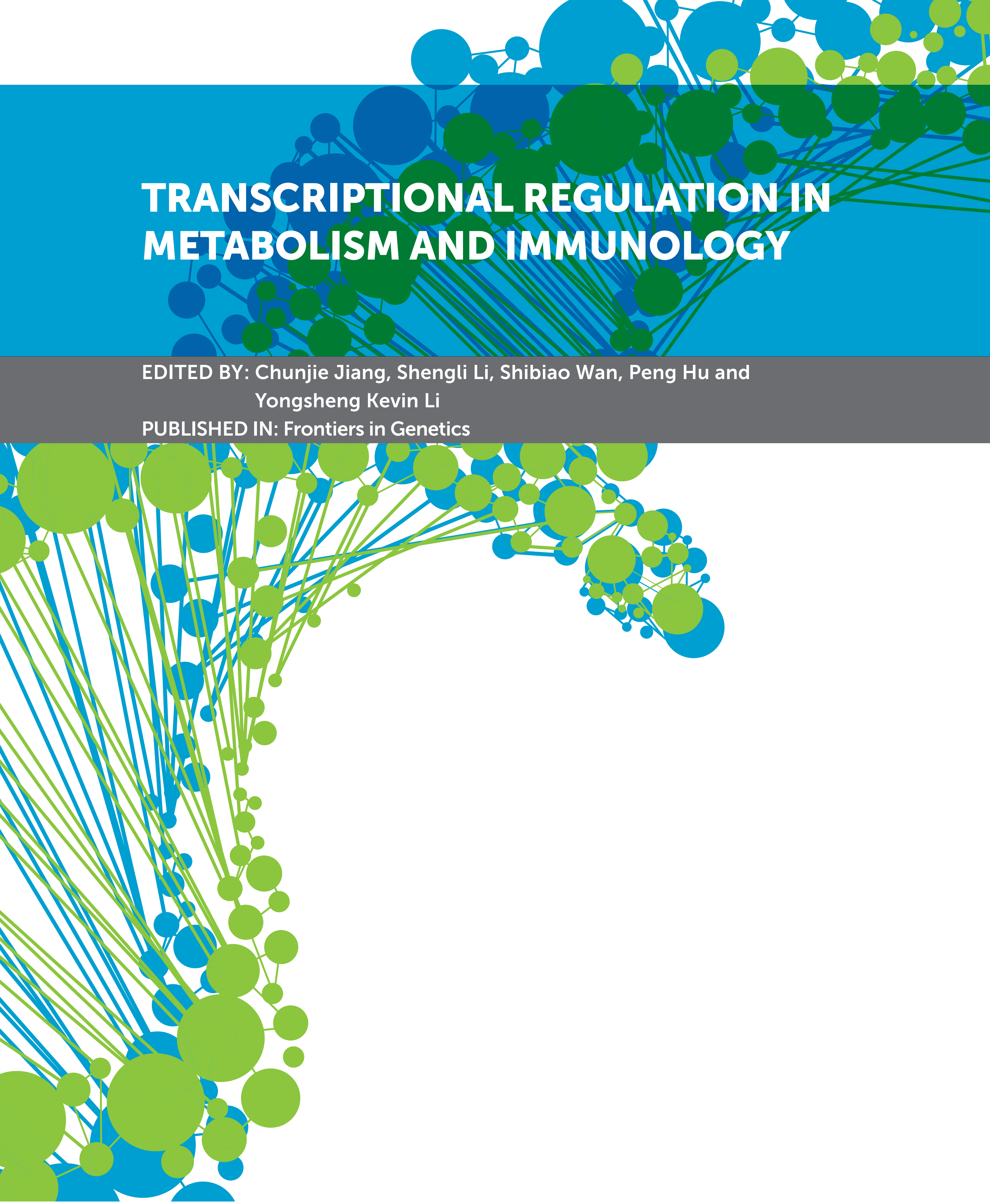
TRANSCRIPTIONAL REGULATION IN METABOLISM AND IMMUNOLOGY

EDITED BY: Chunjie Jiang, Shengli Li, Shibiao Wan, Peng Hu and
Yongsheng Kevin Li

PUBLISHED IN: Frontiers in Genetics



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-573-9
DOI 10.3389/978-2-88974-573-9

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

TRANSCRIPTIONAL REGULATION IN METABOLISM AND IMMUNOLOGY

Topic Editors:
Chunjie Jiang, University of Pennsylvania, United States
Shengli Li, Shanghai Jiao Tong University, China
Shibiao Wan, St. Jude Children's Research Hospital, United States
Peng Hu, University of Pennsylvania, United States
Yongsheng Kevin Li, Hainan Medical University, China

Citation: Jiang, C., Li, S., Wan, S., Hu, P., Li, Y. K., eds. (2022). Transcriptional Regulation in Metabolism and Immunology. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88974-573-9

Table of Contents

frontiers in Genetics

EDITORIAL
published: 02 February 2022
doi: 10.3389/fgene.2022.845697

Editorial: Transcriptional Regulation in Metabolism and Immunology

Chunjie Jiang[1], Shibiao Wan[2], Peng Hu[3], Yongsheng Li[4] and Shengli Li[5]**

[1]Department of Medicine, Division of Diabetes, Endocrinology and Metabolism, Baylor College of Medicine, Houston, TX, United States, [2]Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN, United States, [3]College of Fisheries and Life Science, Shanghai Ocean University, Shanghai, China, [4]Key Laboratory of Tropical Translational Medicine of Ministry of Education, College of Biomedical Information and Engineering, Hainan Medical University, Haikou, China, [5]Precision Research Center for Refractory Diseases, Institute for Clinical Research, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Keywords: transcriptional regulation, next-generation sequencing, transcription factor, metabolism, immunology

Editorial on the Research Topic

Transcriptional Regulation in Metabolism and Immunology

The regulation of transcription that converts DNA to RNA is a vital process in all living organisms to orchestrate gene activities (Weingarten-Gabbay and Segal, 2014; Cramer, 2019). Transcription factors (TFs) are important factors to orchestrate transcription by binding to specific DNA sequences to activate or repress wide repertoires of downstream target genes that control a wide variety of biological processes (Spitz and Furlong, 2012; Lambert et al., 2018), including metabolic and immune systems. A large number of TFs that play critical roles in regulating transcription in the metabolic and immune systems have been investigated and much has been learned about their mechanisms (Mansueto et al., 2017; Hosokawa and Rothenberg, 2021).

Metabolic homeostasis needs fine tuning to adapt to environmental stimuli, which largely depends on transcriptional-level regulation (Mouchiroud et al., 2014). Maintenance of energy homeostasis is critical in all cells, which is mainly perceived and regulated by the highly conserved AMP-activated protein kinase (AMPK) (Garcia and Shaw, 2017). AMPK has been shown to phosphorylate specific transcription factors, such as FOXO transcription factors, to restore energy balance and reprogram many metabolic progresses, including the metabolism of glucose, lipid, mTOR, and proteins. Nonalcoholic fatty liver disease (NAFLD) is the most prevalent liver disease worldwide, which may progress to fatal cirrhosis or hepatocellular carcinoma (Foulds et al., 2017). Exposure to endocrine-disrupting chemicals (EDCs) may increase the susceptibility to the development of NAFLD. Imbalance of hepatic lipid homeostasis may lead to the initiation and development of NAFLD. EDCs can recruit co-regulator proteins by physically binding nuclear receptors (NRs), and modulate the transcription of genes involved in hepatic lipid homeostasis.

Trigger of required immune response demands fine transcriptional regulation in cells of the immune system (Roy, 2019). Wu *et al.* applied single-cell RNA sequencing to investigate IL-4-induced I transcription in B cell differentiation (Wu et al., 2017). Their analysis revealed that the early transcription of Iε could induce class switching to IgE. Thus, the transcription regulation of Iε directs the early choice of IgE. In addition, various noncoding RNAs have been found to participate in the regulation of immune processes and immune cells, including circular RNAs and long noncoding RNAs (Hu W. et al., 2021; Fang et al., 2021).

This Research Topic is dedicated to publishing studies revealing the mechanisms of transcriptional regulation in metabolic and/or immune systems based on the data sets from next-generation sequencing and other state-of-art technologies, which will shed light on the deeper understanding of the underlying mechanisms. A total of 19 articles are included in this Research Topic.

OPEN ACCESS

Edited and reviewed by:
*Richard D. Emes,
University of Nottingham,
United Kingdom*

****Correspondence:***
*Chunjie Jiang
chunjie.jiang917@outlook.com
Shengli Li
shengli.li@shsmu.edu.cn*

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

***Received:** 30 December 2021*
***Accepted:** 17 January 2022*
***Published:** 02 February 2022*

Citation:
Jiang C, Wan S, Hu P, Li Y and Li S (2022) Editorial: Transcriptional Regulation in Metabolism and Immunology. Front. Genet. 13:845697. doi: 10.3389/fgene.2022.845697

Four papers contributed to the transcriptional regulation in metabolic system. Zhang *et al.* revealed five metabolism pathway-related circRNAs in prostate cancer (Zhang et al.). Cheng *et al.* found that alterations in lipid metabolism pathway are associated with prognosis of non-small-cell lung cancer patients that were treated with immune checkpoint inhibitors (Cheng et al.). One research performed systematic analysis of nuclear-encoded mitochondrial genes in hypertrophic cardiomyopathy, including the regulation of transcription factors (Tan et al.). Liu *et al.* examined the dysregulation of immune and metabolism-related RNAs in uterine corpus endometrial carcinoma (Liu and Qiu).

For the transcriptional regulation in immune system, two articles contributed to the transcriptional dysregulation in immune cells and their roles as biomarkers in diseases, including macrophage M2 cells (Wang et al.) and neutrophils (Qiu et al.). Several articles identified immune-related prognostic markers in human complex diseases, including stromal-immune scores (Liu et al.), lncRNAs (Wang et al.; Pang et al.; Zhao et al.), immune-related genes (Hu et al.; He et al.; Li et al.; Xu et al.), and transcriptional regulation factors (He et al.; Zhang et al.; Chen et al.).

In addition, the Research Topic also included two methodology articles, one is about a deep learning classifier for determining disease immune subtypes and related immunosuppression genes (Ning et al.), and the other is the comparisons of dimensionality reduction methods in single-cell transcriptomics data (Xiang et al.).

In conclusion, recent studies have precisely highlighted dysregulated TFs in specific contexts by adopting high throughput sequencing and other state-of-the-art technologies. These studies largely extended our current knowledge of the complexity of gene regulation circuitry in metabolism and immunology, and will facilitate further advancement.

