and lncRNAs with a Bonferroni-corrected *p*-value below 0.01 (**Supplementary Table 1**) were considered as DELs, whereas those with a Bonferroni-corrected $p < 0.05$ but higher than 0.01 (**Supplementary Table 2**) were finally abandoned. To assess the potential of the DELs as biomarkers for T1DM, unsupervised hierarchical clustering analysis was carried out using R package based on the Euclidean distance and the complete linkage method.

Identification of lncRNAs Associated With T1DM

To propose a diagnostic lncRNA molecular signature for T1DM, we used a sigmoid kernel-based support vector machine (SVM) (Tang et al., 2018; Lai et al., 2019) and assessed the predictive ability of the model using 3-fold cross-validation, with 62 individuals in cohort GSE35713 being defined as the discovery cohort. The details are as follows:

(i) Individuals in the discovery cohort were equally divided into three nonoverlapping sets.
(ii) Candidate lncRNAs were sorted according to their importance in the random forest classification algorithm. Then, a supervised discriminative model was established, and lncRNAs with a Bonferroni-corrected $p < 0.01$ (**Supplementary Table 1**) were selected.
(iii) Distinguishing T1DM patients from healthy controls using the SVM-based signature based on voting rules: We added one candidate lncRNA each time sequentially according to the rankings of candidate lncRNAs in the list (i.e., the lncRNA ranked first in the list was added first). After the three nonoverlapping sets were applied, the performance of the SVM-based signature was evaluated.
(iv) The optimal number of lncRNAs in the SVM-based signature could be determined after the balance between lncRNA number and discrimination accuracy was achieved.

Performance Evaluation

The difference between healthy and T1DM individuals was plotted (with SVM) (Su et al., 2018; Yang et al., 2018; Zhu et al., 2019), of which the performance was tested by three nonoverlapping sets. A 2 × 2 contingency table was used to calculate the sensitivity, specificity, and accuracy of the area under the curve (AUC). We plotted true-positive rates (sensitivity) against false-positive rates (1—specificity) to generate the receiver operating characteristic (ROC) curve (Lv et al., 2019), which was then used to determine the discrimination efficiency. We employed Bioconductor and R package to conduct all the aforementioned analyses.

Functional Annotations of the 26 lncRNAs in the 26LncSigT1DM Signature

To predict biological functions of the 26 lncRNAs in the 26LncSigT1DM signature, the Pearson correlation coefficient was adopted to determine correlations between expression levels of mRNAs and those of the 26 lncRNAs. Then, the genes encoding the mRNAs whose expression levels are positively correlated with those of the 26 lncRNAs were subjected to the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional annotations using the DAVID (Database for Annotation, Visualization, and Integrated Discovery, version 6.7) (Huang da et al., 2009). Biological processes enriched in the GO analysis with a Benjamini–Hochberg–adjusted $p < 0.01$ and an enrichment score >1.5 were considered significant. Similarly, pathways enriched in the KEGG analysis with a Benjamini–Hochberg–adjusted $p < 0.01$ and an enrichment score >1.5 were also considered significant (**Figure 1**).

RESULTS

Identification of DELs Between Healthy and Diabetic Individuals

Hara's cohort (GSE35713), the larger cohort included in our study, contained 47 RO T1DM patients and 42 unrelated healthy controls (Levy et al., 2012). To identify RO T1DM-related lncRNAs, we selected 20 T1DM patients and 42 unrelated healthy controls and defined them as a discovery cohort ($n = 62$). Then, we compared lncRNA expression levels between the 20 T1DM patients and 42 healthy controls by performing SAM analysis. After abandoning lncRNAs with a false discovery rate (FDR)–adjusted $p < 0.01$, we totally identified 1,326 DELs (log FC >1 or log FC ≤1, FDR-adjusted $p < 0.01$, **Supplementary Table 3**).

Construction of SVM-Based and Multi-lncRNA Signature as a Diagnostic Tool for T1DM Using the Discovery Cohort

Sigmoid kernel-based SVM and 3-fold cross-validation strategies were used to search for a supervised T1DM predictor from the discovery cohort. We identified a signature of 26 lncRNAs, which were downregulated in T1DM patients in Hara's cohort, had the highest discrimination accuracy (**Figures 2A,B**). This signature was named 26LncSigT1DM (**Table 1**). At the same time, we found that 26 lncRNAs is also an optimal number of lncRNAs that balances lncRNA number and discrimination accuracy. By distinguishing healthy and diabetic individuals in Hara's cohort with this 26LncSigT1DM signature, we generated three ROC curves, whose AUCs are 0.9973, 0.9641, and 0.9556 (**Figure 2C**). Hierarchical clustering was then applied to analyze the expression profiles of the 26 lncRNAs in the 26lncSigT1DM signature in the healthy and diabetic individuals. We found that the 20 T1DM patients and the 42 healthy controls can be grouped into two significantly different clusters (the 20 T1DM patients were grouped into Cluster 1, whereas the 42 healthy

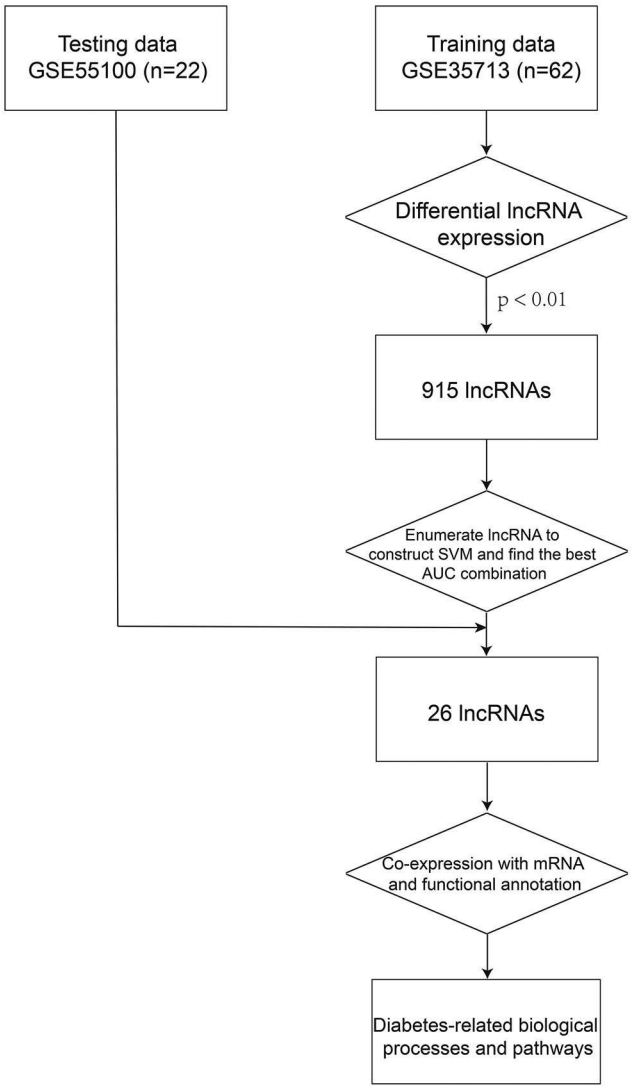


FIGURE 1 | The process of the method.

controls were grouped into Cluster 2) based on the expression levels of the 26 lncRNAs in the 26lncSigT1DM signature ($p = 3.579e\text{-}05$, χ^2-test). Therefore, we successfully distinguished between healthy and diabetic individuals in the discovery cohort using the 26lncSigT1DM signature. These results suggest that the downregulation of the 26 lncRNAs in the 26lncSigT1DM signature is able to reflect the disease progression of T1DM and that the signature has a great potential to be used as a diagnostic tool for T1DM.

Validation of the 26LncSigT1DM Signature

To test the stability and robustness of the 26LncSigT1DM, we introduced another cohort of 22 individuals (including 12 T1DM patients and 10 healthy controls) from Minglan's study (Du et al., 2013). This validation cohort was analyzed with the 26LncSigT1DM signature using sigmoid kernel-based SVM and 3-fold cross-validation strategies. According to our results, by distinguishing healthy and diabetic individuals in the validation cohort with the 26LncSigT1DM signature, we generated an ROC curve with an AUC of 0.825 (**Figures 3A,B**), suggesting that the proposed 26LncSigT1DM signature is a reliable diagnostic tool for T1DM.

Exploration of the Biological Functions of the 26 lncRNAs in the 26LncSigT1DM Signature

To predict biological functions of the 26 lncRNAs in the 26LncSigT1DM signature, we adopted the Pearson correlation coefficient to determine correlations between expression levels of mRNAs and those of the 26 lncRNAs. We found 915 mRNAs whose expression levels were positively correlated with those of the 26 lncRNAs. According to our GO analysis of the genes encoding these mRNAs, 470 biological processes were significantly enriched (**Supplementary Table 4**, p-values <0.05, **Figure 4A**). These biological processes can be clustered into four major functional groups, including cellular response to stimulus, cell communication, multicellular organismal process, and cell motility (**Figure 4B**). KEGG analysis of the genes encoding the 915 mRNAs indicates that they are implicated in several pathways including the NOD-like receptor signaling pathway, transforming growth factor β (TGF-β) signaling pathway, autoimmune thyroid disease, and mineral absorption (**Supplementary Table 5**). Given that all the biological processes and signaling pathways enriched in our GO and KEGG analyses are associated with the pathogenesis of T1DM, we speculate that the downregulation of the 26 lncRNAs may have caused the aberrant expression of a wide range of genes, which subsequently contributes to the pathogenesis of T1DM.

DISCUSSION

Type 1 diabetes mellitus, one of the most common childhood-onset chronic diseases, is caused by defects in pancreatic islet β cells. Complications of T1DM are very serious and sometimes fatal. For example, diabetic ketoacidosis is a life-threatening complication of T1DM caused by a shortage of insulin that demands insulin injections and blood glucose monitor. Recent years have witnessed a rising incidence of diabetes, which has been correlated with both environmental conditions and genetic factors (Fourlanos et al., 2008). To cope with this growing problem, great efforts and progress have been made in the last few years to explore the possible mechanisms underlying the pathogenesis of T1DM at the miRNA, mRNA, and protein levels. So far, several miRNA/mRNA/protein-based signatures have been associated with the occurrence of T1DM, which facilitates the development of new diagnostic and prognostic tools (Azhir et al., 2018; Liao et al., 2018; Bertoccini et al., 2019; Cheng et al., 2019; Guay et al., 2019).

Long noncoding RNAs are a novel group of gene expression regulators (Gibb et al., 2011; Kung et al., 2013). Increasing evidence has revealed that lncRNAs control the differentiation and function of innate and adaptive immune cells to coordinate

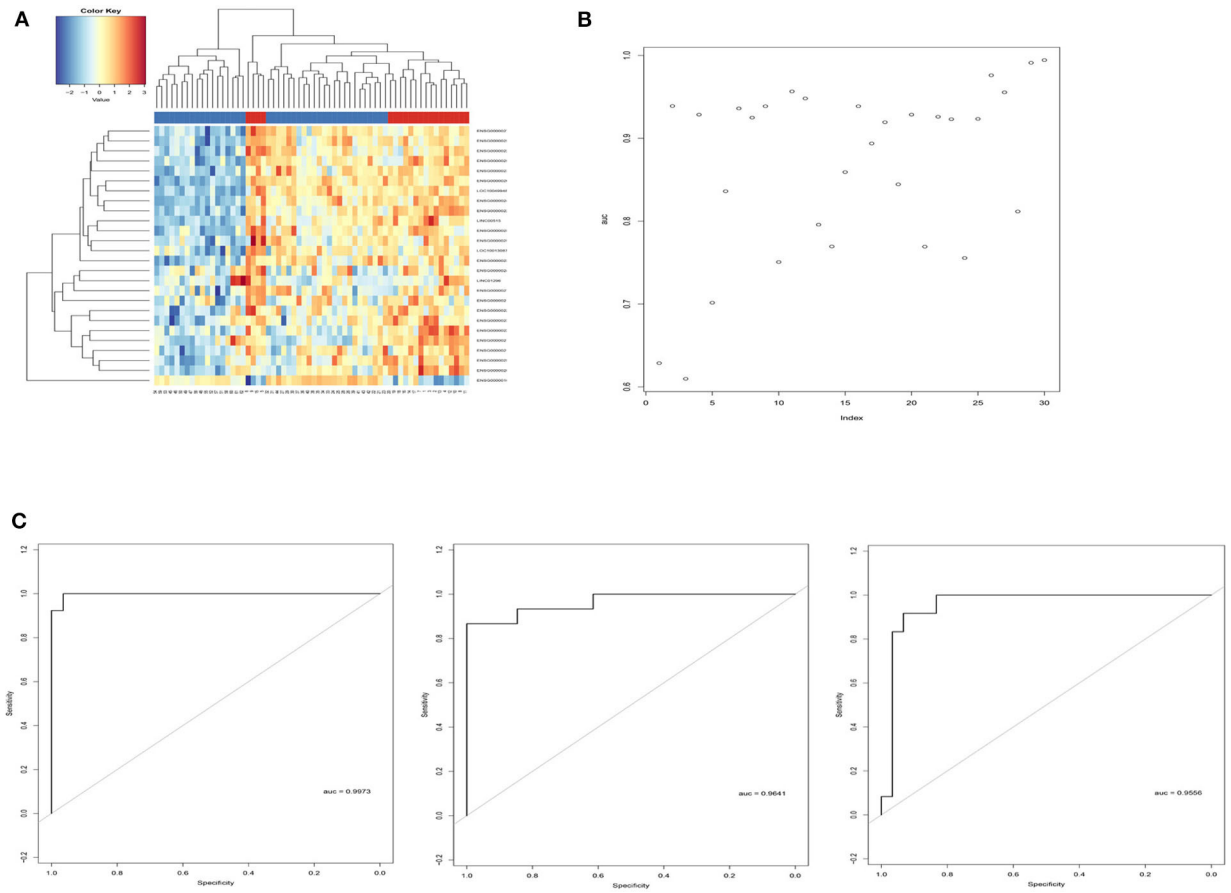


FIGURE 2 | Identification of the SVM and 3-fold cross-validation–based multi-lncRNA signature and its application in T1DM diagnosis. **(A)** Hierarchical clustering analysis of the 62 individuals in the discovery cohort based on the expression levels of the 26 lncRNAs in the 26LncSigT1DM signature. **(B)** Performance of different lncRNA numbers in distinguishing healthy and diabetic individuals of the discovery cohort. **(C)** Performance of the SVM and 3-fold cross-validation–based 26LncSigT1DM signature in distinguishing healthy and diabetic individuals of the discovery cohort.

several aspects of immune functions (Carpenter et al., 2013; Hu et al., 2013; Atianand and Fitzgerald, 2014; Chen et al., 2017). Therefore, their associations with autoimmune diseases have become a research hotspot. To date, several studies have demonstrated the potential of lncRNAs as novel diagnostic or prognostic tool for various types of cancer (Du et al., 2013; Zhou et al., 2015, 2017; Bao et al., 2019). Although several lncRNAs, including HILNC25, lncRNA MEG3, and MALAT-1, have been found to contribute to diabetes (Arnes et al., 2016; Lu et al., 2016; You et al., 2016), little is known about the expression profiles of lncRNAs in T1DM patients and whether lncRNAs can be used as diagnostic or prognostic tool for T1DM.