AUTHOR CONTRIBUTIONS

SL and CJ wrote the manuscript with comments from all the other listed authors. All authors listed approved it for publication.

FUNDING

This work was supported by Shanghai General Hospital Startup Funding (02.06.01.20.06) and National Natural Science Foundation of China (32100517). This research was also supported in part by the National Cancer Institute grant P30 CA021765. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Cramer, P. (2019). Organization and Regulation of Gene Transcription. *Nature* 573, 45–54. doi:10.1038/s41586-019-1517-4

Fang, Z., Jiang, C., and Li, S. (2021). The Potential Regulatory Roles of Circular RNAs in Tumor Immunology and Immunotherapy. *Front. Immunol.* 11, 1–13. doi:10.3389/fimmu.2020.617583

Foulds, C. E., Treviño, L. S., York, B., and Walker, C. L. (2017). Endocrine-disrupting Chemicals and Fatty Liver Disease. *Nat. Rev. Endocrinol.* 13, 445–457. doi:10.1038/nrendo.2017.42

Garcia, D., and Shaw, R. J. (2017). AMPK: Mechanisms of Cellular Energy Sensing and Restoration of Metabolic Balance. *Mol. Cell* 66, 789–800. doi:10.1016/j.molcel.2017.05.032

Hosokawa, H., and Rothenberg, E. V. (2021). How Transcription Factors Drive Choice of the T Cell Fate. *Nat. Rev. Immunol.* 21, 162–176. doi:10.1038/s41577-020-00426-6

Hu, W., Wang, Y., Fang, Z., He, W., and Li, S. (2021b). Integrated Characterization of lncRNA-Immune Interactions in Prostate Cancer. *Front. Cell Dev. Biol.* 9, 1–12. doi:10.3389/fcell.2021.641891

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The Human Transcription Factors. *Cell* 172, 650–665. doi:10.1016/j.cell.2018.01.029

Mansueto, G., Armani, A., Viscomi, C., D'Orsi, L., De Cegli, R., Polishchuk, E. V., et al. (2017). Transcription Factor EB Controls Metabolic Flexibility during Exercise. *Cell Metab.* 25, 182–196. doi:10.1016/j.cmet.2016.11.003

Mouchiroud, L., Eichner, L. J., Shaw, R. J., and Auwerx, J. (2014). Transcriptional Coregulators: Fine-tuning Metabolism. *Cell Metab.* 20, 26–40. doi:10.1016/j.cmet.2014.03.027

Roy, A. L. (2019). Transcriptional Regulation in the Immune System: One Cell at a Time. *Front. Immunol.* 10, 1–8. doi:10.3389/fimmu.2019.01355

Spitz, F., and Furlong, E. E. M. (2012). Transcription Factors: From Enhancer Binding to Developmental Control. *Nat. Rev. Genet.* 13, 613–626. doi:10.1038/nrg3207

Weingarten-Gabbay, S., and Segal, E. (2014). The Grammar of Transcriptional Regulation. *Hum. Genet.* 133, 701–711. doi:10.1007/s00439-013-1413-1

Wu, Y. L., Stubbington, M. J. T., Daly, M., Teichmann, S. A., and Rada, C. (2017). Intrinsic Transcriptional Heterogeneity in B Cells Controls Early Class Switching to IgE. *J. Exp. Med.* 214, 183–196. doi:10.1084/jem.20161056

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 18 December 2020
doi: 10.3389/fgene.2020.616988

Hyperglycemia Decreases Epithelial Cell Proliferation and Attenuates Neutrophil Activity by Reducing ICAM-1 and LFA-1 Expression Levels

*Dongxu Qiu[1], Lei Zhang[1], Junkun Zhan[2], Qiong Yang[2], Hongliang Xiong[3], Weitong Hu[3], Qiao Ji[4] and Jiabing Huang[3]**

[1] Xiangya Hospital, Central South University, Changsha, China, [2] Department of Geriatrics, The Second Hospital of Xiangya, Hunan, China, [3] Department of Cardiology, The Second Affiliated Hospital of Nanchang University, Nanchang, China, [4] The Second Affiliated Hospital of Nanchang University, Nanchang, China

Delayed repair is a serious public health concern for diabetic populations. Intercellular adhesion molecule 1 (ICAM-1) and Lymphocyte function-associated antigen 1 (LFA-1) play important roles in orchestrating the repair process. However, little is known about their effects on endothelial cell (EC) proliferation and neutrophil activity in subjects with hyperglycemia (HG). We cultured ECs and performed a scratch-closure assay to determine the relationship between ICAM-1 and EC proliferation. Specific internally labeled bacteria were used to clarify the effects of ICAM-1 and LFA-1 on neutrophil phagocytosis. Transwell assay and fluorescence-activated cell sorting analysis evaluated the roles of ICAM-1 and LFA-1 in neutrophil recruitment. ICAM-1$^{+/+}$ and ICAM-1$^{-/-}$ mice were used to confirm the findings *in vivo*. The results demonstrated that HG decreased the expression of ICAM-1, which lead to the low proliferation of ECs. HG also attenuated neutrophil recruitment and phagocytosis by reducing the expression of ICAM-1 and LFA-1, which were strongly associated with the delayed repair.

Keywords: hyperglycemia, ICAM-1, LFA-1, neutrophil, phagocytosis

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Jing Li,
Hunan University, China
Shuyu Fu,
Wistar Institute, United States

***Correspondence:**
Jiabing Huang
Neurology1988@aliyun.com

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

Received: *13 October 2020*
Accepted: *23 November 2020*
Published: *18 December 2020*

Citation:
Qiu D, Zhang L, Zhan J, Yang Q, Xiong H, Hu W, Ji Q and Huang J (2020) Hyperglycemia Decreases Epithelial Cell Proliferation and Attenuates Neutrophil Activity by Reducing ICAM-1 and LFA-1 Expression Levels. Front. Genet. 11:616988. doi: 10.3389/fgene.2020.616988

INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by inappropriate hyperglycemia (HG) (American Diabetes, 2013). Uncontrolled HG can lead to a host of diabetic complications, including delayed injury repair, which is a serious public health concern for subjects with diabetes. The tendon injury repair process consists of four phases: coagulation, inflammation, granular tissue formation, and remodeling (Gosain and DiPietro, 2004; Falanga, 2005). All the phases rely strongly on cellular and metabolic components of the inflamed microenvironment. However, the diabetic injury microenvironment is hostile and characterized by markedly elevated levels of inflammatory cytokines, which contribute to the dysfunction of these components. Recent studies have shown that endothelial cell (EC) proliferation at sites of injury is crucial for injury repair. Specifically, intercellular adhesion molecule 1 (ICAM-1) plays a key role in EC proliferation (Nagaoka et al., 2000; Gay et al., 2011; Sumagin et al., 2016). ICAM-1 regulates EC permeability in inflamed tissues by inducing the activation of extracellular signal-regulated kinase 1/2 (ERK1/2) (Han et al., 2016). However, due to the complexity of the immune response under HG conditions, the potential effects of ICAM-1 on ECs remain poorly understood. In the present study, we cultured ECs and performed a scratch-closure assay to evaluate the effects of ICAM-1 on EC proliferation.

Bacteria at injury sites can cause tissue infection and generate biofilms, leading to delayed injury repair (Hurlow et al., 2015; Qiu et al., 2018, 2019). HG is considered the best culture condition for bacterial growth (Frykberg, 2002). HG increases pathogen accumulation. This in turn prevents the proliferation of keratinocytes and angiogenesis at the site of injury (Gosain and DiPietro, 2004; Bandyk, 2018). Both type 1 and type 2 diabetes cause HG, indicating a high risk of insufficient injury repair within the diabetic population. Neutrophils are the main leukocytes involved in the defense against invasion by exogenous pathogens (Everett and Mathioudakis, 2018). Enhanced recruitment of neutrophils promotes injury repair in subjects with HG. Lymphocyte function-associated antigen 1 (LFA-1) is an integrin that is mainly expressed on the surface of lymphocytes (Xingyuan et al., 2006). ICAM-1 and LFA-1 expression levels are critical for neutrophil trafficking into inflamed tissues (Basit et al., 2006). However, the effects of ICAM-1 and LFA-1 on neutrophil recruitment in subjects with HG remain poorly understood. Recent studies have strongly associated neutrophil phagocytosis with ICAM-1/LFA-1 interaction (Lefort and Ley, 2012; Woodfin et al., 2016).

In the present study we explored the effects of ICAM-1 and LFA-1 on neutrophil recruitment and phagocytosis under HG conditions *in vitro* and *in vivo*. The objectives of this study were to evaluate ICAM-1 expression as well as its involvement in EC proliferation and the combined effects of LFA-1 on neutrophil recruitment and phagocytosis under HG conditions. The findings provide new insights and will inform novel therapeutic approaches for the repair of diabetic injuries.

MATERIALS AND METHODS

Cell Culture

Injured tissue from C57BL/6 mice was harvested for EC isolation. The tissue was minced into pieces 0.3–0.4 mm in size. The pieces were enzymatically digested with trypsin and collagenase (Witkowska and Borawska, 2004). Neutrophils were freshly isolated from bone marrow. Overlying muscle and skin were removed from the tibia and femur, and the tissue was placed in Hank's Balanced Salt Solution (HBSS) buffer on ice until needed. Bone marrow tissue was flushed with fresh HBSS for 8 min using a 10 mL sterile syringe. After rinsing, a single-cell suspension was obtained by careful pipetting. ECs and neutrophils were isolated using a magnetic separator. ECs were characterized by CD105 and CD31. Neutrophils were characterized by CD45, Ly6G, and CD11b. ECs were cultured in 500 mL complete mouse endothelial cell medium with a kit (Cell Biologics Inc., Chicago, IL, United States) supplemented with 15% fetal bovine serum (FBS; Hyclone, Logan, UT, United States), 2 mM L-glutamine, 100 mg/mL heparin, 15 mg/mL EC growth supplement, 100 mg/mL streptomycin, and 100 U/mL penicillin. Cells were grown at 37.5°C in an atmosphere of 5% CO_2 and 95% relative humidity, and seeded in a wells of a 24-well culture plate at a density of 2×10^5 cells/well.

Scratch-Closure Assay

ECs were pre-treated with high (25 mM) or low (5 mM) glucose concentrations for 6 days. In some cases, anti-ICAM-1/LFA-1 neutralizing antibody (ab109361, ab52895; Abcam, Cambridge, MA, United States) was added to the culture medium and confluent monolayer cells were scraped off using a 200-μL pipette tip. For the *in vitro* assay, we gently removed the debris, cleaned the scratch border and replaced the volume with growth medium (Becker et al., 1991). To determine the number of ECs that had migrated into the scraped area, photographs were taken at various times and analyzed using NIS-Elements D image analysis software (Nikon, Tokyo, Japan).

EC Proliferation Assay

EC proliferation was detected using 5-ethynyl-2-deoxyuridine (EdU) with a Click-iT Cell Proliferation imaging kit (Thermo Fisher Scientific, Waltham, MA, United States). Briefly, the indicated cells were cultured in triplicate in 24-well plates for 24 h and were then treated with 50 μM of EdU for 2 h at 37°C. Then they were fixed in 4% formaldehyde for 10 min and permeabilized with 0.5% Triton X-100 for 10 min at room temperature, the cells were treated with 1 × Apollo reaction cocktail for 30 min. For *in vitro* analysis, ECs were pre-treated with a low (5 mM) or high (25 mM) concentration of glucose and incubated with anti-ICAM-1 neutralizing antibody (15 μg/mL) or isotype IgG as a control. At least six random fields per subgroup were measured in three parallel assays. The data are expressed as the percentage of all proliferating cells in a single field. Triplicate technical replicates were assigned to each group.

Transwell Migration Assay

Confluent ECs were continuously stimulated for 18 h with tumor necrosis factor-alpha (TNF-α) to induce EC activation prior to transmigration assays (Campos, 2012; Kolluru et al., 2012). Confluent neutrophils were inoculated into the upper chambers of the Transwell system. ECs were also added into the lower chamber of the device and incubated in fresh medium. In some cases, anti-ICAM-1 neutralizing antibody was added to the culture medium. Transwell inserts were incubated at 37.5°C in an atmosphere of 5% CO_2 and 95% relative humidity for 20 h. Cells that migrated to the lower side of the membrane were attached, fixed with 2% paraformaldehyde (PFA) and stained with 0.5% crystal violet. Cells at the upper side of the membrane were scraped off using a cotton swab. Digital images were obtained using a light microscope system.

Western Blotting

Skin injury tissue was isolated using the Mammalian Cell Lysis Kit (Sigma-Aldrich, St. Louis, MO, United States). Samples were adjusted to equal total protein amounts and transferred to polyvinylidene fluoride or polyvinylidene difluoride membranes. Membranes were blocked with 5% (wt/vol) blocking reagent (Roche, Basel, Switzerland) in Tris-buffered saline for 1 h. The blots were probed with rabbit monoclonal anti-ICAM-1/CD11a and β-actin antibodies (Thermo Fisher Scientific). Alkaline phosphatase conjugated to goat anti-mouse/rabbit IgG (Abcam) was added as the secondary antibody after incubation with the primary antibody.

Animal Model

ICAM-1^{+}/$^{+}$ and ICAM-1^{-}/$^{-}$ mice were obtained from the Jackson Laboratory (Bar Harbor, ME, United States). All mice were housed under specific pathogen-free conditions in full compliance with the Animal Use and Care Committee of Central South University, Changsha, Hunan Province, China. A type 1 diabetic model was induced by continuous low-dose streptozotocin (STZ) intraperitoneal injection (50 mg/kg; Sigma-Aldrich) for 5 days. The normal control (NG) was injected with an identical dose of phosphate-buffered saline (PBS). Mice were identified as diabetic based on a blood glucose level > 250 mg/dL. Skin injury was performed after the mice had maintained a diabetic status for longer than 3 weeks. Prior to surgery, mice were anaesthetized by intraperitoneal injection with a ketamine–xylazine solution (80 mg/kg ketamine, 5 mg/kg xylazine). We used a 3.0-mm biopsy punch to perform symmetrical full-thickness excisional injury on the skin. Mice were euthanized with CO_2 and injured tissue was collected 4 and 8 days after surgery. Seven mice per group were analyzed at each time point.

Histology and Immunofluorescence Staining

Collected samples were fixed with formalin (10%; Sigma-Aldrich) for 20 h at 4°C, followed by slow decalcification in 10% EDTA solution for 4 weeks. Each specimen was bisected evenly, and half of the tissues were embedded in paraffin blocks for histological analysis. Slices of 5 μm thickness were prepared for hematoxylin and eosin (H&E) staining. For ICAM-1 analysis, paraffin slides were subjected to immunofluorescence staining. Slides were incubated with an ICAM-1 monoclonal antibody (Thermo Fisher Scientific) at 4°C overnight. The slide was mounted with 4,6-diamidino-2-phenylindole (DAPI) for nuclear counterstaining. Histomorphometry of the injured tissue was performed using a Nikon digital camera coupled to a microscope, followed by analysis using the associated Nikon AR software.

Flow Cytometry Analysis

The tissue surrounding the injury edge was collected using a 4-mm punch and minced into pieces 0.1–0.2 mm in size. The pieces were transferred to conical tubes containing 5 mL digestion medium (collagenase type IV, DNase, and dispase II). The suspensions were transferred to a shaking incubator (200 rpm) at 37°C for 1 h after digestion. A 70 μm strainer was used to filter the suspended solution after shaking. The solution was centrifuged at 4°C and 400 × g for 8 min. The supernatant was removed, the pellets were resuspended in 150 μL washing buffer (3% FBS RPMI) and the cells were counted. Fluorescence-labeled murine monoclonal antibodies were obtained from BioLegend (San Diego, CA, United States) and eBioscience (San Diego, CA, United States). The isolate solution was dispensed in flow cytometry tubes (100 μL/tube). Anti-CD16/CD32 antibodies (Fc blocker; BioLegend) and 2c/100 μL cells were added for 10 min. A master mix containing CD45-Pacific Blue, CD11b allophycocyanin (APC) and Ly6G^{+} APC was created as a neutrophil panel. The mixture was centrifuged at 300 × g for 8 min at 4°C and the supernatant was gently removed. PBS was added to a volume of 200 μL and run the flow for these samples. To detect neutrophil phagocytic function, diabetic mice were intraperitoneally injected with lipopolysaccharide (LPS) to induce neutrophilia. Fluorescent zymosan-Texas-Red (ZymTR) or PBS was administered to the mice 8 and 16 h prior to tissue collection. The enzymatically digested injured tissue was analyzed by flow cytometry.

Neutrophil Phagocytosis Assay

Bacterial phagocytosis was induced as described previously (Habas and Shang, 2018), with some modifications. Briefly, a suspension of 5 × 10^{6} neutrophils/mL was co-cultured with Staphylococcus aureus labeled with carboxy fluorescein succinimides (CFSE; Thermo Fisher Scientific). Add 50 μL of HBSS to at least one tube to create a negative (i.e., no bacteria) control for flow cytometry gating. Mix solutions very gently by inverting tubes several times. Place tubes in an incubated oven and rotate very gently (∼5–10 rpm) for 10 min. Remove tubes from incubator and immediately place on ice to arrest the phagocytosis process. Immediately add 0.55 mL of cold 4% paraformaldehyde to each tube, mix gently by inverting tubes, and incubate on ice for 30 min. Rinse cells once with cold HBSS (no Ca/Mg) by centrifugation at 400 × g for 10 min. Resuspend cells in 0.2 mL of cold HBSS (no Ca/Mg). Measure cell-associated fluorescence by flow cytometry. Neutrophil-associated bacterias were evaluated by co-localizing CFSE-labeled S. aureus. Internalized bacteria and neutrophils associated with the labeled bacteria were counted using fluorescent microscopy.

Quantitative Real-Time Polymerase Chain Reaction (qPCR) Analysis

We performed qPCR analysis to detect the expression of ICAM-1 and LFA-1 (CD11a). Four copies per sample were analyzed and the results were averaged. The following primers were used for the PCR reactions: ICAM-1, forward primer: TTCAAGCTGAGCGACATTGG; reverse primer: CGCTC TGGGAACGAATACACA; matrix metalloproteinase-1 (MMP-1), forward primer: AGCTAGCTCAGGATGACATTGATG; reverse primer: GCCGATGGGCTGGACAG; MMP-2, forward primer: TGGCGATGGATACCCCTTT; reverse primer: TCCTCCCAAGGTCCATAGCTCAT and MMP-9, forward primer: CCTGGGCAGATTCCAAACCT; reverse primer: GCAACTCTTCCGAGTAGTTTCCAT.

Statistical Analyses

All data are expressed as mean ± standard deviation (SD). Differences were assessed using Student's t-test or paired one-way analysis of variance (ANOVA) using GraphPad Prism ver. 4.0 software (GraphPad, La Jolla, CA, United States). Statistical significance was indicated at $P < 0.05$.

RESULTS

HG Reduces ICAM-1 Expression

ECs increase the release of ICAM-1 during inflammation (van der Zijpp et al., 2003; Awla et al., 2011). However, little is known about the expression of ICAM-1 by ECs in subjects with HG. Since HG has been causally associated with endothelial dysfunction (Prokopowicz et al., 2012; Rada, 2019), we explored the release of ICAM-1 under hyperglycemic conditions. ECs were cultured and incubated for 16 h together with stimulation by the TNF-α pro-inflammatory cytokine. ICAM-1 expression was decreased in the HG group. However, no significant differences were detected among the non-activated counterparts (**Figure 1A**). To better assess ICAM-1 release by ECs, we repeated this assay using a Transwell system, which allowed us to measure levels of ICAM-1 in independent culture media. ECs were seeded as a monolayer on the Transwell interfaces. The total amount of ICAM-1 decreased in the basolateral chamber in the HG group. The rate of ICAM-1 increase was also lower in this group (**Figure 1B**). We inferred that MMPs are involved in ICAM-1 expression in response to ECs. To identify the effects of MMPs on the induction of ICAM-1 expression in the HG group, we focussed specifically on MMP-9, MMP-1, and MMP-2, which have been associated with ICAM-1 expression. In contrast to previous findings, no difference was detected in the levels of these MMPs between the HG and control groups (**Supplementary Figures S1A–C**). Based on these observations, HG appeared to reduce the release rate of ICAM-1 *in vitro*.