In the present study, we retrospectively analyzed a published microarray data set and determined the expression levels of lncRNAs in 62 individuals (the discovery cohort), including 20 T1DM patients and 42 healthy controls. Using this discovery cohort, we identified a supervised multi-lncRNA T1DM diagnostic signature, 26LncSigT1DM, based on SVM and 3-fold cross-validation strategies. This 26LncSigT1DM signature consists of 26 lncRNAs, whose expression levels were downregulated in the 20 T1DM patients as compared with the 42 healthy controls. Using the 26LncSigT1DM signature, we accurately distinguished between healthy and diabetic individuals in the discovery cohort. To test the stability and robustness of the 26LncSigT1DM signature, we introduced a 22-individual validation cohort (a cohort from Minglan's study) and found that the 26LncSigT1DM signature was also able to accurately distinguish between healthy and diabetic individuals in the validation cohort. According to the tree traversal algorithms, the number of combined lncRNAs was found to be not correlated with the model effects. Therefore, to avoid overfitting or underfitting, the combination of 26 lncRNA models was finally selected to build the classifier. These results suggest that the proposed 26LncSigT1DM signature has a great potential to be used as a diagnostic tool for T1DM. To the best of our knowledge, this is the first multi-lncRNA signature capable of diagnosing T1DM early in its development. However, there are several limitations in this study. First, only two microarray

TABLE 1 | Detailed information of the 26 lncRNAs in the 26LncSigT1DM signature.

Gene name	Gene. Name	Score	Numerator	Denominator	Fold change	*Q*-value
ENSG00000224020	MIR181A2HG	6.193	0.57	0.092	1.103	0
ENSG00000253165	–	2.504	0.185	0.074	1.066	0
ENSG00000259150	LINC00929	3.433	0.313	0.091	1.092	0
ENSG00000248118	LINC01019	3.363	0.305	0.091	1.104	0
ENSG00000231365	WARS2-AS1	3.151	0.19	0.06	1.06	0
LINC01296	DUXAP9	2.359	0.391	0.166	1.154	0.085
LINC00515	LINC00515	4.022	0.336	0.083	1.138	0
ENSG00000203999	LINC01270	2.826	0.25	0.088	1.088	0
ENSG00000227540	–	3.261	0.251	0.077	1.053	0
ENSG00000278156	TSC22D1-AS1	2.35	0.173	0.074	1.056	0.085
ENSG00000236519	LINC01424	2.352	0.195	0.083	1.082	0.085
ENSG00000257242	LINC01619	3.532	0.4	0.113	1.077	0
LOC100130872	–	2.785	0.253	0.091	1.049	0
ENSG00000215417	MIR17HG	4.629	0.342	0.074	1.107	0
ENSG00000186594	MIR22HG	−7.497	−0.701	0.093	0.924	0
ENSG00000275549	STPG3-AS1	2.528	0.193	0.076	1.043	0
ENSG00000214401	KANSL1-AS1	2.933	0.221	0.075	1.055	0
ENSG00000237940	LINC01238	2.842	0.269	0.095	1.094	0
ENSG00000212978	–	2.538	0.205	0.081	1.055	0
ENSG00000246339	EXTL3-AS1	2.152	0.135	0.063	1.039	0.085
ENSG00000281649	EBLN3P	3.874	0.226	0.058	1.029	0
ENSG00000223478	–	6.081	0.566	0.093	1.113	0
ENSG00000258573	–	2.564	0.204	0.079	1.048	0
LOC100499489	–	3.018	0.244	0.081	1.065	0
ENSG00000254813	–	3.51	0.336	0.096	1.082	0
ENSG00000229589	ACVR2B-AS1	3.157	0.36	0.114	1.091	0

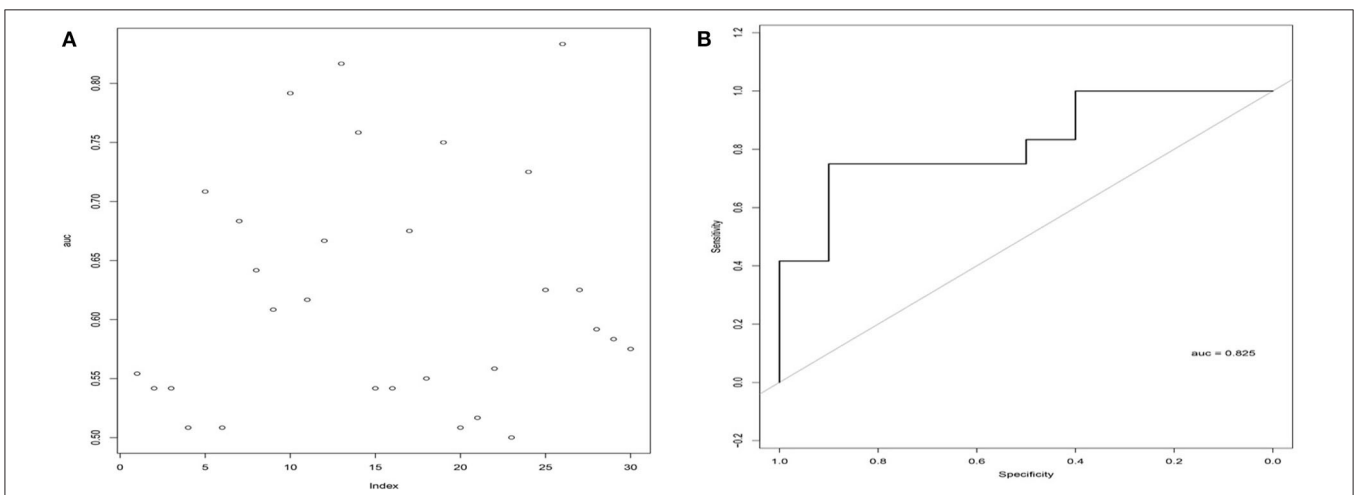


FIGURE 3 | Validation of the 26LncSigT1DM signature using an additional independent cohort. **(A)** Performance of different lncRNA numbers in distinguishing healthy and diabetic individuals of the validation cohort. **(B)** Performance of the SVM and 3-fold cross-validation–based 26LncSigT1DM signature in distinguishing healthy and diabetic individuals of the validation cohort.

datasets are available online, limiting the sample size analyzed in this study. Second, since the signature of T1DM was derived from newly-onset patients without the data after onset, the 26LncSigT1DM cannot be applied to the prognosis analysis. Last but not least, due to limited available data of T2DM patients, this signature is not capable of distinguishing T1DM from T2DM. Future studies related to these questions are worth conducting.

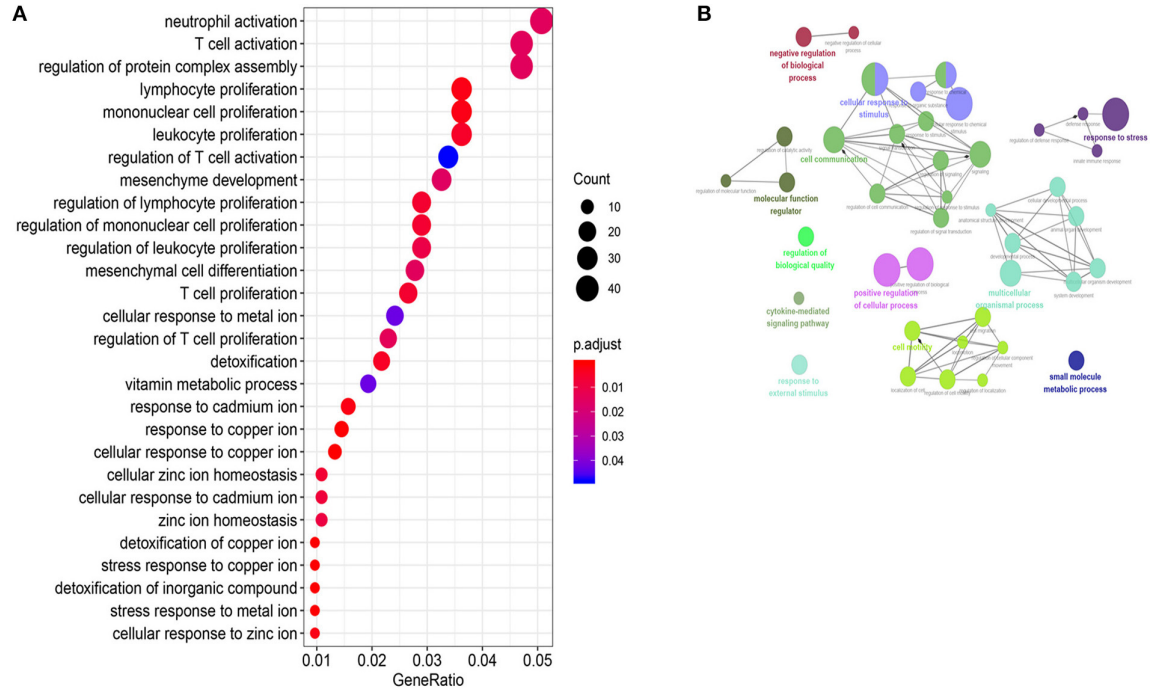


FIGURE 4 | Function annotations of the 26LncSigT1DM lncRNAs. **(A)** Gene Ontology and Kyoto Encyclopedia of Genes and Genomes annotations of the genes encoding the mRNAs whose expression levels are positively correlated with those of the 26 lncRNAs in the 26LncSigT1DM signature. The results indicate that the 26LncSigT1DM lncRNAs may affect immune processes. **(B)** The lncRNAs involved in regulating human biological processes.

Prior studies have confirmed that lncRNAs are important gene expression regulators because they modulate the expression of a wide range of functional genes involved in multiple biological processes (Guo et al., 2013; Liu et al., 2015). To predict biological functions of the 26 lncRNAs in the 26LncSigT1DM signature, we used the Pearson correlation coefficient to identify correlations between expression levels of mRNAs and those of the 26 lncRNAs. We found 915 mRNAs whose expression levels were positively correlated with those of the 26 lncRNAs. According to our GO and KEGG analyses, the genes encoding these mRNAs are involved in multiple T1DM-related biological processes and signaling pathways. These results are consistent with previous findings. In a previous study, it was reported that thioredoxin-interacting protein (TXNIP), an activator of NOD, LRR, and PYD domains-containing protein 3 (NLRP3) inflammasome, is associated with nonalcoholic fatty liver disease and T1DM (Wang et al., 2013). In addition, it was demonstrated that mitochondrial DNA-mediated NLRP3 activation can induce IL-1β secretion in the pancreas of STZ-induced T1DM mice (Carlos et al., 2017). The TGF-β signaling pathway inside T cells, which coordinates immune responses, has already been proved to play a critical role in the pathogenesis of T1DM (Green et al., 2003). Type 1 diabetes mellitus patients have already been thought to have increased risk to suffer from autoimmune thyroid disease compared with healthy individuals. Our KEGG analysis implies that autoimmune thyroid disease and mineral absorption may be closely related to T1DM. Among these 26 lncRNAs, LINC01619 alteration has been proved to influence the diabetic nephropathy by inducing oxidative stress and podocyte damage via regulating miR-27a (Bai et al., 2018). However, except for LINC01619, the mecanisms of other lncRNAs affecting T1DM still remain unclear, which need to be clarified. The results of the GO and KEGG analyses of the 26LncSigT1DM signature in this study pave the way for further studies to investigate the relationship between these lncRNAs and T1DM as well as its complications. The underlying mechanisms remain to be further studied. Thus, the results of our study may provide suggestive information for future research.

In summary, we conducted a comparative analysis of the lncRNA expression profiles between T1DM patients and healthy controls. And a dysregulated lncRNA-mRNA coexpression network was built to enrich our knowledge of T1DM-related lncRNAs. More importantly, we proposed and validated a 26LncSigT1DM signature that has a great potential to be used as a diagnostic tool for T1DM using sigmoid kernel-based SVM and 3-fold cross-validation strategies. This study is the first to use a multi-lncRNA signature to diagnose T1DM. Therefore, the 26LncSigT1DM signature proposed by our study may represent a good complement to the existing clinical diagnostic indicators for T1DM. Lastly, this study also improves our understanding of the mechanisms underlying the pathogenesis of T1DM and may provide other options for the prevention and treatment of T1DM.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE35713, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35713, GSE55100, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55100.

AUTHOR CONTRIBUTIONS

GG, ZZ, and LC designed and performed the experiments, analyzed the data, wrote the manuscript, organized the figures, and revised the manuscript. LC also improved the experimental performance, revised, finalized, and approved the article. All authors contributed to the article and approved the submitted version.

FUNDING

The Tou-Yan Innovation Team Program of the Heilongjiang Province (2019-15). National Natural Science Foundation of China (Grant No. 61871160). The Heilongjiang Postdoctoral Fund (Grant No. LBH-TZ20). Young Innovative Talents in Colleges and Universities of Heilongjiang Province (2018-69). The Fundamental Research Funds for the Provincial Universities (2018-KYYWF-0473).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00553/full#supplementary-material

REFERENCES

Alvarez-Dominguez, J. R., Bai, Z., Xu, D., Yuan, B., Lo, K. A., Yoon, M. J., et al. (2015). *De novo* reconstruction of adipose tissue transcriptomes reveals long non-coding RNA regulators of brown adipocyte development. *Cell Metabol.* 21, 764–776. doi: 10.1016/j.cmet.2015.04.003

Arnes, L., Akerman, I., Balderes, D. A., Ferrer, J., and Sussel, L. (2016). betalinc1 encodes a long noncoding RNA that regulates islet beta-cell formation and function. *Genes Devel.* 30, 502–507. doi: 10.1101/gad.273821.115

Atianand, M. K., and Fitzgerald, K. A. (2014). Long non-coding RNAs and control of gene expression in the immune system. *Trends Mol. Med.* 20, 623–631. doi: 10.1016/j.molmed.2014.09.002

Azhir, Z., Dehghanian, F., and Hojati, Z. (2018). Increased expression of microRNAs, miR-20a and miR-326 in PBMCs of patients with type 1 diabetes. *Mol. Biol. Rep.* 45, 1973–1980. doi: 10.1007/s11033-018-4352-z

Bai, X., Geng, J., Li, X., Wan, J., Liu, J., Zhou, Z., et al. (2018). Long noncoding RNA LINC01619 regulates MicroRNA-27a/Forkhead box protein O1 and endoplasmic reticulum stress-mediated podocyte injury in diabetic nephropathy. *Antioxid. Redox. Signal.* 29, 355–376. doi: 10.1089/ars.2017.7278

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief Bioinform*. doi: 10.1093/bib/bbz118. [Epub ahead of print].

Bertoccini, L., Sentinelli, F., Incani, M., Bailetti, D., Cimini, F. A., Barchetta, I., et al. (2019). Circulating miRNA-375 levels are increased in autoantibodies-positive first-degree relatives of type 1 diabetes patients. *Acta Diabetol.* doi: 10.1007/s00592-019-01297-7

Bluestone, J. A., Herold, K., and Eisenbarth, G. (2010). Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature* 464, 1293–1300. doi: 10.1038/nature08933

Carlos, D., Costa, F. R., Pereira, C. A., Rocha, F. A., Yaochite, J. N., Oliveira, G. G., et al. (2017). Mitochondrial DNA Activates the NLRP3 Inflammasome and Predisposes to Type 1 Diabetes in Murine Model. *Front. Immunol.* 8:164. doi: 10.3389/fimmu.2017.00164

Carpenter, S., Aiello, D., Atianand, M. K., Ricci, E. P., Gandhi, P., Hall, L. L., et al. (2013). A long noncoding RNA mediates both activation and repression of immune response genes. *Science* 341, 789–792. doi: 10.1126/science.1240925

Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *J. Mol. Diagnos.* 5, 73–81. doi: 10.1016/S1525-1578(10)60455-2

Chen, Y. G., Satpathy, A. T., and Chang, H. Y. (2017). Gene regulation in the immune system by long noncoding RNAs. *Nat. Immunol.* 18, 962–972. doi: 10.1038/ni.3771

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

Du, Z., Fei, T., Verhaak, R. G., Su, Z., Zhang, Y., Brown, M., et al. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* 20, 908–913. doi: 10.1038/nsmb.2591

Fourlanos, S., Varney, M. D., Tait, B. D., Morahan, G., Honeyman, M. C., Colman, P. G., et al. (2008). The rising incidence of type 1 diabetes is accounted for by cases with lower-risk human leukocyte antigen genotypes. *Diab. Care* 31, 1546–1549. doi: 10.2337/dc08-0239

Gibb, E. A., Brown, C. J., and Lam, W. L. (2011). The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer* 10:38. doi: 10.1186/1476-4598-10-38

Green, E. A., Gorelik, L., McGregor, C. M., Tran, E. H., and Flavell, R. A. (2003). $CD^{4+}CD^{25+}$ T regulatory cells control anti-islet CD^{8+} T cells through TGF-beta-TGF-beta receptor interactions in type 1 diabetes. *Proc. Natl. Acad. Sci. U. S. A.* 100, 10878–10883. doi: 10.1073/pnas.1834400100

Groop, L., and Pociot, F. (2014). Genetics of diabetes–are we missing the genes or the disease? *Mol. Cell. Endocrinol.* 382, 726–739. doi: 10.1016/j.mce.2013.04.002

Guay, C., Kruit, J. K., Rome, S., Menoud, V., Mulder, N. L., Jurdzinski, A., et al. (2019). Lymphocyte-derived exosomal MicroRNAs promote pancreatic beta cell death and may contribute to type 1 diabetes development. *Cell Metabol.* 29, 348–361. doi: 10.1016/j.cmet.2018.09.011

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., et al. (2013). Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucl. Acids Res.* 41:e35. doi: 10.1093/nar/gks967

Guo, X. L., Gao, L., Wang, Y., Chiu, D. K. Y., Wang, B. B., Deng, Y., et al. (2018). Large-scale investigation of long noncoding RNA secondary structures in human and mouse. *Curr. Bioinform.* 13, 450–460. doi: 10.2174/1574893613666180118111019

Hu, G., Tang, Q., Sharma, S., Yu, F., Escobar, T. M., Muljo, S. A., et al. (2013). Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat. Immunol.* 14, 1190–1198. doi: 10.1038/ni.2712

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Prot.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, Y., Liu, N., Wang, J. P., Wang, Y. Q., Yu, X. L., Bin Wang, Z., et al. (2012). Regulatory long non-coding RNA and its functions. *J. Physiol. Biochem.* 68, 611–618. doi: 10.1007/s13105-012-0166-y

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

Jiang, H., and Wong, W. H. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 24, 2395–2396. doi: 10.1093/bioinformatics/btn429

Jin, F., Wang, N., Zhu, Y., You, L., Wang, L., De, W., et al. (2017). Downregulation of long noncoding RNA Gas5 affects cell cycle and insulin

secretion in mouse pancreatic beta cells. *Cell. Physiol. Biochem.* 43, 2062–2073. doi: 10.1159/000484191

Kung, J. T., Colognori, D., and Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics* 193, 651–669. doi: 10.1534/genetics.112.146704

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. Nucl. Acids* 17, 337–346. doi: 10.1016/j.omtn.2019.05.028

Lennon, G. P., Bettini, M., Burton, A. R., Vincent, E., Arnold, P. Y., Santamaria, P., et al. (2009). T cell islet accumulation in type 1 diabetes is a tightly regulated, cell-autonomous event. *Immunity* 31, 643–653. doi: 10.1016/j.immuni.2009.07.008

Levy, H., Wang, X., Kaldunski, M., Jia, S., Kramer, J., Pavletich, S. J., et al. (2012). Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes. *Genes Immun.* 13, 593–604. doi: 10.1038/gene.2012.41

Liao, Z. J., Li, D. P., Wang, X. R., Li, L. S., and Zou, Q. (2018). Cancer diagnosis through IsomiR expression with machine learning method. *Curr. Bioinform.* 13, 57–63. doi: 10.2174/1574893611666160609081155

Liu, Y., Zhang, R., Qiu, F., Li, K., Zhou, Y., Shang, D., et al. (2015). Construction of a lncRNA-PCG bipartite network and identification of cancer-related lncRNAs: a case study in prostate cancer. *Mol. BioSyst.* 11, 384–393. doi: 10.1039/C4MB00439F

Lu, W., Huang, S. Y., Su, L., Zhao, B. X., and Miao, J. Y. (2016). Long noncoding RNA LOC100129973 suppresses apoptosis by targeting miR-4707-5p and miR-4767 in vascular endothelial cells. *Sci. Rep.* 6:21620. doi: 10.1038/srep21620

Luo, G., Liu, D., Huang, C., Wang, M., Xiao, X., Zeng, F., et al. (2017). LncRNA GAS5 inhibits cellular proliferation by targeting P27(Kip1). *Mol. Cancer Res.* 15, 789–799. doi: 10.1158/1541-7786.MCR-16-0331

Lv, H., Zhang, Z. M., Li, S. H., Tan, J. X., Chen, W., and Lin, H. (2019). Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform.* 982–995. doi: 10.1093/bib/bbz048

Mirza, A. H., Kaur, S., and Pociot, F. (2017). Long non-coding RNAs as novel players in beta cell function and type 1 diabetes. *Hum. Genom.* 11:17. doi: 10.1186/s40246-017-0113-7

Moran, I., Akerman, I., van de Bunt, M., Xie, R., Benazra, M., Nammo, T., et al. (2012). Human beta cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metabol.* 16, 435–448. doi: 10.1016/j.cmet.2012.08.010

Motterle, A., Gattesco, S., Caille, D., Meda, P., and Regazzi, R. (2015). Involvement of long non-coding RNAs in beta cell failure at the onset of type 1 diabetes in NOD mice. *Diabetologia* 58, 1827–1835. doi: 10.1007/s00125-015-3641-5

Petersmann, A., Nauck, M., Muller-Wieland, D., Kerner, W., Muller, U. A., Landgraf, R., et al. (2018). Definition, classification and diagnosis of diabetes mellitus. *Exp. Clin. Endocrinol. Diab.* 126, 406–410. doi: 10.1055/a-0584-6223

Polychronakos, C., and Li, Q. (2011). Understanding type 1 diabetes through genetics: advances and prospects. *Nat. Rev. Genet.* 12, 781–792. doi: 10.1038/nrg3069

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell* 136, 629–641. doi: 10.1016/j.cell.2009.02.006

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics*. 1–9. doi: 10.1093/bioinformatics/bty508

Sun, J., Zhou, M., Mao, Z. T., Hao, D. P., Wang, Z. Z., and Li, C. X. (2013). Systematic analysis of genomic organization and structure of long non-coding RNAs in the human genome. *FEBS Lett.* 587, 976–982. doi: 10.1016/j.febslet.2013.02.036

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, 957–64. doi: 10.7150/ijbs.24174

Wang, W., Wang, C., Ding, X. Q., Pan, Y., Gu, T. T., Wang, M. X., et al. (2013). Quercetin and allopurinol reduce liver thioredoxin-interacting protein to alleviate inflammation and lipid accumulation in diabetic rats. *Br. J. Pharmacol.* 169, 1352–1371. doi: 10.1111/bph.12226

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens. *J. Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004

Yang, M., Ye, L., Wang, B., Gao, J., Liu, R., Hong, J., et al. (2015). Decreased miR-146 expression in peripheral blood mononuclear cells is correlated with ongoing islet autoimmunity in type 1 diabetes patients 1miR-146. *J. Diab.* 7, 158–165. doi: 10.1111/1753-0407.12163

Yao, Y. H., Li, X. H., Geng, L. L., Nan, X. Y., Qi, Z. H., and Liao, B. (2018). Recent progress in long noncoding RNAs prediction. *Curr. Bioinform.* 13, 344–351. doi: 10.2174/1574893612666170905153933

You, L., Wang, N., Yin, D., Wang, L., Jin, F., Zhu, Y., et al. (2016). Downregulation of long noncoding RNA Meg3 affects insulin synthesis and secretion in mouse pancreatic beta cells. *J. Cell. Physiol.* 231, 852–862. doi: 10.1002/jcp.25175

Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., et al. (2017). RNALocate: a resource for RNA subcellular localizations. *Nucl. Acids Res.* 45, D135–D138. doi: 10.1093/nar/gkw728

Zhou, M., Guo, M., He, D., Wang, X., Cui, Y., Yang, H., et al. (2015). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J. Translat. Med.* 13:231. doi: 10.1186/s12967-015-0556-3

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

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. *Know. Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* 9:515. doi: 10.3389/fgene.2018.00515

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 © 2020 Geng, Zhang and Cheng. 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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 02 September 2020
doi: 10.3389/fbioe.2020.00679

IGF2BP3 May Contributes to Lung Tumorigenesis by Regulating the Alternative Splicing of PKM

*Huang Xueqing, Zhang Jun, Jiang Yueqiang, Liao Xin, Hu Liya, Fang Yuanyuan, Zhang Yuting, Zeng Hao, Wu Hua, Liu Jian and Yin Tiejun**

Department of Geriatrics, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

RNA binding proteins (RBPs) play a key role in genome regulation. Here we report the post-transcript regulation of IGF2BP3, which belongs to the insulin-like growth factor 2 mRNA binding protein family. We used iRIP-seq and RNA-seq to analyze the transcript regulation and alternative splicing on IGF2BP3 treated with overexpression cells and control. Overexpressed IGF2BP3 has broadly increased genes expression which involved in G-protein coupled receptor signaling pathway, positive regulation of cell proliferation, and signal transduction. IGF2BP3 regulated alternative splicing of multiple genes mainly clustered at response to hypoxia, negative regulation of transcription, and embryonic development. This study first provides alternative splicing analysis on transcription level of IGF2BP3 regulation, which laid the foundation for later research on IGF2BP3 critical functions.