HG Decreases EC Proliferation by Reducing ICAM-1 Expression

Recent studies have demonstrated the critical role played by ICAM-1 in EC proliferation (Tamanini et al., 2003; Dragoni et al., 2017). However, little is known about the effects of ICAM-1 on EC proliferation under hyperglycemic conditions. Using NG and HG culture media with or without TNF-α stimulation, we observed decreased EC proliferation in the HG culture medium as the expression of EdU was reduced in that group. Accordingly, little proliferation occurred in the absence of exogenous stimulation (**Figures 1C,D**). This result was verified by the introduction of anti-ICAM-1 neutralizing antibody. ICAM-1 levels decreased in both the NG and HG groups (**Supplementary Figure S1D**), indicating the efficiency of ICAM-1 inhibition. Notably, the proliferation rate markedly declined in the NG group following exposure to the ICAM-1 inhibitor. No significant differences were detected between the NG and HG groups (**Figure 1E**). Thus, EC proliferation decreased in the hyperglycemic condition through reduced expression of ICAM-1. These findings established that increased EC proliferation is a major step in injury repair (Liang et al., 2007; Bourland et al., 2019). Given our observation that HG reduced EC proliferation via ICAM-1, we hypothesized that low levels of ICAM expression would decrease injury closure in the HG group. To test this hypothesis, we introduced anti-ICAM-1 neutralizing antibody to scratch-closure EC monolayers. The effects were evaluated in terms of scratch-closure area and gap distance. The closure area was less in the HG group at both 24 and 48 h post-scratching (**Figures 1F,G**). The gap distance tended to be wider than that of the NG group (**Figure 1H**), whereas no differences in scratch-closure area or gap distance were observed following treatment with the ICAM-1 inhibitor. The collective findings indicated that injury closure was markedly delayed in the HG group due to reduced ICAM-1 expression.

HG Attenuates Neutrophil Migration via ICAM-1 and LFA-1

Neutrophil recruitment is strongly associated with bacterial clearance at injury sites. Therefore, efficient neutrophil migration is an essential step in injury repair. ICAM-1/LFA-1 interaction stimulates signaling pathways involved in neutrophil migration to the inflamed tissue (Lefort and Ley, 2012). To explore the effects of ICAM-1 and LFA-1 on neutrophil migration under hyperglycemic conditions, we modeled neutrophil migration under inflammation using the Transwell system. Both ICAM-1 and LFA-1 were decreased in the HG group (**Figures 2A–C**). Concomitantly, the Transwell migration assay revealed fewer migrating neutrophils in the HG group ($P < 0.03$) (**Figures 2D,E**). To independently explore the role of LFA-1 in neutrophil migration, we introduced an LFA-1 inhibitor to block the function of LFA-1. As expected, the level of LFA-1 sharply decreased in both the NG and HG groups (**Figure 2F**). Strikingly, the number of migrating neutrophils in the NG group was halved following exposure to the LFA-1 inhibitor (**Figure 2G**). To elucidate the crosslink between LFA-1 and ICAM-1 in neutrophil migration, we blocked the function of ICAM-1 using anti-ICAM-1 neutralizing antibody. As expected, the expression of LFA-1 and ICAM-1 were both reduced (**Figures 2H,I**) and little migration was detected in either group following the administration of the ICAM-1 inhibitor (**Figure 2J**). These findings indicated that HG attenuates neutrophil migration via ICAM-1 and LFA-1.

ICAM-1 and LFA-1 Regulate Neutrophil Phagocytosis in the HG Group

LFA-1 has been implicated in the regulation of neutrophil phagocytosis (Sigal et al., 2000). However, little is known about the role of LFA-1 in neutrophil phagocytosis under hyperglycemic conditions. Therefore, we introduced internally labeled bacteria to evaluate neutrophil phagocytosis. As exhibited in **Figure 3A**, the CFSE-labeled S. aureus staining were presented in the left side. The CFSE-labeled S. aureus/DAPI merged images were presented in the middle. Images in upper side were from hyperglycemia treated group. Images in lower side were from normoglycemia treated group. Neutrophil phagocytosis was evaluated by the clearance index. The results showed that the clearance index was 60% lower in the HG group ($P < 0.05$) (**Figures 3A,B**). Accordingly, the total number of neutrophils involved in phagocytosis of bacteria was also lower in the HG group ($P < 0.05$) (**Figure 3C**). These results indicated that HG attenuates neutrophil phagocytosis of bacterial pathogens. To further confirm the role of LFA-1 in neutrophil phagocytic activity, we introduced an LFA-1 inhibitor to block LFA-1 expression. The number of positive phagocytic neutrophils was

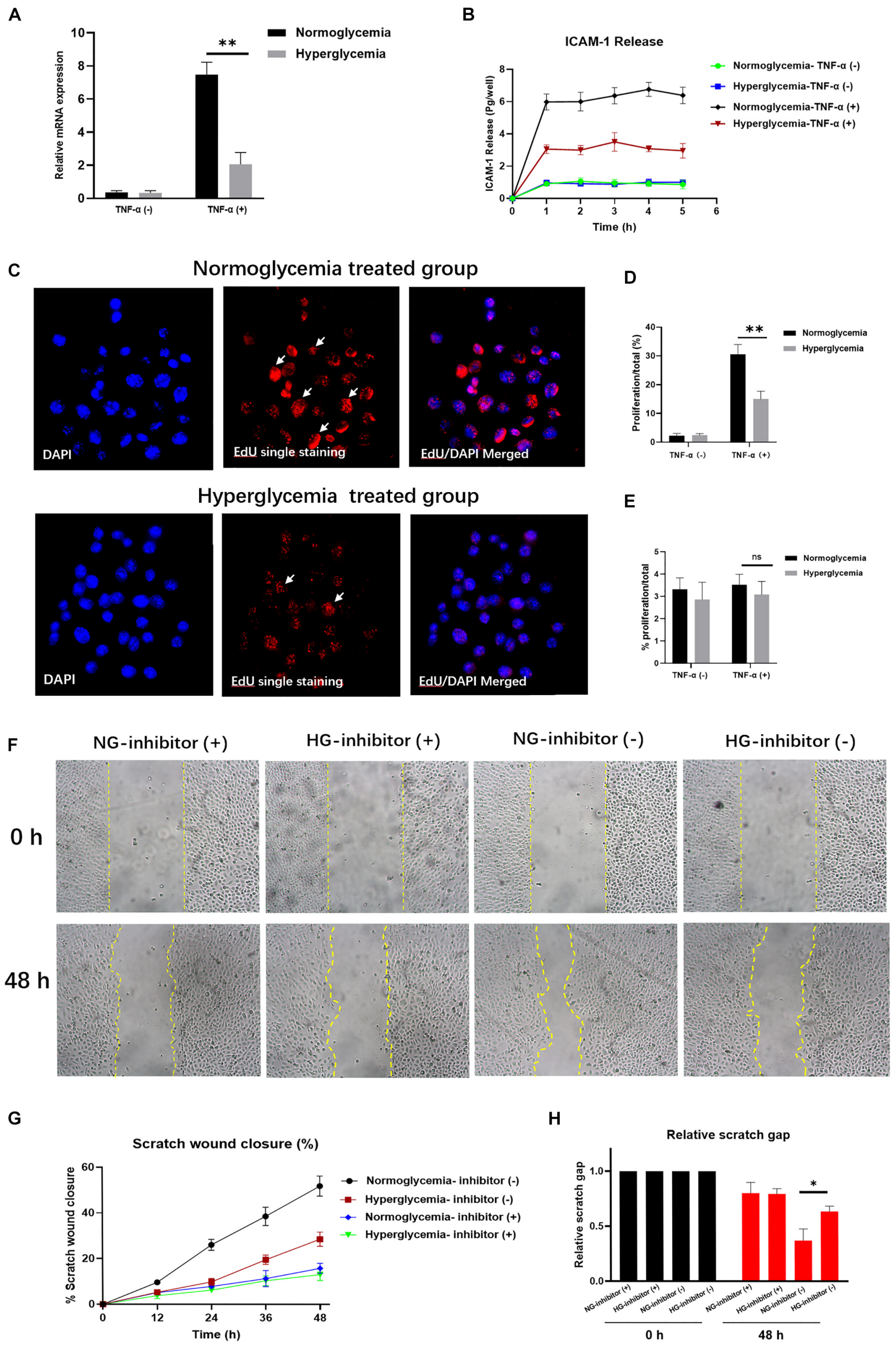


FIGURE 1 | Hyperglycemia (HG) reduces ICAM-1 expression and attenuates endothelial cell (EC) proliferation. **(A)** ICAM-1 expression was lower in the HG group ($P < 0.01$). No significant differences were detected in their non-activated counterparts (NG; $P > 0.05$). **(B)** The total amount of ICAM-1 released into the basolateral chamber was decreased in the HG group. **(C)** EC proliferation was decreased in HG cultural medium. The contrast nuclear was stained with DAPI (blue) and presented in the left side. The cell proliferation was measured by 5-Ethynyl-2'-deoxyuridine (EdU) staining (red) and presented in the middle. White arrows indicate positive staining of EdU. The merged images were presented in the right. Images in upper side were from normoglycemia group. Images in lower side were from hyperglycemia group. **(D)** The expression of EdU was reduced in HG group, which indicated the low proliferation in HG culture medium. Little proliferation occurred in the absence of exogenous stimulation. **(E)** Proliferation rates declined markedly following exposure to an ICAM-1 inhibitor in the NG group. **(F,G)** In the HG group, the closure area was decreased at both 24 and 48 h post-scratching compared with the NG group. **(H)** The scratch gap distance tended to be wider in the HG group. The yellow line demarcates the closure area after scratching. Bars represent mean ± SD. $^*P < 0.05$; $^{**}P < 0.01$.

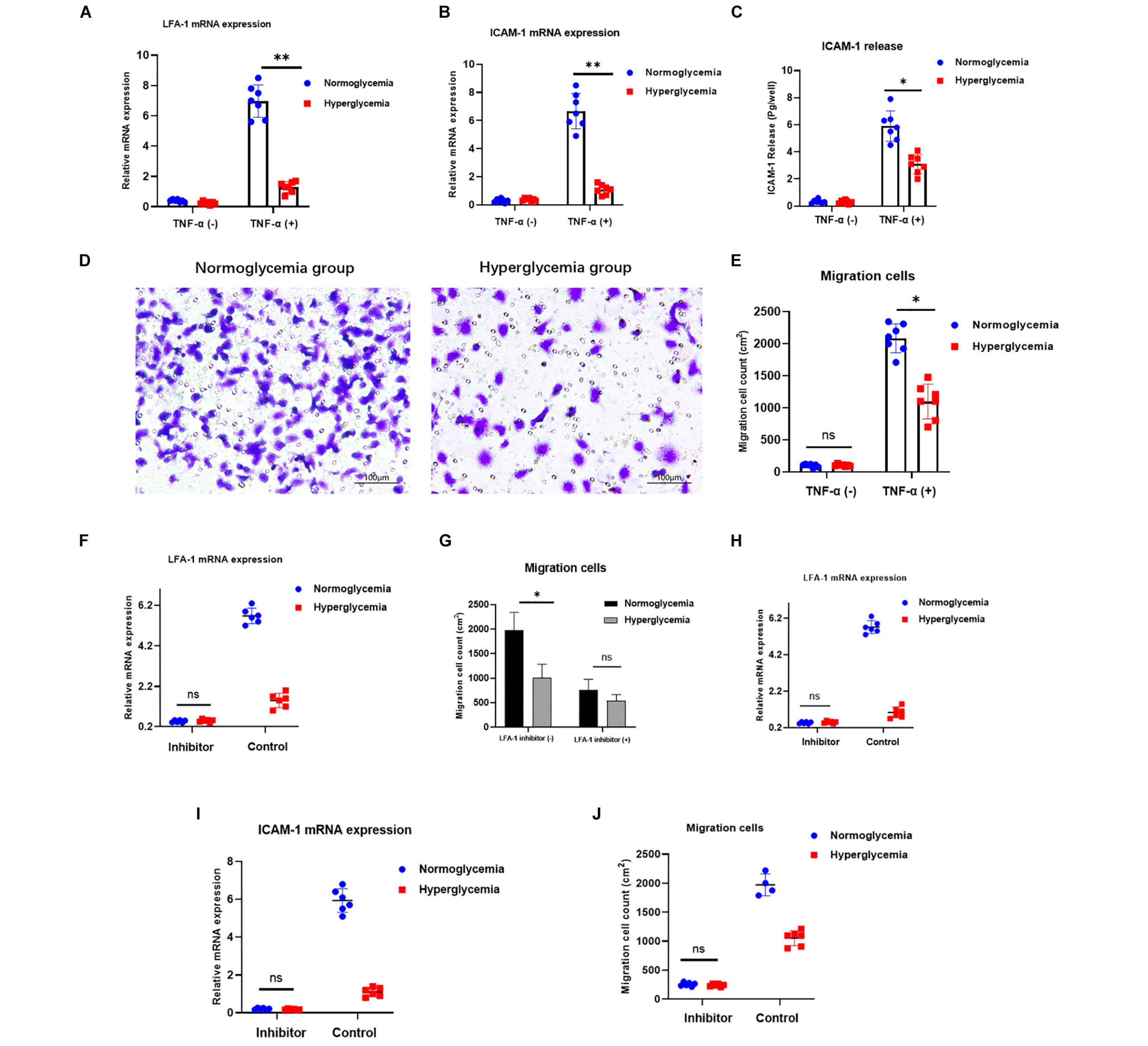


FIGURE 2 | HG attenuates neutrophil migration via ICAM-1 and LFA-1. **(A–C)** Expression of ICAM-1 and LFA-1 was decreased in the HG group. **(D,E)** A Transwell migration assay revealed lower numbers of migrating neutrophils in the HG group. Images in left side were from normoglycemia group. Images in right side were from hyperglycemia group. **(F)** LFA-1 expression decreased sharply in both the NG and HG groups following exposure to the LFA-1 inhibitor. **(G)** The number of migrating neutrophils in the NG group was halved following exposure to the LFA-1 inhibitor. **(H,I)** ICAM-1 and LFA-1 expression levels decreased in both the HG and NG groups. **(J)** Little migration was observed in either group following administration of the ICAM-1 inhibitor ($P > 0.05$). Bars represent mean ± SD. $^*P < 0.05$; $^{**}P < 0.01$.

significantly reduced in the NG group after exposure to the LFA-1 inhibitor (**Figures 3D,E**). Neutrophil activation enhances the efficiency of pathogen clearance, which is associated with the upregulation of CD11b. As shown in **Figures 3G,H**, CD11b expression was lower in the HG group. However, in the absence of LFA-1, neutrophils exhibited low levels of CD11b in both the NG and HG groups. Myeloperoxidase (MPO) is another neutrophil phagocytic biomarker (Muller, 2003; Ley et al., 2007). MPO generates hypochlorous acid, which aids neutrophil phagocytosis. Therefore, we extended this experiment by examining MPO activity in the NG and HG groups. An enzyme activity assay revealed that MPO activity was 1.2-fold lower in the HG group. MPO levels were further reduced in the presence of LFA-1 inhibitor (**Figure 3F**). To elucidate the nature of the association between LFA-1 and ICAM-1 in neutrophil phagocytosis, we blocked the function of ICAM-1 using anti-ICAM-1 neutralizing antibody. Interestingly, the expression of LFA-1was decreased in both the NG and HG groups (**Figure 3I**). In addition, neutrophil phagocytosis was attenuated following the administration of the ICAM-1 inhibitor (**Figure 3J**). These findings provided direct evidence that ICAM-1 and LFA-1 regulate neutrophil phagocytosis in hyperglycemic conditions.

HG Decreases ICAM-1 and LFA-1 Expression *in vivo*

To independently confirm the role of ICAM-1 and LFA-1 in the regulation of neutrophil phagocytosis *in vivo*, we induced injury in ICAM-$1^{+}/^{+}$ mice. HG was induced by continuous STZ injection for 5 days. The average blood glucose levels were 345.1 and 331.5 mg/dL in wild type (ICAM-$1^{+}/^{+}$) and ICAM-1 deletion (ICAM-$1^{-}/^{-}$) mice, respectively (**Figure 4A**). Skin injury repair was monitored at 2, 4, and 8 days post-surgery. Intriguingly, injury closure was significantly delayed in ICAM-$1^{+}/^{+}$HG mice (**Figure 4B**). Microscopy of tissues stained with H&E showed that ICAM-$1^{+}/^{+}$-HG mice displayed delayed injury repair with incomplete re-epithelialization and greater epithelium distance (**Figures 4C,D**). The deposition of new granular tissue was also decreased in the ICAM-$1^{+}/^{+}$HG group (**Figure 4E**), indicating insufficient injury repair in the hyperglycemic condition. Scratch-injury closure was markedly attenuated in the HG group due to reduced ICAM-1 expression. Immunofluorescence staining (**Figures 4F–H**) and ELISA analysis (**Figure 4I**) revealed that ICAM-1 expression was decreased in ICAM-$1^{+}/^{+}$HG injury tissue. White arrows indicated the positive staining of ICAM-1. Similarly, LFA-1 expression also decreased in ICAM-$1^{+}/^{+}$HG injury tissue (**Figures 4J,K**), suggesting the strong interaction between LFA-1 and ICAM-1 in injury repair *in vivo*.

HG Impairs Neutrophil Phagocytosis and Recruitment via ICAM-1/LFA-1

LFA-1 has been shown to induce neutrophil migration *in vitro*. Since LFA-1 expression levels were reduced in ICAM-$1^{+}/^{+}$HG injury tissue, we hypothesized that decreased LFA-1 expression would reduce neutrophil infiltration into ICAM-$1^{+}/^{+}$HG injury sites. To evaluate neutrophil recruitment *in vivo*, Ly6G^{+} granulocytic subsets of CD11b^{+} myeloid cells were detected by fluorescence-activated cell sorting (FACS) analysis. The gating strategy used in this analysis is shown in **Figure 5A**. As expected, the proportion of neutrophil granulocytes (CD45^{+}CD11b^{+}Ly6G^{+}) was decreased in the ICAM-$1^{+}/^{+}$-HG group (**Figures 5B,C**). As LFA-1 expression was implicated in neutrophil phagocytic activity, we also explored the effects of LFA-1 on neutrophil phagocytosis in ICAM-$1^{+}/^{+}$-HG injury tissue. The ICAM-$1^{+}/^{+}$-HG and -NG groups were treated with LPS to induce neutrophilia. Injured tissue was collected 8 and 16 h following injection of ZymTR. The number of ZymTR positive neutrophils was markedly decreased in the ICAM-$1^{+}/^{+}$HG group (**Figures 5D,E**). To confirm the critical role of LFA-1 in neutrophil phagocytosis and recruitment *in vivo*, we topically injected the LFA-1 inhibitor at both ICAM-$1^{+}/^{+}$-HG and ICAM-$1^{+}/^{+}$-NG injury sites. Blocking LFA-1 decreased neutrophil infiltration, with no difference detected between the ICAM-$1^{+}/^{+}$-HG and -NG groups ($P > 0.05$) (**Figures 6A,B**). No significant difference in ZymTR positive neutrophils was detected between the groups ($P > 0.05$) (**Figures 6C,D**). As described above, ICAM-1 induced LFA-1 expression and was implicated in neutrophil migration and phagocytosis *in vitro*. However, the exact interactions between ICAM-1 and LFA-1 *in vivo* required further elucidation. To clarify the ICAM-1/LFA-1 association *in vivo*, we used ICAM-$1^{-}/^{-}$ mice as an injury model. Unlike the ICAM-$1^{+}/^{+}$ mice, both the ICAM-$1^{-}/^{-}$-NG and -HG groups of mice displayed a decreased frequency of neutrophil infiltration (**Figure 6E**), the release of ICAM-1 was also detected by ELISA analysis, and no difference were detected between ICAM-$1^{-}/^{-}$-NG and -HG group (**Figure 6F**). Parallel results were observed in ZymTR positive neutrophils (**Figure 6G**). Thus, although injury repair was dramatically delayed in ICAM-$1^{-}/^{-}$-NG mice (**Figures 6H,I**), no difference was observed between groups. Notably, decreased LFA-1 expression was observed in both the ICAM-$1^{-}/^{-}$-NG and -HG groups (**Figure 6J**). The collective results provided direct evidence that HG affects the expression of ICAM-1 and LFA-1, which results in insufficient injury repair. Furthermore, changes in ICAM-1 and LFA-1 expression levels impair neutrophil phagocytosis and decrease neutrophil recruitment in the injured tissue.