Keywords: lung tumorigenesis, IGF2BP3, iRIP-seq, alternative splicing, cancer therapy 3, IGF2BP3/IMP3

OPEN ACCESS

Edited by:
Xiaotian Zhang,
Van Andel Institute (VAI),
United States

Reviewed by:
Xiaobin Zheng,
Carnegie Institution for Science (CIS),
United States
Diogo F. Troggian Veiga,
The Jackson Laboratory for Genomic Medicine, United States

***Correspondence:**
Yin Tiejun
tjyin@tjh.tjmu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *06 January 2020*
Accepted: *01 June 2020*
Published: *02 September 2020*

Citation:
Xueqing H, Jun Z, Yueqiang J, Xin L, Liya H, Yuanyuan F, Yuting Z, Hao Z, Hua W, Jian L and Tiejun Y (2020) IGF2BP3 May Contributes to Lung Tumorigenesis by Regulating the Alternative Splicing of PKM. Front. Bioeng. Biotechnol. 8:679. doi: 10.3389/fbioe.2020.00679

INTRODUCTION

Cancer is a complex disease caused by the malfunction of the cells regulation which is led by genetic and epigenetic mutations (Moore et al., 2018). Simply put, cancer is the abnormal growth of cells. Cancer can originate from any organ or body structure and consists of tiny cells that lose growth (Roy and Saikia, 2016). The basis of cancer hallmarks are genome unstable, which generate the genetic diversity and inflammation (Hanahan and Weinberg, 2011). Gene mutations cause cancer to be known to the public, but epigenetics is not. Epigenetic is a gene expression mutation that can be inherited but independent of DNA sequence (Kelly and Issa, 2017). Post-transcriptional gene regulation is also essential for maintaining cellular metabolism, coordinating RNA maturation, transport, stability, and degradation (Schmiedel et al., 2016).

A growing body of evidence shows that IGF2BP3 has shown its promising value on cancer therapy (Lochhead et al., 2012; Hsu et al., 2015). Numerous cancers have been identified the overexpression of IGF2BP3 including lung cancer (Findeis-Hosey and Xu, 2012), breast cancer (Scott et al., 2018), colorectal cancer (Wei et al., 2017; Xu et al., 2019), hepatocellular carcinoma (Shaalan et al., 2018), pancreatic cancer (Chen et al., 2019), glioblastoma and ovarian clear cell carcinoma (Bi et al., 2016; Dutoit et al., 2018). IGF2BP3 is a carcinoembryonic protein that is highly expressed during embryogenesis, lowly expressed in adult tissues, and re-expressed in malignant tissues (Palanichamy et al., 2016). In addition it's the key role in fetal and adult hemoglobin

expression (Tangprasittipap et al., 2018). It has been identified that IGF2BP3 can inhibit miRNA-3614 maturation thereby increased TRIM25 expression and promotes the cell proliferation of breast cancer (Wang et al., 2019). Studies showed that IGF2BP3 can induce cell proliferation and invasion through post-transcriptional regulation of IGF2 and promotes lung tumorigenesis via attenuating P53 stability which give us a hint to do more research on it (Taniuchi et al., 2014; Zhao et al., 2017). In some cases, overexpression is associated with a worse prognosis or a later stage disease (Shantha Kumara et al., 2015).

Although its function in cancer haven't been told clearly yet, it has been investigated in other fields. IGF2BP3 (also known as IMP3) is a member of the insulin-like growth factor 2 mRNA binding protein family (Schmiedel et al., 2016). IGF2BP3 is a 580 amino acid protein which has two RNA recognition motifs and four K homology domains, which was encoded by a 4350-bp mRNA transcript generated by the IGF2BP3 gene on chromosome 7p11.5 (Shantha Kumara et al., 2015). RNA binding protein (often abbreviated as RBP), as a crucial molecule, involves in almost all stages of post-transcriptional regulation, thus determining the destiny and function of each transcript in the cell and ensuring cell homeostasis (Lukong et al., 2008; Pereira et al., 2017). RBP, can be classified by the ability of binding to all transcripts (usually by common RNA elements) or recognizing specific transcripts (by specific motifs) (Rissland, 2017). Over the past few years, RBPs have been discovered serve as a hub molecule and identified to be hallmarks in more than one cancer (Lukong et al., 2008; Gerstberger et al., 2014; Brinegar and Cooper, 2016). RBPs also. RBPs involved in all aspects of RNA metabolism and play a pivotal role in cancer through a wide range of mechanisms, including genomic alterations, and post-transcriptional control (Fu and Ares, 2014; Mitchell and Parker, 2014; Brinegar and Cooper, 2016). Therefore, understanding the mechanisms by which RBP regulate cancer development is critical to reducing its morbidity and mortality. Since alternative splicing (AS) is a mechanism of post-transcriptional RNA processing that affords a significant evolutionary superiority by producing multiple mRNAs during pre-mRNA maturation, resulting in the production of multiple proteins with different functions from a single gene (Le et al., 2015; Brander et al., 2018; Liu et al., 2018). Accordingly, we hypothesized IGF2BP3 may also play a role in alternative splicing.

Herein, we aimed at the potential role of IGF2BP3 in mRNA binding and alternative splicing in A549 cells and its regulatory functions. We used RIP-seq technology to capture IGF2BP3-bound RNAs and their interaction sites. The results indicated IGF2BP3 has a good capability on RNA binding and strongly associates with mRNAs more at CDS regions than at 3′UTRs. In line with these, we performed high-throughput RNA sequencing (RNA-seq) for overexpressed IGF2BP3 and control cells to investigate the impact of IGF2BP3 on gene expression levels. The results showed that there are distinct changes in transcript profiles and IGF2BP3 strongly associates with mRNAs more at 5′UTR regions than at 3′UTRs. Then we delve into the role of IGF2BP3 in alternative splicing. Totally, we identified 11,418 alternative splicing events (p-value cutoff <0.05), including 6,804 novel alternative splicing events (RASEs) and 61 IGF2BP3-bound and –regulated alternative splicing genes were identified. We also found the alternative splicing pattern of the PKM gene, a kinase of pyruvate and its splicing pattern highly expressed in lung cancer cells (Mi et al., 2017), which is regulated by IGF2BP3. Taken together, our work clearly identified a complex genome-wide IGF2BP3-RNA interaction map in human lung cancer cell line and indicated that IGF2BP3 binds to CDS regions and regulates multiple alternative splicing events, which will help to understand the regulatory mechanisms of IGF2BP3 at the pre-RNA splicing level and provide insight into the role of IGF2BP3 in multiple biological processes.

RESULTS

IGF2BP3 Expression Level Was Significantly Increased in Multiple Cancer Types

According to previous studies, IGF2BP3 is highly expressed in numerous cancer tissues, such as esophageal cancer, lung cancer, prostate cancer, gastric cancer, and colorectal malignancies (Shantha Kumara et al., 2015; Zhang et al., 2017). At the tissue level, previous study has shown IGF2BP3 is higher expressed in squamous cell carcinoma and adenocarcinoma (Zhao et al., 2017). So we first studied the expression level of IGF2BP3 in different cancer types through the Oncomine database (**Figure 1A**). In the grid in the center of the table, blue indicates low expression of IGF2BP3 in the corresponding tumor, red indicates high expression, and gray indicates no data. And the numbers in the table represent the number of studies that meet the screening criteria which P-value is 1E-5 and the fold change is 3. Then, we found that IGF2BP3 is highly expressed in lung cancer tissues and low in normal lung tissues (**Figures 1B,C**). Taken together, these results indicate that IGF2BP3 is highly expressed in multiple cancer tissues, especially in lung squamous cell carcinoma and lung adenocarcinoma.

IGF2BP3 Strongly Associates With mRNAs More at 5′UTR Regions and 3′UTR Regions

To our knowledge, a classic RBP is evaluated by its ability of binding and the amount of bound-reads, as well as its efficiency. For these reasons, we attempt to explore IGF2BP3's binding ability by using RNA Immunoprecipitation (RIP) method. We used the improved RIP and high throughput sequencing approach (iRIP-seq, see "Materials and Methods for detail information) to identify the transcripts which are interacting with IGF2BP3 in A549 cells. As shown in **Figure 2A**, we can see that the quality control of iRIP-sequence is good. Then, we utilized Hi-seq 2000 platform to sequence the cDNA libraries of RNAs from anti-IGF2BP3 and IgG immunoprecipitates. After removing the reads of low-quality and adaptor sequences, IGF2BP3 and IgG immunoprecipitates totally recovered 27,872,842 and 27,850,918 reads. It turned out that there were about 39.58% reads can align when we use these data to locate to the GRCh38 genome by using the Tophat2 (**Supplementary Table S1**). Due to the

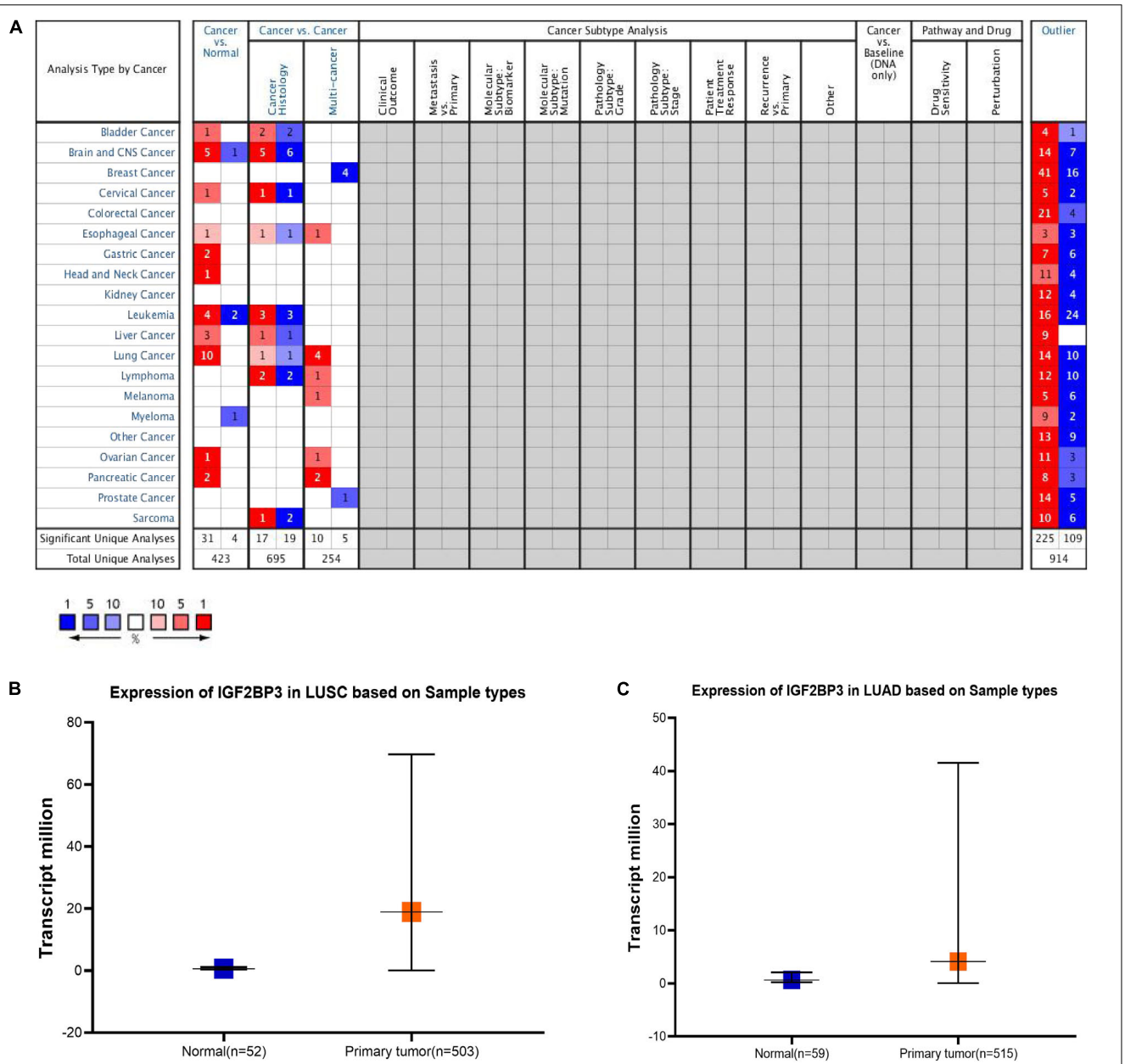


FIGURE 1 | IGF2BP3 expression level was significantly increased in multiple cancer types. **(A)** Different expression of IGF2BP3 in multiple cancer types. The data was downloaded from Oncomine website. **(B,C)** Plot of the expression of IGF2BP3 in normal and LUAD and LUSC samples. The data was downloaded from TCGA website.

existence of broken adaptor and primer sequences, there were a lot of unaligned reads. Most anti-IGF2BP3 reads can be uniquely mapped, however, many fewer IgG reads were either uniquely mapped or mapped.

To gain insight into the particularity of IGF2BP3 pull-downs, IgG was set as the blank control of immunoprecipitation. To assure the reliability of the results, we conducted two reduplicative experiments. As shown in **Figure 2B**, the correlation of IGF2BP3 samples are high and obvious different with the control. It is worth noting that the results of the two tests are almost identical, indicating that the sample is reproducible. Collectively, these results demonstrate the ability of IGF2BP3 samples to bind to RNA has a good specificity. Usually, in CLIP-seq protocols, ribonuclease was used to develop short-length reads from the sites that protected by RBPs, whose function is to digest the immunoprecipitated RNAs from the RBP-RNA complexes (Xue et al., 2009). But in RIP protocols, there was no RNase digestion step and doesn't recover intact transcripts. Surprisingly, the results showed that when the distribution of uniquely mapped anti-IGF2BP3 reads were potted on the whole

A High-quality clean reads of RIP-seq

Sample	Raw	Clean	Raw base	Clean base	Unique Tag	Q30	DUP
IGF2BP3-IP-1	28496270	24584658 (86.27%)	4.28 G	2.51G (58.64%)	4496073 (18.29%)	96.29%	94.67%
IGF2BP3-IP-2	31818188	31161027 (97.93%)	4.77 G	3.53G (74.00%)	9951293 (31.94%)	96.70%	85.85%
IGF2BP3-input-1	31480662	28556308 (90.71%)	4.72 G	2.64G (55.93%)	9014580 (31.57%)	96.20%	90.60%
IGF2BP3-input-2	28959586	27145528 (93.74%)	4.34 G	2.55G (58.76%)	7605534 (28.02%)	95.77%	91.99%

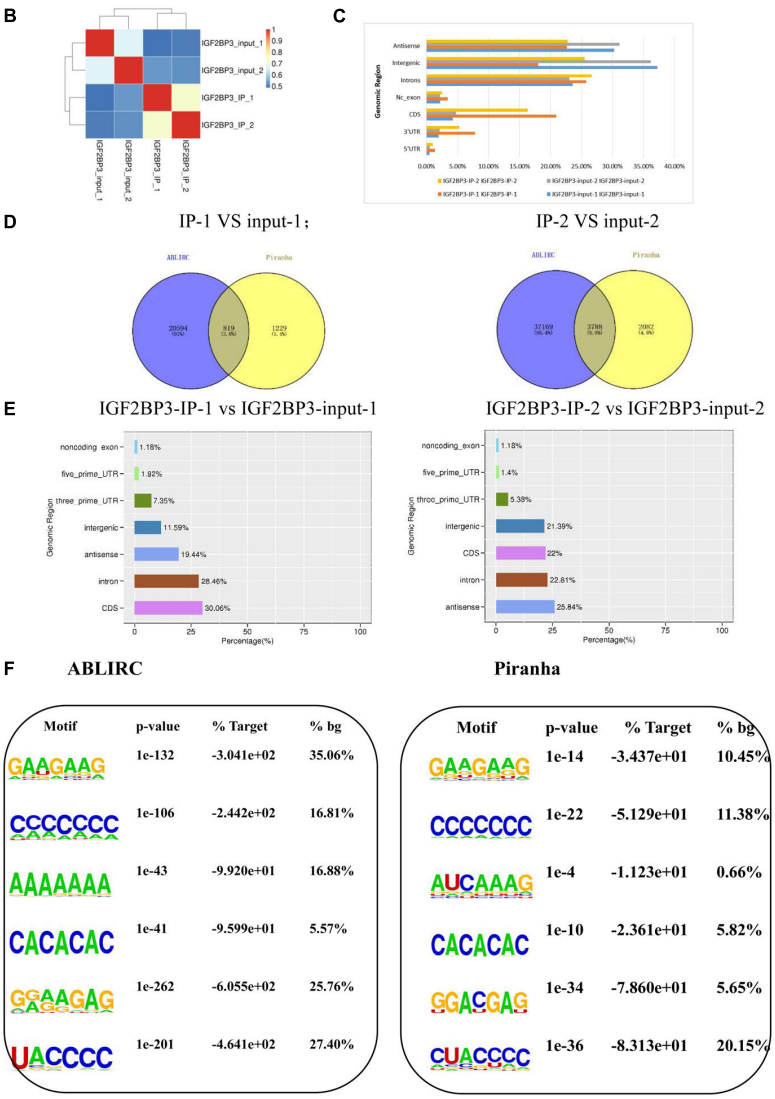


FIGURE 2 | IGF2BP3 strongly associates with mRNAs more at CDS regions than at 3′UTRs. **(A)** The form shows the quality control of our iRIP-sequence is good. (1) Raw data: the number of original sequences transformed from the original image data obtained by sequencing through base calling; (2) Clean reads: the raw reads are stripped of the adapter sequence, and the number of valid sequences obtained after low-quality bases is used for subsequent analysis; (3) Raw base: the count the number of bases it contains, based on the number and length of raw reads, in G; (4) Clean base: according to the number and length of clean reads, count the number of bases it contains, in G; (5) Unique tag: unique tag, the number of non-repeating reads and its proportion of clean reads; (6) Q30: Proportion of bases whose sequencing error rate is less than 0.1%; (7) DUP: duplication level. The ratio of duplicate reads to total reads. **(B)** Heat map showing the hierarchically clustered Pearson correlation matrix resulting from comparing the transcript expression values for control and IGF2BP3 IP samples. **(C)** Bar plot of the genomic region distribution of the control and uniquely mapped IGF2BP3 IP reads. **(D)** The peak distribution across reference genomic region of the IGF2BP3 IP and control group. **(E)** Venn diagram analysis from the comparative result of ABLIRC and Piranha peak calling methods. **(F)** Extracted IGF2BP3 peaks motifs using ABLIRC or Piranha.

human genome, the RIP-seq reads were much higher enriched in 5′UTR regions than in 3′UTR regions, and also enriched in CDS regions (**Figure 2C**).

From what we know, a gene with a low transcript abundance is less likely to be caught by an RNA-binding protein during IP (Chi et al., 2009). Recent days, we noticed that there is a software tool, ABLIRC, which can be used to recover the IGF2BP3 binding sites from the RIP-seq reads. Based on the results identified by the ABLIRC algorithm, we observed 77.36% sense peaks and 22.64% anti-sense peaks from IGF2BP3-associated RNAs and control (**Figure 2D**). After filtered the overlapping peaks of IGF2BP3 and IgG samples, 11,238 genes were distributed by a total of 38,608 (sense and antisense) peaks. The number of intronic peak in IGF2BP3 sample is in significantly contrast to the IgG sample. To verify these conclusions, we used Piranha, a published software to identify the RNA-protein interaction sites from the data of high-throughput sequencing, to call IGF2BP3-bound peaks. As shown in **Figure 2E**, Piranha average validated 2,303 peaks from ABLIRC averagely, representing 7.39% and 58.18% of ABLIRC and Piranha peaks, correspondingly. Furthermore, we found that 5′ss motif, 3′ss motif, and GA-rich motif were high enriched in Piranha peaks (**Figure 2F**), which was likely with those in ABLIRC peaks.

In summary, we successfully extracted the binding motif of IGF2BP3 RNA-binding protein by ABLIRC algorithm, which will help us understand the comprehensive regulation of IGF2BP3-RNA interaction during gene expression of A549 cells.

IGF2BP3-Bound Genes Are Distributed in Various Pathways

We identified IGF2BP3 binding genes in A549 cells by ABLIRC. **Supplementary Table S2** lists a total of 11,238 genes were detected with IGF2BP3 binding signal by our iRIP-seq data. As shown in **Supplementary Table S3**, we can see the ranks of candidates of IGF2BP3 targets, such as KRT7, PTMS, PKM, and IFI6. In order to investigate more functions of IGF2BP3, we selected some IGF2BP3 binding targets for RIP-PCR validation. According to the results of the quantitative RIP-PCR, we found that all candidate RNA targets were significantly enriched with anti-IGF2BP3 immunoprecipitates relative to the IgG control group (**Figure 3A**). For example, BTF3 relatives to IgG immunoprecipitates, which have been identified as previous IGF2BP3s' targets, were observed significantly enriched in IGF2BP3 immunoprecipitates. And we also found that HNRNPDL RNA was just enriched in one group of IGF2BP3 immunoprecipitate. Furthermore, we noticed that PKM RNA, which can generate a marker gene PKM2 for lung cancer (Rzechonek et al., 2017), is efficiently enriched in IGF2BP3 immunoprecipitates. We then assigned the identified peak genes of IGF2BP3 targets to the pathways in the Gene Ontology database and KEGG database, resulting in 867 GO biological processes and 289 KEGG biochemical pathways (**Supplementary Dataset 1**). The top 10 pathways are shown in **Figures 3B,C**. As shown in **Figure 3B**, IGF2BP3 involved in gene expression, nuclear mRNA splicing which via spliceosome and RNA splicing activities. On the other hand, based on KEGG annotation, we find that the spliceosome pathway ranks first and IGF2BP3 targets also involved in the processes like focal adhesion, proteoglycans in cancer, pathways in cancer, and RNA transport. Therefore we hypothesis IGF2BP3 has a strong regulation on the alternative splicing process.

IGF2BP3 Regulates Multiple Alternative Splicing Events

As the research progressed, we decided to explore the potential function of IGF2BP3 in A549 cells by using transcriptome sequencing (RNA-seq) with two biological replicates. We first examine the quality of data which shows in the **Supplementary Table S4**. As shown in **Figure 4A**, quantitative PCR and Western blot experiments both confirmed that the expression of IGF2BP3 was significantly increased in transfected A549 cells. The genes expressed differentially was identified by using edgeR (Robinson et al., 2010). Interestingly, we identified that among all expressed genes, there are 611 up-regulated and 605 down-regulated genes, which indicate that IGF2BP3 regulates gene transcription broadly (**Figure 4B**). We can found the details of the differentially expressed genes (DEGs) in **Supplementary Table S5**. As shown in the heatmap analysis based on the expression patterns of DEGs which in RNA-seq samples, there are a high consistency of the IGF2BP3-regulated transcription in both sets (**Figure 4C**). To further characterize the potential biological functions of these DEGs, DEGs resulted from whole expressed genes were subjected to GO and KEGG database. As shown in **Figure 4D**, shows several pathways which up-regulated genes annotated with categories of GO, including G-protein coupled receptor signaling pathway, positive regulation of cell proliferation and signal transduction. In line with these, the up-regulated genes annotated with KEGG categories in A549 cells are associated with retinol metabolism, neuroactive ligand-receptor interaction and Ras signaling pathway (**Figure 4E**).

Since one key aim of our study was to investigate the role of IGF2BP3 on alternative splicing (AS) regulation, thus we used transcriptome sequencing to analyze the AS events of IGF2BP3-dependent in A549 cells. The first transcriptome was generated by sequencing 47,749,031 in the original data, there is 44.38% is the junction reads. After analysis by the ABLAS tool, 9745 alternative splicing events (ASE) were obtained. These results enlarged the role of IGF2BP3 on the regulation of alternative splicing. In line with these, we performed a deeper transcriptome sequencing to see the distribution sites of alternative splicing events. **Figure 4F** showed that 121 AS events which changed significantly have been detected. Among them, a lot of splicing events belonged to Intron R (intron retention, 54), A5SS (alternative 5′ splice site, 15), ES (exon skipping, 15), and A3SS (alternative 3′ splice site, 14) categories. The other IGF2BP3-bound and IGF2BP3–affected splicing events included 3pMXE (mutually exclusive exons, 2), 5pMXE (mutually exclusive 5′ UTRs, 3), and Cassette Exon (number: 12). Taken together, we can know that Intron R, A5SS, ES are main splicing events regulated by IGF2BP3, which indicated that IGF2BP3 regulates alternative splicing events all over in A549 cells.