DISCUSSION

ICAM-1 is a key member of the immunoglobulin superfamily and is centrally involved in EC proliferation and neutrophil trafficking (Gay et al., 2011; Sumagin et al., 2016; Qiu et al., 2020). ICAM-1 is expressed at low levels on the surface of ECs, but is upregulated in response to a variety of inflammatory cytokines. ICAM-1 has been recently implicated in the regulation of injury repair by promoting EC proliferation. However, due to the complex immune response in HG, the potential effects of ICAM-1 on EC proliferation remain unclear. While investigating the natural status of ICAM-1 release, we unexpectedly found that HG decreased ICAM-1 expression, resulting in the dramatic attenuation of EC proliferation. These findings were confirmed by introducing the ICAM-1 inhibitor to rule out other factors that might contribute to EC proliferation. Similar to the results of the ICAM-1 release assay, we found that the EC proliferation rate of the NG group was markedly attenuated following exposure to the ICAM-1 inhibitor, and injury closure was decreased by HG at 24 and 48 h following creation of the scratch assay. However, no effect was observed following exposure to the ICAM-1 inhibitor. Thus, we conclude that HG can reduce the expression of ICAM-1 and prolong injury closure *in vitro*. These findings extend our knowledge of ICAM-1 function in the HG injury repair process.

We also evaluated the potential mechanism of ICAM-1 expression regulation in HG. Previous studies reported the involvement of MMPs, including MMP-9, MMP-1, and MMP-2, in the release of ICAM-1. However, a further experiment revealed no significant differences in these MMPs between the NG and HG groups, suggesting that MMPs are unlikely to play a role in ICAM-1 expression in HG. Other important kinases, such as mitogen-activated protein kinase (MAPK), c-Jun N-terminal kinase and ERK1/2, are reportedly involved

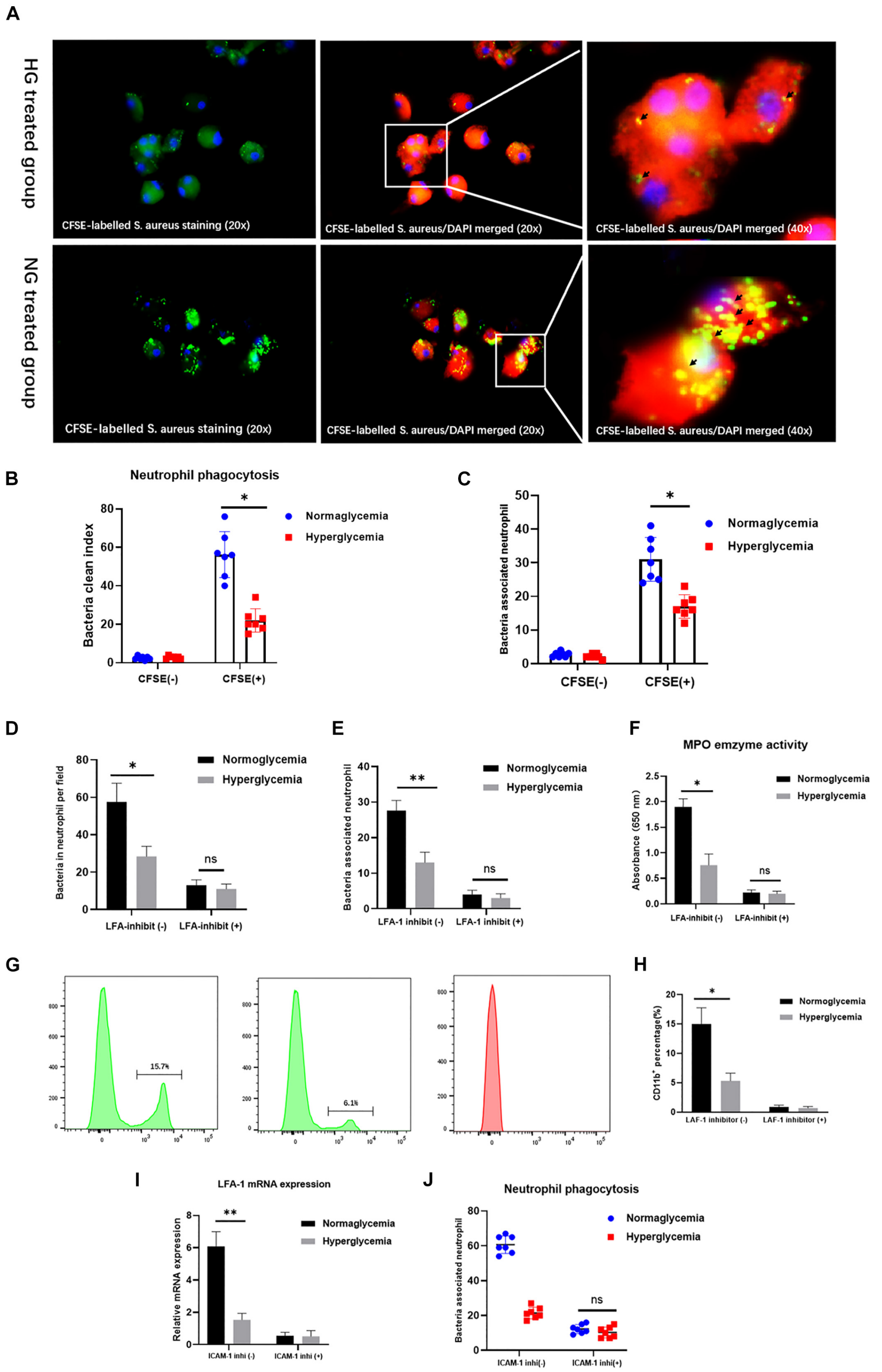


FIGURE 3 | ICAM-1 and LFA-1 regulate neutrophil phagocytosis in HG culture. **(A,B)** The neutrophil clearance index was reduced in the HG group. The CFSE-labeled S. aureus staining (20x) were presented in the left side. The CFSE-labeled S. aureus/DAPI merged images (20x) were presented in the middle. Images in upper side were from hyperglycemia treated group. Images in lower side were from normoglycemia treated group. Short black arrows indicate labeled bacteria cleared by neutrophils. **(C)** The total number of positive phagocytic neutrophils associated with labeled bacteria was decreased in the HG group. **(D,E)** The number of positive phagocytic neutrophils was significantly reduced in the NG group following exposure to the LFA-1 inhibitor. **(F)** Myeloperoxidase (MPO) enzyme activity was reduced 1.2-fold in the HG group, and further decreased by exposure to the LFA-1 inhibitor. **(G,H)** CD11b expression was elevated in the NG group. However, neutrophils in the absence of LFA-1 exhibited low CD11b levels in both the NG and HG groups. **(I)** LFA-1 expression levels were decreased in both the HG and NG groups following exposure to the ICAM-1 inhibitor. **(J)** Neutrophil phagocytosis was attenuated after administration of the ICAM-1 inhibitor. Bars represent mean ± SD. **P* < 0.05; ***P* < 0.01.