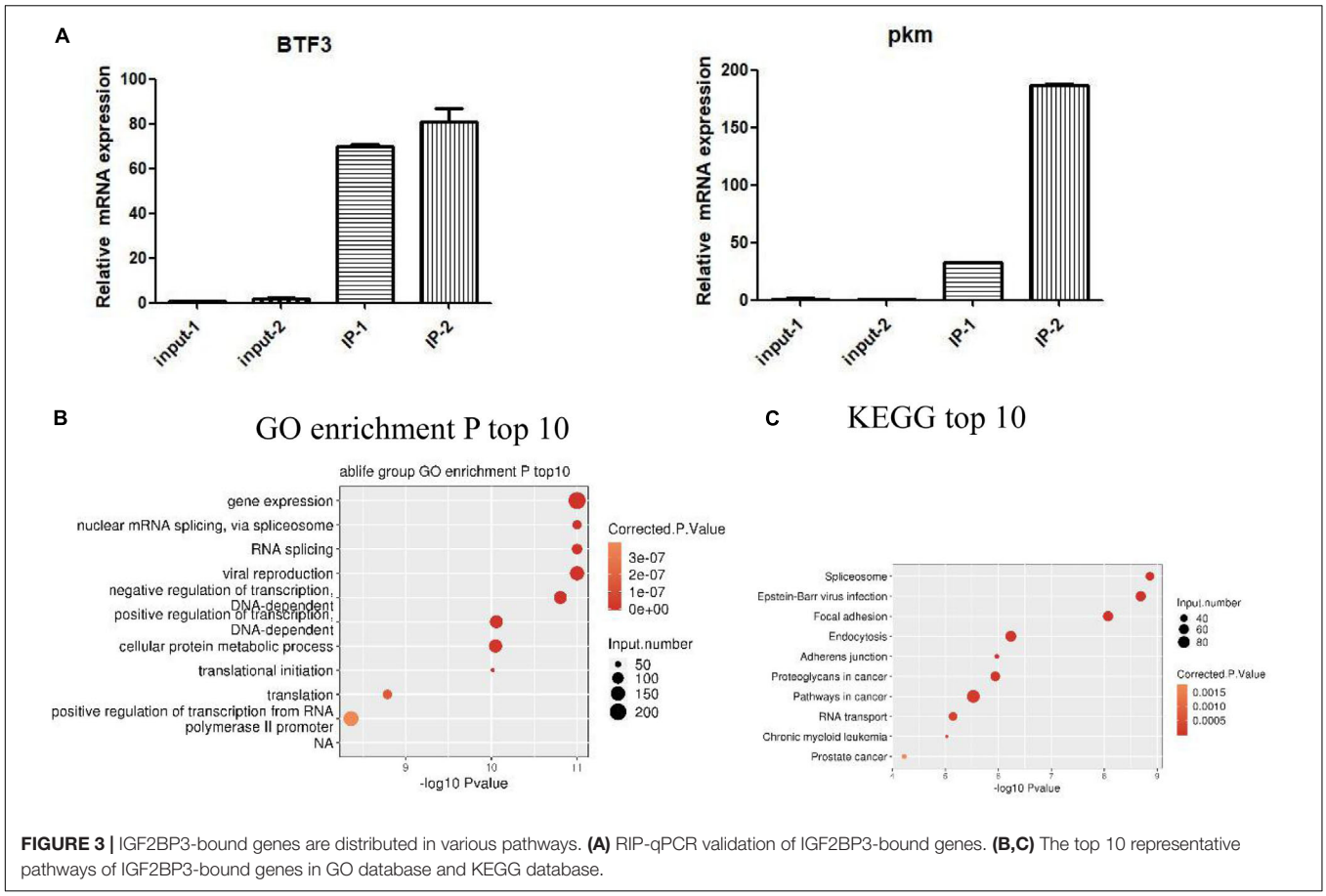


FIGURE 3 | IGF2BP3-bound genes are distributed in various pathways. **(A)** RIP-qPCR validation of IGF2BP3-bound genes. **(B,C)** The top 10 representative pathways of IGF2BP3-bound genes in GO database and KEGG database.

To further investigate the relationship between IGF2BP3 binding and the regulation of alternative splicing, we performed overlap analysis of the DEGs from the overexpressed RNA-seq analysis and the peak genes from the iRIP-seq analysis. The results showed that there were 61 overlap genes. Next, we performed functional clustering of genes with both binding and alternative splicing. The results showed that the above genes are mainly clustered at response to hypoxia, negative regulation of transcription, and embryonic development (**Figure 4G**).

IGF2BP3 Administered the Alternative Splicing of PKM

To directly assess the role of IGF2BP3 in these alternative splicing events, we undertook an RT-PCR experiment to examine the changes of the type of alternative splicing between control and A549 cells with overexpressed IGF2BP3. The inclusion/exclusion (In/Ex) ratio was used to measure the quantification of AS pattern. **Figure 5A** and **Supplementary Datasets 2,3** shows some representative affected AS events which has two alternative splicing patterns, including intron retention and alternative 5′splice site. Among those AS genes, PKM and BTF3 came into our notice, whose dysregulation associated with cell proliferation (Dayton et al., 2016). In this study, we first used RIP-seq protocol to analyze the interaction between IGF2BP3 and PKM pre-mRNA. As shown in the **Figure 5** (left panel), there are a large amount of IGF2BP3-bound peaks located in the exonic position. These two experiments' results both implied that PKM is a direct target of IGF2BP3 protein in A549 cells.

To gain insight into the regulation of alternative splicing events of PKM and BTF3 whether regulated by IGF2BP3-bound, we used Q-PCR to investigate the changes of ASEs in A549 cells with overexpressed IGF2BP3. As shown in **Figure 5B**, the ASEs of PKM is showing a decrease trend, whereas BTF3 showed an increase. Taken together, our results indicated that in A549 cells alternative splicing of PKM and BTF3 are regulated by IGF2BP3 directly.

DISCUSSION

Being a significant RBP, IGF2BP3 is involved in the process of mRNA metabolism and contributes to the cell proliferation and invasion in various cancers (Samanta et al., 2012; Taniuchi et al., 2014; Hsu et al., 2015; Palanichamy et al., 2016; Zhang et al., 2017; Xu et al., 2019). Moreover, multiple evidences proved that IGF2BP3 is relate to the processes of cancer development and prognosis (Lochhead et al., 2012; Cao et al., 2018). Given these foundations, some researches have been performed to comprehensively understand the functions of IGF2BP3 (Cao et al., 2018; Mancarella et al., 2018; Wang et al., 2019).

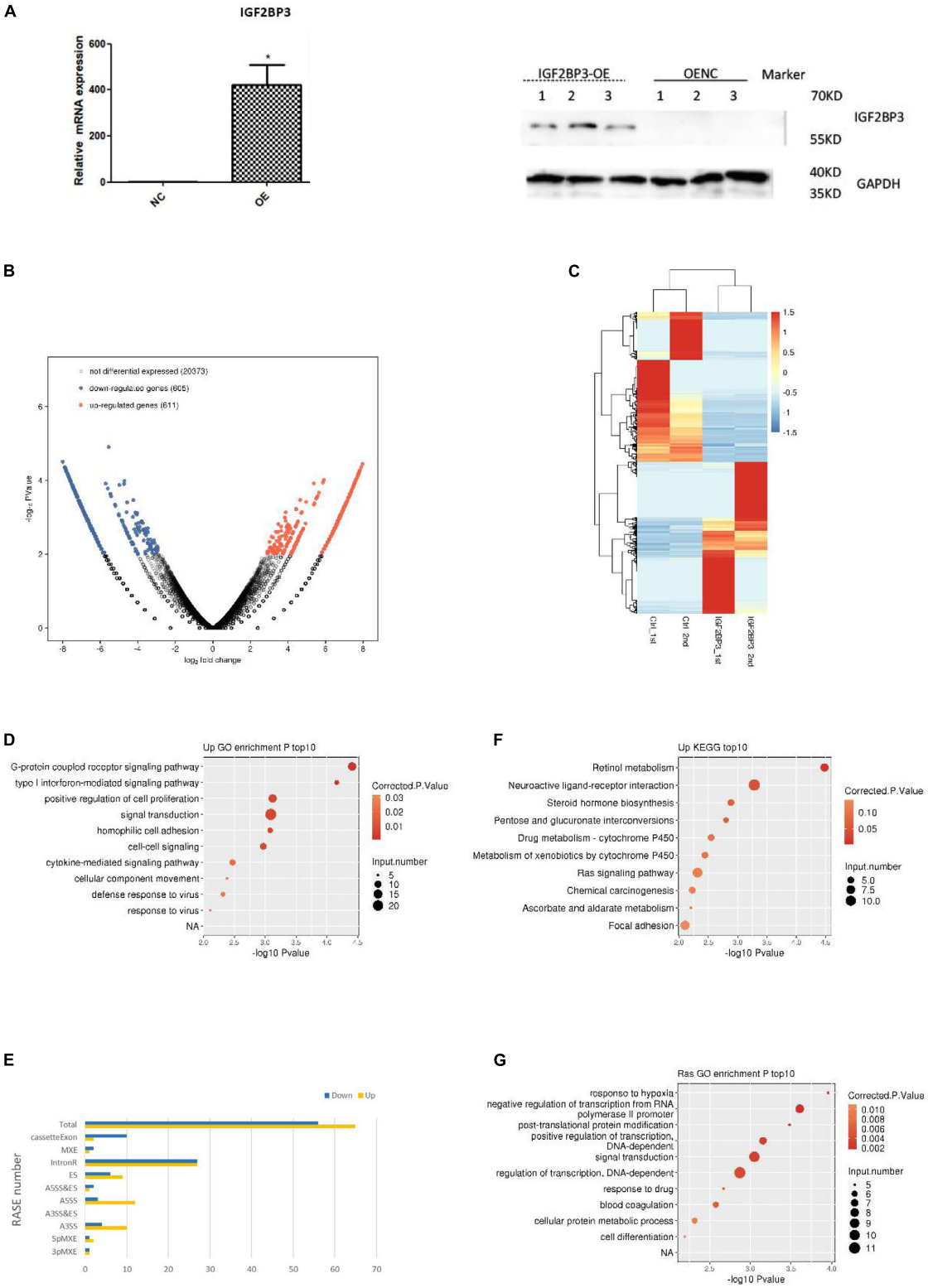


FIGURE 4 | IGF2BP3 regulates multiple alternative splicing events. **(A)** IGF2BP3 expression on mRNA and protein level in A549 cells after transiently transfected with IGF2BP3 specific overexpression RNA or control vector, as determined by qRT-PCR and Western blot analysis. **(B)** Total number of differentially expressed genes. **(C)** Hierarchical clustering of DEGs in overexpressed and control samples. Expression values (FPKM) are log2-transformed and then median-centered by each gene. **(D,E)** The top 10 representative GO Biological Process terms of IGF2BP3-regulated genes. **(F)** Classification of different AS types regulated by IGF2BP3 protein. **(G)** The top 10 GO Pathways for gene aggregation of differential AS events.

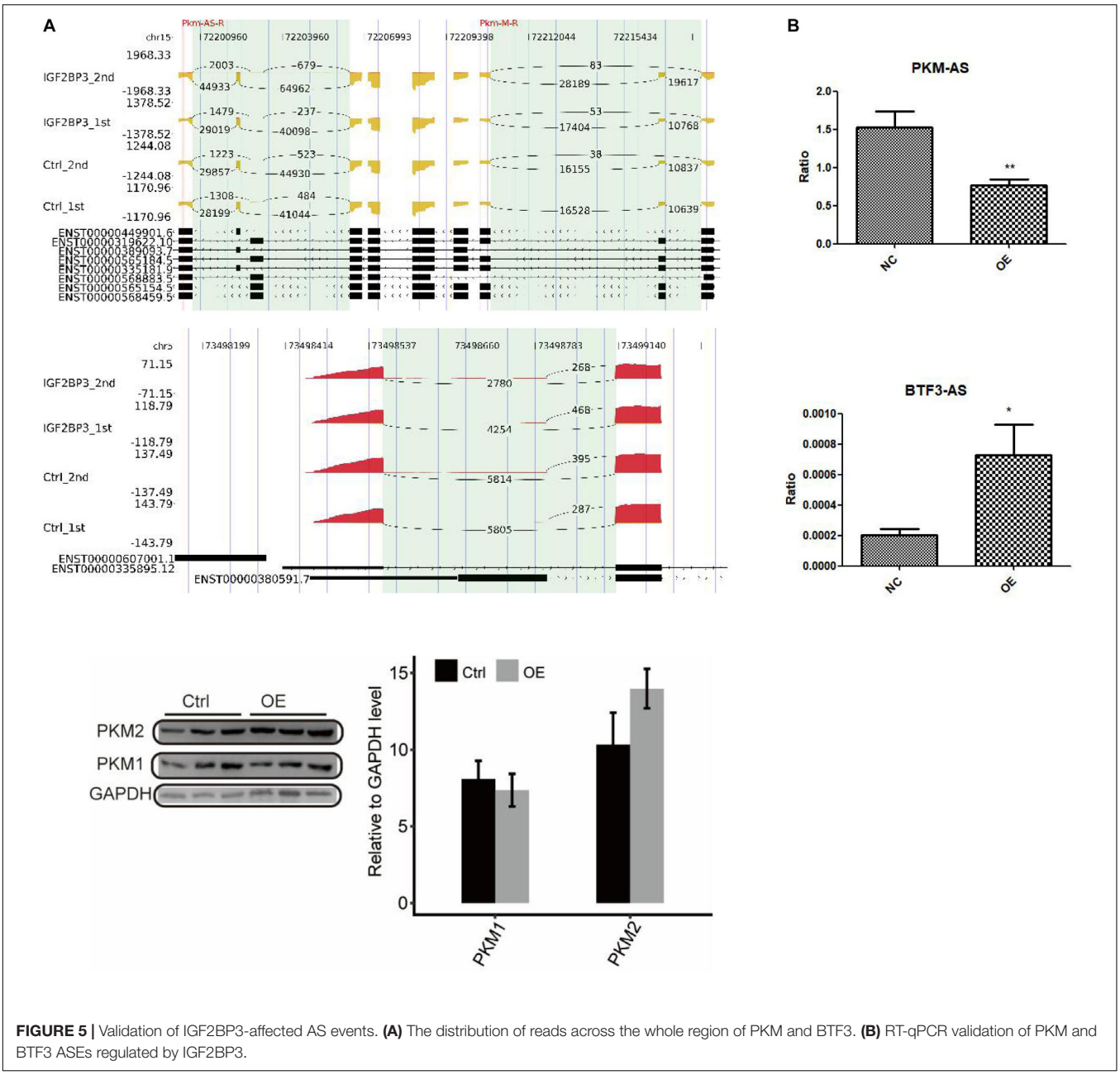


FIGURE 5 | Validation of IGF2BP3-affected AS events. **(A)** The distribution of reads across the whole region of PKM and BTF3. **(B)** RT-qPCR validation of PKM and BTF3 ASEs regulated by IGF2BP3.