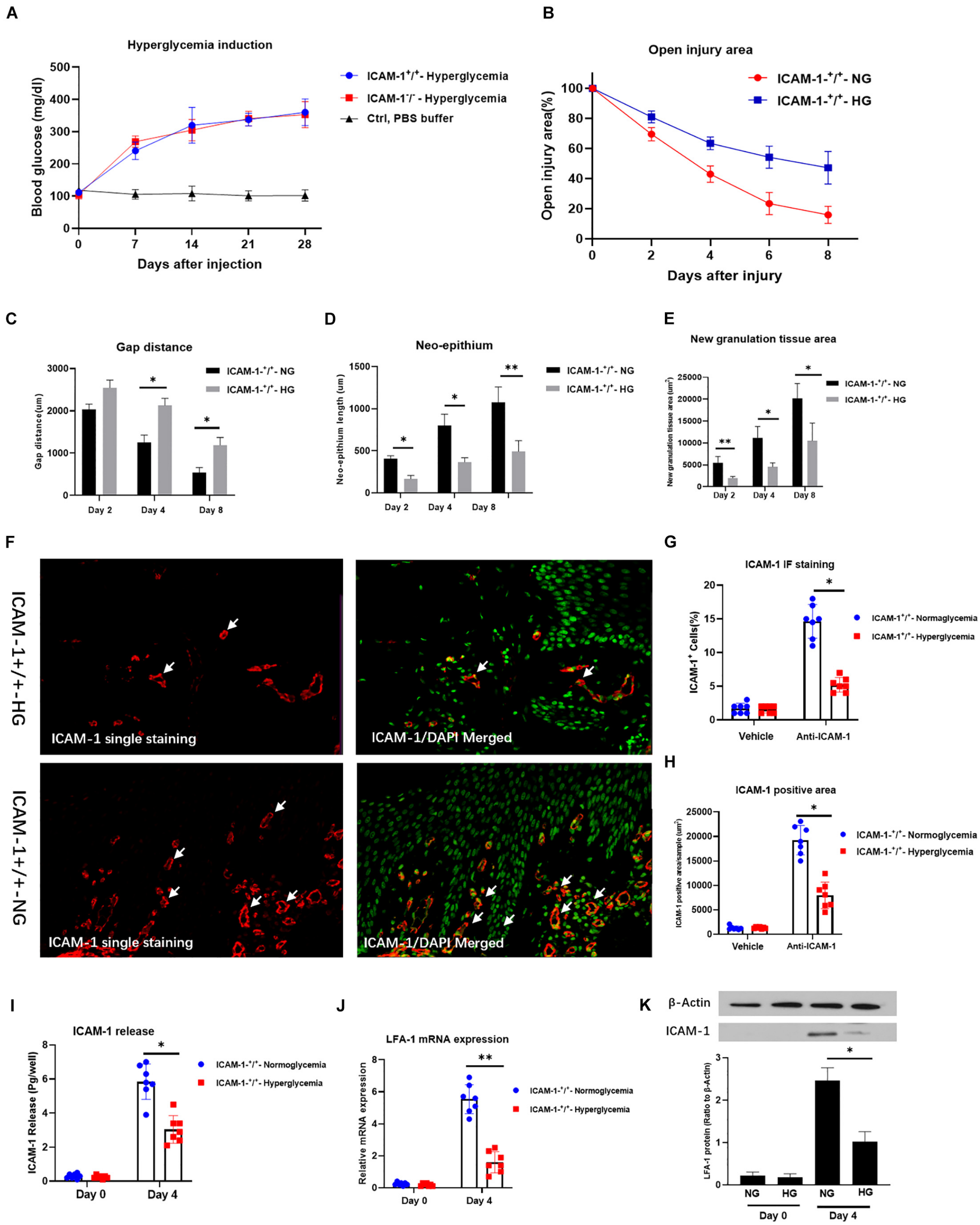


FIGURE 4 | HG decreases ICAM-1 and LFA-1 expression *in vivo*. **(A)** Blood glucose levels in ICAM-1^{-}/$^{-}$-HG and ICAM-1^{+}/$^{+}$-HG mice treated by injection of streptozotocin injection. **(B)** Injury closure was significantly delayed in ICAM-1^{+}/$^{+}$-HG mice. **(C,D)** H and E staining showed that ICAM-1^{+}/$^{+}$-HG mice displayed delayed repair, incomplete re-epithelialization and larger epithelium distance. **(E)** Deposition of new granular tissue was decreased in the ICAM-1^{+}/$^{+}$-HG group. Bars represent mean ± SD. *$P < 0.05$; **$P < 0.01$ **(F)** Immunofluorescence (IF) analysis showed that ICAM-1 expression was decreased in ICAM-1^{+}/$^{+}$-HG injury tissue. IF staining for ICAM-1 in injury tissue were red and presented in the left side. Contrast nuclear staining were green. The merged images were presented in the right side. White arrows indicate positive staining of ICAM-1. Images in upper side were from ICAM-1^{+}/$^{+}$-HG injury group. Images in lower side were from ICAM-1^{+}/$^{+}$-NG injury group. **(G,H)** ICAM-1 expression was decreased in ICAM-1^{+}/$^{+}$-HG injury tissue **(I)** Enzyme-linked immunosorbent assay results revealed reduced release of ICAM-1 from injured tissue in ICAM-1^{+}/$^{+}$-HG mice. **(J,K)** LFA-1 expression was decreased in ICAM-1^{+}/$^{+}$-HG injured tissue. Bars represent mean ± SD. *$P < 0.05$; **$P < 0.01$.

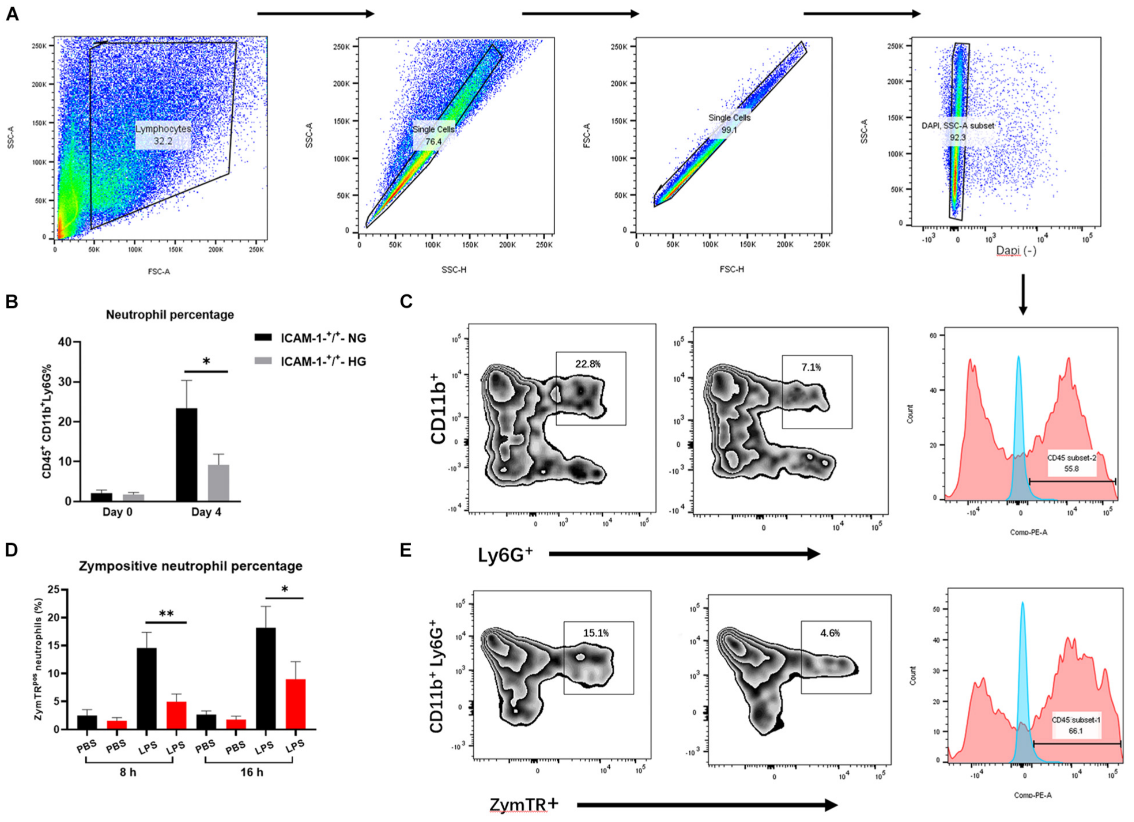


FIGURE 5 | HG impairs neutrophil phagocytosis and recruitment via ICAM-1 and LFA-1 *in vivo*. **(A)** Gating strategies for neutrophils ($CD45^+CD11b^+Ly6G^+$). **(B,C)** The proportion of neutrophils was decreased in the ICAM-$1^+/^+$-HG group. **(D,E)** The number of ZymTR positive neutrophils was reduced in the ICAM-$1^+/^+$-HG group. Bars represent mean ± SD. $^*P < 0.05$; $^{**}P < 0.01$.

in ICAM-1 expression (Christensen and Bruggemann, 2014; Hurabielle et al., 2020). Further studies focussing on these kinases are required to elucidate the potential mechanisms of ICAM-1 expression.

LFA-1 is a heterodimeric integrin consisting of αL and β2 subunits expressed on the surface of neutrophils (Lefort and Ley, 2012). Recent studies have shown that the interaction of LFA-1 with its ligand ICAM-1 mediates several important steps in the cell immune response. For example, LFA-1 integrin is critical for the firm adhesion of neutrophils to ICAM-1 (Meisel et al., 2018) and the expression of ICAM-1 and LFA-1 triggers the activation of myosin light chains, MAPK and Rho GTPase, which enhances neutrophil transmigration into inflamed tissues (Wolcott et al., 2016; Bourland et al., 2019). Neutrophil recruitment has been associated with bacterial clearance at injury sites. Specifically, ICAM-1 and LFA-1 are essential for neutrophil trafficking to inflamed tissue. However, the impact of ICAM-1 and LFA-1 on neutrophil migration in HG remains poorly understood. In this context, we explored the causative involvement of ICAM-1 and LFA-1 by modeling neutrophil migration in an inflammatory stimulation model. As expected, both LFA-1 and ICAM-1 were attenuated in the HG medium. Intriguingly, the Transwell migration assay also revealed fewer migrating neutrophils in the HG group, which was consistent with the low expression of ICAM-1 and LFA-1. The results indicate that HG attenuates neutrophil migration by regulating the expression of ICAM-1 and LFA-1. We further identified the association between LFA-1 and ICAM-1 in neutrophil migration by blocking the function of ICAM-1. Little migration was observed in the NG or HG group following exposure to an ICAM-1 inhibitor. Together, these results provide solid evidence that, under hyperglycemic conditions, ICAM-1 is involved in neutrophil migration by inducing LFA-1 expression.

We further analyzed the role of LFA-1 in neutrophil phagocytosis by introducing internally labeled bacteria. Interestingly, in HG medium, the ability of neutrophils to clear the bacteria was dramatically attenuated and the total number of neutrophils associated with labeled bacteria was reduced. These findings indicated that HG impairs neutrophil phagocytosis. To confirm these findings, we introduced an LFA-1 inhibitor to block LFA-1 expression. Surprisingly, the number of positive phagocytic neutrophils was sharply reduced in the NG group following exposure to the LFA-1 inhibitor, while no difference was detected in the HG group. These results support the hypothesis that HG decreases neutrophil phagocytosis by reducing LFA-1 expression. The findings suggest that it may be possible to promote neutrophil phagocytic activity to enhance LFA-1 expression in subjects with HG. To elucidate the interplay