Herein, we used iRIP-seq to uncover the mysteries of interactions between IGF2BP3 and RNA in human cancer cell line. The results of RIP-seq showed that sequencing reads of IGF2BP3-associated RNAs were mostly mapped to intron region, CDS region and intergenic region, which indicated IGF2BP3 has functional RNA targets *in vivo*. Of note, about 21.175% reads were allocated on intergenic region, which generate microRNA and other kinds of non-coding RNA. This result provide us a new hint that IGF2BP3 might participate in the process of non-coding RNA metabolism, which has never been reported and would be worthy to further study.

In here, as for the analysis of our iRIP-seq results, we utilized the ABLIRC to analyze the abundant peaks of IGF2BP3, which showed a significant fraction of IGF2BP3 binding peaks were enriched at 5′UTR region, 3′UTR region, and CDS. To be specific, IGF2BP3 might regulate the splice junction via the recognition of 3′ UTR region and 5′UTR region in A549 cells. The recognition of 3′UTR region and 5′UTR region in pre-mRNA are essential to the alternative splicing events. By using the iRIP-PCR, we verified that BTF3, PKM, and HNRNPDL are IGF2BP3-bound targets. From the GO and KEGG results we've got, IGF2BP3 involved in various important pathways, including gene expression, RNA splicing activities, and spliceosome pathway which ranks first. Therefore we speculated that IGF2BP3 can regulate alternative splicing in A549 cells. These observations might broad the understanding of the mechanism in the pre-mRNA alternative

splicing regulation of IGF2BP3. We then used RNA-seq to analyze the alternative splicing events regulated by IGF2BP3. We have identified 4614 RASE from 11418 alternative splicing events. We noticed that almost half of the pre-mRNA targets are regulated via directly binding to the IGF2BP3 in A549 cells. Therefore, we concluded that IGF2BP3 involved in different kinds of alternative splicing events regulation, such as NFKB2, LUC7L, BTF3, and PKM.

In the end, we speculated that PKM pre-mRNA may act as a new target of IGF2BP3. Our results indicated that the overexpression of IGF2BP3 has a regulation on the alternative splicing of PKM and BTF3. Very importantly, PKM is universally expressed in cancer and play a pivotal role in maintaining the metabolic program during cancer progression (Yang and Lu, 2013). In addition, PKM2 has been viewed as a potential target of lung cancer and its activity has been found higher in NSCLC patients than in normal subjects (Mi et al., 2017; Rzechonek et al., 2017).

We know that the final rate-limiting step in glycolysis is regulated by pyruvate kinase (PK) and catalyzes the transfer of phosphate groups from phosphoenolpyruvate (PEP) to adenosine diphosphate (Yang and Lu, 2013). As we know, PKM2 is an M2 isoform of PK, together with PKM1, is encoded by PKM. PKM pre-mRNA is spliced alternatively by the poly pyrimidine-tract binding (PTB) protein splicing factors and the heterogeneous nuclear ribonucleoproteins (hnRNPs) A1/2. As a result, PKM1 and PKM2 are generated by the exclusion of exon 9 and the inclusion of exon 10. Through the process of retaining exon 10, PKM2 owns unique properties which is important in the reprogramming of cell metabolism (Wong et al., 2015). PKM2 can be detected in normal tissues such as lung, liver, and kidney and also can expressed in cells with a high rate of nucleic synthesis. The expression of PKM2 is regulated at multiple levels by the regulation of DNA methylation, pre-mRNA splicing of PKM and post-translational modifications of the PKM2 protein (Yang and Lu, 2015).

In most cancer cells, the expression of PKM2 is increased and the expression of PKM2 in cancer cells affects aerobic glycolysis (Dayton et al., 2016). This phenomenon known as Warburg effect which tell the special way of energy production of tumor cells is very special: most tumor cells pass relatively low yields of glycolic acid to itself, whereas this kind of mechanism of action neither require oxygen nor mitochondria involvement (Koppenol et al., 2011). Recent studies showed that the replacement from PKM2 to PKM1 has been verified the ability to suppress aerobic glycolysis and tumor growth. PKM2 also can regulate tumor formation and growth by acting in gene transcription (Guo et al., 2017). It has also been proved that the splicing of PKM2 exist in drug resistance-pancreatic ductal adenocarcinoma (Calabretta et al., 2016). These evidence positions PKM2 as a promising target for cancer therapy. Aggregating our results, our observation might offer a new evidence about the mechanisms of how IGF2BP3 having an impact on lung cancer.

For the first time, we have demonstrated IGF2BP3 regulation of alternative splicing by successfully applied RIP-seq and RNA-seq in A549 cell line. Although we have identified the direct regulation in alternative splicing between IGF2BP3 and PKM, the exact mechanism by which IGF2BP3 regulate the alternative splicing of PKM has not been completely elucidated, which needs further investigations. Collectively, our work paves the way for seek the new biological function in the progress of tumorigenesis and provide new clues of the target therapy in lung cancer.

MATERIALS AND METHODS

Cell Culture and Transfections

A549 cells were cultured under standard conditions with Dulbecco's modified Eagle's medium (DMEM) with 10% fetal bovine serum (FBS), 100 μg/mL streptomycin, and 100 U/mL penicillin.

According to the manufacturer's instructions, the transfection of A549 cells with an IGF2BP3-overexpressing plasmid was performed using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, United States). After 48 h, q-PCR and western blot assay were performed to confirm the overexpression of IGF2BP3 mRNA and protein expression level (Tu et al., 2019).

Western Blotting Analysis

We loaded protein samples into 10% or 12% SDS-PAGE gels depending on molecular weight and transferred them onto 0.45 mm PVDF membranes. Then the PVDF membranes were blocked with 5% skim milk (in a buffer containing 10 mM Tris, pH 8.0, 150 mM NaCl, 0.05% Tween 20) for an hour. Next we incubated overnight with primary antibody at 4°C and later incubated with horseradish peroxidase-conjugated secondary antibody for 1 h at room temperature. Then, through the chemiluminescence, membranes were visualized. We also have quantitated some of the WB bands by the software Image J. Antibodies: The following antibodies were purchased from commercial sources from ABlife Company.

iRIP-seq Library Preparation and Sequencing

The IGF2BP3-binding RNAs were isolated by using TRIzol (Invitrogen). Complementary DNA (cDNA) libraries were drew up with the ScriptSeq RNA-seq Library Preparation Kit (SSV21124, Illumina). The libraries were prepared on the basis of the manufacturer's instructions and used to the Illumina HiSeq X Ten system for 150 nt paired-end sequencing.

iRIP-seq Peak Calling Analysis

After we aligned reads onto the genome, only uniquely mapped reads were applied for the following analysis. "ABLIRC" strategy was applied to identify the binding regions of TTP on genome (Xia et al., 2017). At least 1 bp overlapped reads were clustered as peaks. For each gene, we used computational simulation to randomly produced reads with the same number and lengths as reads in peaks. The outputting reads were further mapped to the same genes to generate random max peak height from overlapping reads. The whole process was repeated for 500 times. All the observed peaks with heights

higher than those of random max peaks (p-value < 0.05) were selected. The IGF2BP3 and Input samples were analyzed by the simulation independently, and the IGF2BP3 peaks that have overlap with Input peaks were removed. The target genes of IGF2BP3 were finally determined by the peaks and the binding motifs of IGF2BP3 protein were called by Homer software (Heinz et al., 2010).

RIP-qPCR

We used TRIzol (Invitrogen) to extract total RNAs from the immunoprecipitate of IGF2BP3 based on the manufacturer's instructions. Random primer was used for the cDNA synthesis. To detect whether IGF2BP3 target genes were significantly and specifically enriched in the IGF2BP3 immunoprecipitate, we then used normal PCR as a validation. At the meantime, we used input RNA as a reference and performed quantitative RT-PCR as characterized to determine the relative level of specific RNAs in the IGF2BP3 immunoprecipitates and IgG (Zhao et al., 2010). Q-PCR data represents the mean values from at least three independent experiments. Several genes were selected for PCR-amplification both in IGF2BP3 and IgG immunoprecipitates.

RNA Extraction and High-Throughput Sequencing

Total RNA was extracted by the TRIZOL (Ambion) and was purified with two phenol-chloroform treatments later. In order to remove DNA, the purified RNA was treated with RQ1 DNase (RNase free) (Promega, Madison, WI, United States) and its quality and quantity were reassured by measuring the absorbance at 260 nm/280 nm (A260/A280) using Smartspec Plus (Bio-Rad, United States). The integrity of RNA was then confirmed by 1.5% agarose gel electrophoresis. 10 μg of the total RNA for each sample was used to preparing directional RNA-seq library. Ahead of that, the polyadenylated mRNAs were condensed with oligo (dT)-conjugated magnetic beads (Invitrogen, Carlsbad, CA, United States). Next, the concentrated mRNAs were iron fragmented at 95°C, end repaired and 5′ adaptor ligated with 5′ adaptor. In line with it, we performed reverse transcription (RT) with RT primer harboring 3′ adaptor sequence and randomized hexamer. Before they were used for sequencing, the purified cDNAs were amplified and stored at −80°C (Li et al., 2019). As reported by the manufacturer's instructions, the libraries were ready for high-throughput sequencing. We used Illumina HiSeq4000 system to collect data from 151-bp pair-end sequencing (ABlife Inc., Wuhan, China).

RNA-seq Raw Data Clean and Alignment

We first discarded raw reads containing more than 2-N bases. Then we trimmed adaptors and low quality bases from raw sequencing reads using FASTX-Toolkit (Version 0.0.13). The short reads less than 16 nt were dropped too. Following that, clean reads were aligned to the GRch38 genome by TopHat2 (Kim et al., 2013) with 4 mismatches. Uniquely mapped reads were applied to calculate reads number and FPKM value (FPKM represents fragments per kilobase and per million) for each gene.

Differentially Expressed Genes (DEG) Analysis

FPKM (paired-end fragments per kilobase of exon per million fragments mapped), was used to assess the expression level of genes. In order to screen out the DEGs, we used the software edgeR (Robinson et al., 2010), which was specifically used for analyzing the differential expression of genes by using raw RNA-seq reads.

So as to determine whether a gene was differentially expressed, we analyzed the results on the basis of the fold change (fold change ≥ 2 or ≤ 0.5) and false discovery rate (FDR < 0.05). To gain insight into the gene function and measure the functional category distribution frequency, Gene Ontology (GO) analyses and enriched KEGG pathway were analyzed using KOBAS 2.0 server (Xie et al., 2011). We used Hypergeometric test and Benjamini-Hochberg FDR controlling procedure to define the enrichment of each pathway (corrected p-value < 0.05).

Alternative Splicing Analysis

As described previously, the ABLas pipeline (Jin et al., 2017; Xia et al., 2017) was used to define and quantify the ASEs (alternative splicing events) and RASEs (regulated alternative splicing events) between the samples. In short, detection of seven types of canonical ASEs in each sample was on the basis of the splice junction reads. These ASEs were exon skipping (ES), cassette exon (cassette Exon, CE), alternative 3′splice site (A3SS), alternative 5′splice site (A5SS), mutual exclusive exon skipping (MXE), the MXE combined with alternative polyadenylation site (3pMXE), and with alternative 5′ promoter (5pMXE). Then, the significant p-value was calculated by fisher's exact test, with the model reads of samples and alternative reads as input data, separately. RASE ratio was defined by the changed ratio of alternatively spliced reads and constitutively spliced reads between compared samples. The RASE ratio >0.2 and p-value < 0.05 were set as the threshold for RASEs detection.

Reverse Transcription qPCR Affirmation of Alternative Splicing Events

To exemplify the validity of ASEs in A549 cells, quantitative reverse-transcription polymerase chain reaction (RT-qPCR) was used in some selected RASEs, and standardized with the reference gene GAPDH. The primers which were used for detecting the pre-mRNA splicing are shown in **Supplementary Table S6**. In order to quantitatively analyzing the two different splicing isoforms of a specific ASE using a qPCR approach, we arranged two pairs of primers to specifically amplify each of these two isoforms after the initial synthesis of the first strand cDNA using random primers. To achieve this specificity, we designed a primer which can pair the splice junction of the constitutive exon and alternative exon. The RNA samples used for RT-qPCR and for RNA-seq too. The PCR conditions included denaturing at 95°C for 10 min, and 40 cycles of denaturing at 95°C for 15 s, then annealing and extension at 60°C for 1 min.

PCR amplifications were performed in triplicate for control and IGF2BP3-OE samples, respectively.

Statistical Analysis

All the statistics were showed as mean ± standard deviation (SD) and processed by SPSS 16.0 statistical software (Chicago, IL, United States). All experiments were operated at least three times independently, and values of $P < 0.05$ were considered considerable.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

YT and ZJ conceived the project. HX, FY, ZY, and ZH implemented the experiments and analyzed the data. ZH, HX, ZJ, and JY prepared the data and performed literature search. WH, LJ, and HL guided the experimental plan and writing of the subject. HX wrote the manuscript. All authors approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00679/full#supplementary-material

REFERENCES

Bi, R., Shen, X., Zhang, W., Cheng, Y., Feng, Z., Cai, X., et al. (2016). Clear cell carcinomas of the ovary: a mono-institutional study of 73 cases in China with an analysis of the prognostic significance of clinicopathological parameters and IMP3 expression. *Diagn. Pathol.* 11:17.

Brander, D. M., Friedman, D. R., Volkheimer, A. D., Christensen, D. J., Rassenti, L. Z., Kipps, T. J., et al. (2018). SET alpha and SET beta mRNA isoforms in chronic lymphocytic leukaemia. *Br. J. Haematol.* 184, 605–615. doi: 10.1111/bjh.15677

Brinegar, A. E., and Cooper, T. A. (2016). Roles for RNA-binding proteins in development and disease. *Brain Res.* 1647, 1–8. doi: 10.1016/j.brainres.2016.02.050

Calabretta, S., Bielli, P., Passacantilli, I., Pilozzi, E., Fendrich, V., Capurso, G., et al. (2016). Modulation of PKM alternative splicing by PTBP1 promotes gemcitabine resistance in pancreatic cancer cells. *Oncogene* 35, 2031–2039. doi: 10.1038/onc.2015.270

Cao, J., Mu, Q., and Huang, H. (2018). The roles of insulin-like growth factor 2 mRNA-Binding protein 2 in Cancer and Cancer Stem Cells. *Stem Cells Int.* 2018:4217259.

Chen, H., Kong, Y., Yao, Q., Zhang, X., Fu, Y., Li, J., et al. (2019). Three hypomethylated genes were associated with poor overall survival in pancreatic cancer patients. *Aging* 11, 885–897. doi: 10.18632/aging.101785

Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* 460, 479–486. doi: 10.1038/nature08170

Dayton, T. L., Jacks, T., and Vander Heiden, M. G. (2016). PKM2, cancer metabolism, and the road ahead. *EMBO Rep.* 17, 1721–1730. doi: 10.15252/embr.201643300

Dutoit, V., Migliorini, D., Ranzanici, G., Marinari, E., Widmer, V., Lobrinus, J. A., et al. (2018). Antigenic expression and spontaneous immune responses support the use of a selected peptide set from the IMA950 glioblastoma vaccine for immunotherapy of grade II and III glioma. *Oncoimmunology* 7:e1391972. doi: 10.1080/2162402x.2017.1391972

Findeis-Hosey, J. J., and Xu, H. (2012). Insulin-like growth factor II-messenger RNA-binding protein-3 and lung cancer. *Biotech. Histochem.* 87, 24–29. doi: 10.3109/10520295.2011.591831

Fu, X.-D., and Ares, M. Jr. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Na. Rev. Genet.* 15, 689–701. doi: 10.1038/nrg3778

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. *Nat. Rev. Genet.* 15, 829–845. doi: 10.1038/nrg3813

Guo, C. Y., Yan, C., Luo, L., Goto, S., Urata, Y., Xu, J. J., et al. (2017). Enhanced expression of PKM2 associates with the biological properties of cancer stem cells from A549 human lung cancer cells. *Oncol. Rep.* 37, 2161–2166. doi: 10.3892/or.2017.5438

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

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Hsu, K. F., Shen, M. R., Huang, Y. F., Cheng, Y. M., Lin, S. H., Chow, N. H., et al. (2015). Overexpression of the RNA-binding proteins Lin28B and IGF2BP3 (IMP3) is associated with chemoresistance and poor disease outcome in ovarian cancer. *Br. J. Cancer* 113, 414–424. doi: 10.1038/bjc.2015.254

Jin, L., Li, G., Yu, D., Huang, W., Cheng, C., Liao, S., et al. (2017). Transcriptome analysis reveals the complexity of alternative splicing regulation in the fungus Verticillium dahliae. *BMC Genomics* 18:130. doi: 10.1186/s12864-017-3507-y

Kelly, A. D., and Issa, J.-P. J. (2017). The promise of epigenetic therapy: reprogramming the cancer epigenome. *Curr. Opin. Genet. Dev.* 42, 68–77. doi: 10.1016/j.gde.2017.03.015

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.

Koppenol, W. H., Bounds, P. L., and Dang, C. V. (2011). Otto Warburg's contributions to current concepts of cancer metabolism. *Nat. Rev. Cancer* 11, 325–337.

Le, K. Q., Prabhakar, B. S., Hong, W. J., and Li, L. C. (2015). Alternative splicing as a biomarker and potential target for drug discovery. *Acta Pharmacol. Sin.* 36, 1212–1218. doi: 10.1038/aps.2015.43

Li, R. Z., Hou, J., Wei, Y., Luo, X., Ye, Y., and Zhang, Y. (2019). hnRNPDL extensively regulates transcription and alternative splicing. *Gene* 687, 125–134. doi: 10.1016/j.gene.2018.11.026

Liu, F., Dai, M., Xu, Q., Zhu, X., Zhou, Y., Jiang, S., et al. (2018). SRSF10-mediated IL1RAP alternative splicing regulates cervical cancer oncogenesis via mIL1RAP-NF-κB-CD47 axis. *Oncogene* 37, 2394–2409. doi: 10.1038/s41388-017-0119-6

Lochhead, P., Imamura, Y., Morikawa, T., Kuchiba, A., Yamauchi, M., Liao, X., et al. (2012). Insulin-like growth factor 2 messenger RNA binding protein 3 (IGF2BP3) is a marker of unfavourable prognosis in colorectal cancer. *Eur. J,. Cancer* 48, 3405–3413. doi: 10.1016/j.ejca.2012.06.021

Lukong, K. E., Chang, K. W., Khandjian, E. W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends Genet.* 24, 416–425. doi: 10.1016/j.tig.2008.05.004

Mancarella, C., Pasello, M., Manara, M. C., Toracchio, L., Sciandra, E. F., Picci, P., et al. (2018). Insulin-like growth factor 2 mRNA-binding protein 3 influences sensitivity to Anti-IGF system agents through the translational regulation of IGF1R. *Front. Endocrinol.* 9:178. doi: 10.3389/fendo.2018.00178

Mi, Y., He, M., and Liu, B. (2017). [MiR-133b Affect the Proliferation and Drug Sensitivity in A549 Lung Cancer Stem Cells by Targeting PKM2]. *Zhongguo Fei Ai Za Zhi* 20, 376–381.

Mitchell, S. F., and Parker, R. (2014). Principles and Properties of Eukaryotic mRNPs. *Mol. Cell* 54, 547–558. doi: 10.1016/j.molcel.2014.04.033

Moore, S., Järvelin, A. I., Davis, I., Bond, G. L., and Castello, A. (2018). Expanding horizons: new roles for non-canonical RNA-binding proteins in cancer. *Curr. Opin. Genet. Dev.* 48, 112–120. doi: 10.1016/j.gde.2017.11.006

Palanichamy, J. K., Tran, T. M., Howard, J. M., Contreras, J. R., Fernando, T. R., Sterne-Weiler, T., et al. (2016). RNA-binding protein IGF2BP3 targeting of oncogenic transcripts promotes hematopoietic progenitor proliferation. *J. Clin. Invest.* 126, 1495–1511. doi: 10.1172/jci80046

Pereira, B., Billaud, M., and Almeida, R. (2017). RNA-Binding proteins in cancer: old players and new actors. *Trends Cancer* 3, 506–528. doi: 10.1016/j.trecan.2017.05.003

Rissland, O. S. (2017). The organization and regulation of mRNA-protein complexes. *Wiley Interdiscip,. Rev. RNA* 8:e1369. doi: 10.1002/wrna.1369

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, 139–140. doi: 10.1093/bioinformatics/btp616

Roy, P., and Saikia, B. (2016). Cancer and cure: a critical analysis. *Indian J. Cancer* 53, 441–442.

Rzechonek, A., Kaminska, A., Mamczur, P., Drapiewski, A., and Budzynski, W. (2017). Limited clinical significance of dimeric form of pyruvate kinase as a diagnostic and prognostic biomarker in non-small cell lung cancer. *Adv. Exp. Med. Biol.* 955, 51–57. doi: 10.1007/5584_2016_92

Samanta, S., Sharma, V. M., Khan, A., and Mercurio, A. M. (2012). Regulation of IMP3 by EGFR signaling and repression by ERβ: implications for triple-negative breast cancer. *Oncogene* 31, 4689–4697. doi: 10.1038/onc.2011.620

Schmiedel, D., Tai, J., Yamin, R., Berhani, O., Bauman, Y., and Mandelboim, O. (2016). The RNA binding protein IMP3 facilitates tumor immune escape by downregulating the stress-induced ligands ULPB2 and MICB. *eLife* 5:e13426. doi: 10.7554/eLife.13426

Scott, C. M., Wong, E. M., Joo, J. E., Dugue, P. A., Jung, C. H., O'Callaghan, N., et al. (2018). Genome-wide DNA methylation assessment of 'BRCA1-like' early-onset breast cancer: data from the australian breast cancer family registry. *Exp. Mol. Pathol.* 105, 404–410. doi: 10.1016/j.yexmp.2018.11.006

Shaalan, Y. M., Handoussa, H., Youness, R. A., Assal, R. A., El-Khatib, A. H., Linscheid, M. W., et al. (2018). Destabilizing the interplay between miR-1275 and IGF2BPs by Tamarix articulata and quercetin in hepatocellular carcinoma. *Nat. Prod. Res.* 32, 2217–2220. doi: 10.1080/14786419.2017.1366478

Shantha Kumara, H., Kirchoff, D., Caballero, O. L., Su, T., Ahmed, A., Herath, S. A., et al. (2015). Expression of the cancer testis antigen IGF2BP3 in colorectal cancers. IGF2BP3 holds promise as a specific immunotherapy target. *Oncoscience* 2, 607–614. doi: 10.18632/oncoscience.174

Tangprasittipap, A., Kaewprommal, P., Sripichai, O., Sathirapongsasuti, N., Satirapod, C., Shaw, P. J., et al. (2018). Comparison of gene expression profiles between human erythroid cells derived from fetal liver and adult peripheral blood. *PeerJ* 6:e5527. doi: 10.7717/peerj.5527

Taniuchi, K., Furihata, M., Hanazaki, K., Saito, M., and Saibara, T. (2014). IGF2BP3-mediated translation in cell protrusions promotes cell invasiveness and metastasis of pancreatic cancer. *Oncotarget* 5, 6832–6845. doi: 10.18632/oncotarget.2257

Tu, Y., Wu, X., Yu, F., Dang, J., Wang, J., Wei, Y., et al. (2019). Tristetraprolin specifically regulates the expression and alternative splicing of immune response genes in HeLa cells. *BMC Immunol.* 20:13. doi: 10.1186/s12865-019-0292-1

Wang, Z., Tong, D., Han, C., Zhao, Z., Wang, X., Jiang, T., et al. (2019). Blockade of miR-3614 maturation by IGF2BP3 increases TRIM25 expression and promotes breast cancer cell proliferation. *EBioMedicine* 41, 357–369. doi: 10.1016/j.ebiom.2018.12.061

Wei, Q., Zhou, H., Zhong, L., Shi, L., Liu, J., Yang, Q., et al. (2017). IMP3 expression in biopsy specimens as a diagnostic biomarker for colorectal cancer. *Hum. Pathol.* 64, 137–144. doi: 10.1016/j.humpath.2017.03.013

Wong, N., Ojo, D., Yan, J., and Tang, D. (2015). PKM2 contributes to cancer metabolism. *Cancer Lett.* 356(2 Pt A), 184–191. doi: 10.1016/j.canlet.2014.01.031

Xia, H., Chen, D., Wu, Q., Wu, G., Zhou, Y., Zhang, Y., et al. (2017). CELF1 preferentially binds to exon-intron boundary and regulates alternative splicing in HeLa cells. *Biochim. Biophys. Acta Gene Regul. Mech.* 1860, 911–921. doi: 10.1016/j.bbagrm.2017.07.004

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322.

Xu, W., Sheng, Y., Guo, Y., Huang, Z., Huang, Y., Wen, D., et al. (2019). Increased IGF2BP3 expression promotes the aggressive phenotypes of colorectal cancer cells in vitro and vivo. *J. Cell. Physiol.* 234, 18466–18479. doi: 10.1002/jcp.28483

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y. S., et al. (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell* 36, 996–1006. doi: 10.1016/j.molcel.2009.12.003

Yang, W., and Lu, Z. (2013). Regulation and function of pyruvate kinase M2 in cancer. *Cancer Lett.* 339, 153–158. doi: 10.1016/j.canlet.2013.06.008

Yang, W., and Lu, Z. (2015). Pyruvate kinase M2 at a glance. *J. Cell Sci.* 128, 1655–1660. doi: 10.1242/jcs.166629

Zhang, J., Ji, Q., Jiao, C., Ren, L., Zhao, Y., Chen, Y., et al. (2017). IGF2BP3 as a potential tissue marker for the diagnosis of esophageal high-grade intraepithelial neoplasia. *Onco Targets Ther.* 10, 3861–3866. doi: 10.2147/ott.s141179

Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., et al. (2010). Genome-wide Identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* 40, 939–953. doi: 10.1016/j.molcel.2010.12.011

Zhao, W., Lu, D., Liu, L., Cai, J., Zhou, Y., Yang, Y., et al. (2017). Insulin-like growth factor 2 mRNA binding protein 3 (IGF2BP3) promotes lung tumorigenesis via attenuating p53 stability. *Oncotarget* 8, 93672–93687. doi: 10.18632/oncotarget.21280

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 © 2020 Xueqing, Jun, Yueqiang, Xin, Liya, Yuanyuan, Yuting, Hao, Hua, Jian and Tiejun. 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.

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