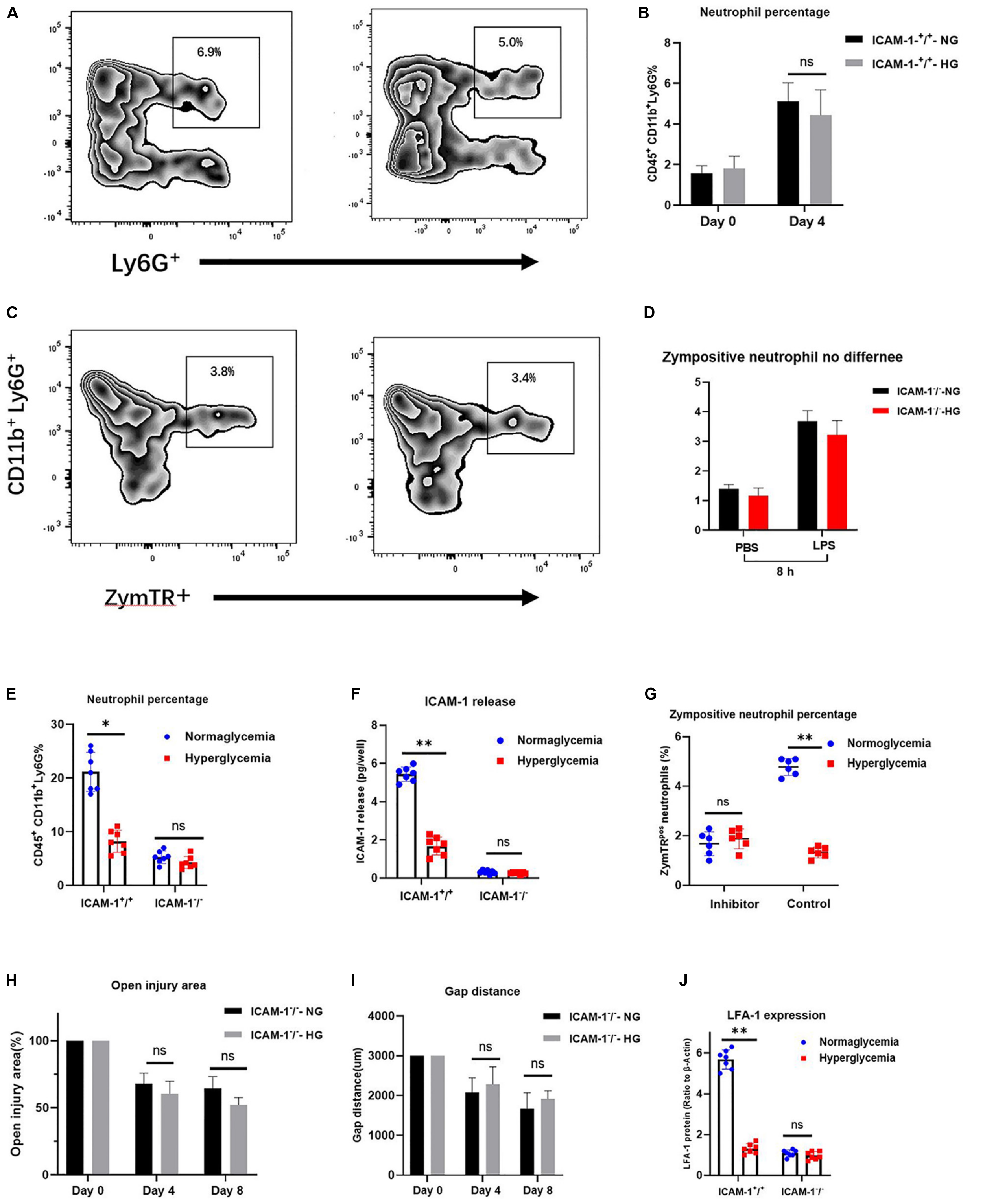


FIGURE 6 | Alteration of ICAM-1 and LFA-1 expression levels attenuates neutrophil phagocytosis and decreases neutrophil recruitment in injured tissue. **(A,B)** Blocking of LFA-1 decreased neutrophil infiltration in both the ICAM-1$^{+/+}$-HG and -NG groups ($P > 0.05$). **(C,D)** ZymTR positive neutrophil levels did not differ significantly between ICAM-1$^{+/+}$-HG and -NG groups following administration of the LFA-1 inhibitor ($P > 0.05$). **(E)** No difference was observed between groups in ICAM-1 releasing. **(F)** Neutrophil infiltration frequency was decreased in both the ICAM-1$^{-/-}$-NG and -HG groups. **(G)** ZymTR positive neutrophil levels were reduced in both the ICAM-1$^{-/-}$-NG and -HG groups. **(H,I)** Open injury area and epithelium gap distance did not differ between the ICAM-1$^{-/-}$-NG and -HG groups ($P > 0.05$). **(J)** LFA-1 expression was decreased in ICAM-1$^{-/-}$-NG and -HG injured tissue. Bars represent mean ± SD. $^{*}P < 0.05$; $^{**}P < 0.01$.

between LFA-1 and ICAM-1 in neutrophil phagocytosis, we blocked the function of ICAM-1 using anti-ICAM-1 neutralizing antibody. LFA-1 expression was dramatically decreased in both groups. Neutrophil phagocytosis was also attenuated following the administration of the ICAM-1 inhibitor. Together, these findings demonstrate the interconnection between ICAM-1 and LFA-1, neutrophil phagocytosis and HG in an *in vitro* model. Both type 1 and type 2 diabetes cause HG, which contributes to the accumulation of pathogens at injury sites, leading to insufficient injury repair (Standiford et al., 1990; Cunningham and Kirby, 1995; Liang et al., 2007; Herrera et al., 2015). Enhancing neutrophil phagocytosis by the regulation of ICAM-1/LFA-1 expression may provide novel therapeutic approaches for diabetic injury repair.

To independently confirm these observations *in vivo*, we induced skin injury using ICAM-1^{+}/$^{+}$ and ICAM-1^{-}/$^{-}$ mouse models. Consistent with our *in vitro* results, ICAM-1^{+}/$^{+}$-HG mice exhibited delayed injury repair with incomplete re-epithelialization and larger epithelium distance as well as decreased neutrophil recruitment and phagocytic activity. Importantly, the frequency of neutrophil infiltration declined dramatically in ICAM-1^{+}/$^{+}$-HG injured tissue. Similar results were also obtained in ZymTR positive neutrophils, which showed decreased levels of LFA-1. Together, these findings confirmed our *in vitro* results and indicate that HG attenuates skin injury repair and decreases neutrophil phagocytosis and recruitment by regulating ICAM-1 and LFA-1 expression.

Notably, plenty of researches showed that HG increased the expression of ICAM-1 in umbilical vein as well as in microvascular endothelial cell. In disagree with the previously findings, our study obtained the opposite results indicating that HG decreased the ICAM-1 expression in HG condition. The reasons might be as following: after the skin injury, invading pathogens and necrotic debris triggers an acute inflammatory response, which contributed to the pathogen defensing and the debris removing. Unlike the situation in microvascular endothelial or umbilical vein endothelial cells, the endothelial cells in injury tissue was exposure directly to the outside environment, and the pathogen around the injury site was easily invading into the internal injury tissue (Grice et al., 2010). Thus, the acute inflammation in injury area was critical to defense against the bacterial infections. Therefore, a source of proinflammatory cytokines, including the ICAM-1, was highly expressed in injury site to response to exogenous pathogens. ICAM-1 is constitutively presented on endothelial cells and reported to be a pro-inflammatory cytokine involved in the acute inflammatory process. ICAM-1 is also critical for the firm arrest and transmigration of neutrophil out of blood vessels into the injury tissue. Neutrophils are efficiently entering tissue and enable to engulf invading pathogens. Additionally, neutrophil release antimicrobial peptides, ROS, and cytotoxic enzymes to defense against extracellular pathogens (Wolcott et al., 2008; Su and Richmond, 2015). Therefore, the expression of ICAM-1 is closely related to the recruitment of neutrophil and worked as the protective factors within the injury repair. Taken together, as the inflammatory microenvironment was distinctive and complex in injury tissue, the expression of ICAM-1 might not be the same as the previous study and could be changed according to the specific reality. However, although we obtained the valid results based on rigorous experimental design, more studies are still required to further confirm it.

CONCLUSION

The scratch-closure assays of NG and HG cultured tissues demonstrated that HG decreases ICAM-1 expression, which results in low EC proliferation. A Transwell assay and FACS analysis further indicated that HG attenuates neutrophil recruitment and phagocytosis by reducing ICAM-1 and LFA-1 expression. These observations were confirmed *in vivo* in ICAM-1^{+}/$^{+}$ and ICAM-1^{-}/$^{-}$ mouse injury models. Together, these results highlight the important roles of ICAM-1 and LFA-1 in EC proliferation and neutrophil activity in HG culture. Targeting ICAM-1 and/or LFA-1 may provide an alternative approach for improving injury repair in diabetic populations.

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.

ETHICS STATEMENT

The animal study was reviewed and approved by the Central South University of Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

DXQ wrote the manuscript. JBH designed and supervised the study. LZ, JKZ, QY, and HLX performed the statistical analyses. DXQ, LZ, JKZ, QY, HLX, WTH, QJ, and JBH critically revised the manuscript for intellectual content. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

This work was supported by the grants from the National Natural Science Foundation of China (No. 81974223).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.616988/full#supplementary-material

Supplementary Figure 1 | (A–C) MMP-9, MMP-1 and MMP-2 levels did not differ significantly between the NG and HG groups ($P > 0.05$). **(D)** ICAM-1 expression was decreased in both the NG and HG groups ($P > 0.05$). Bars represent mean ± SD.

REFERENCES

American Diabetes, A. (2013). Diagnosis and classification of diabetes mellitus. *Diabetes Care* 36(Suppl. 1), S67–S74.

Awla, D., Abdulla, A., Zhang, S., Roller, J., Menger, M. D., Regner, S., et al. (2011). Lymphocyte function antigen-1 regulates neutrophil recruitment and tissue damage in acute pancreatitis. *Br. J. Pharmacol.* 163, 413–423. doi: 10.1111/j.1476-5381.2011.01225.x

Bandyk, D. F. (2018). The diabetic foot: Pathophysiology, evaluation, and treatment. *Semin. Vasc. Surg.* 31, 43–48. doi: 10.1053/j.semvascsurg.2019.02.001

Basit, A., Reutershan, J., Morris, M. A., Solga, M., Rose, C. E. Jr., and Ley, K. (2006). ICAM-1 and LFA-1 play critical roles in LPS-induced neutrophil recruitment into the alveolar space. *Am. J. Physiol. Lung Cell Mol. Physiol.* 291, L200–L207.

Becker, J. C., Dummer, R., Hartmann, A. A., Burg, G., and Schmidt, R. E. (1991). Shedding of ICAM-1 from human melanoma cell lines induced by IFN-gamma and tumor necrosis factor-alpha. Functional consequences on cell-mediated cytotoxicity. *J. Immunol.* 147, 4398–4401.

Bourland, J., Mayrand, D., Tremblay, N., Moulin, V. J., Fradette, J., and Auger, F. A. (2019). Isolation and Culture of Human Dermal Microvascular Endothelial Cells. *Methods Mol. Biol.* 1993, 79–90.

Campos, C. (2012). Chronic hyperglycemia and glucose toxicity: pathology and clinical sequelae. *Postgrad. Med.* 124, 90–97. doi: 10.3810/pgm.2012.11.2615

Christensen, G. J., and Bruggemann, H. (2014). Bacterial skin commensals and their role as host guardians. *Benef. Microb.* 5, 201–215. doi: 10.3920/bm2012.0062

Cunningham, A. C., and Kirby, J. A. (1995). Regulation and function of adhesion molecule expression by human alveolar epithelial cells. *Immunology* 86, 279–286.

Dragoni, S., Hudson, N., Kenny, B. A., Burgoyne, T., McKenzie, J. A., Gill, Y., et al. (2017). Endothelial MAPKs Direct ICAM-1 Signaling to Divergent Inflammatory Functions. *J. Immunol.* 198, 4074–4085. doi: 10.4049/jimmunol.1600823

Everett, E., and Mathioudakis, N. (2018). Update on management of diabetic foot ulcers. *Ann. N. Y. Acad. Sci.* 1411, 153–165. doi: 10.1111/nyas.13569

Falanga, V. (2005). Wound healing and its impairment in the diabetic foot. *Lancet* 366, 1736–1743. doi: 10.1016/s0140-6736(05)67700-8

Frykberg, R. G. (2002). Diabetic foot ulcers: pathogenesis and management. *Am. Fam. Physician.* 66, 1655–1662.

Gay, A. N., Mushin, O. P., Lazar, D. A., Naik-Mathuria, B. J., Yu, L., Gobin, A., et al. (2011). Wound healing characteristics of ICAM-1 null mice devoid of all isoforms of ICAM-1. *J. Surg. Res.* 171, e1–e7.

Gosain, A., and DiPietro, L. A. (2004). Aging and wound healing. *World J. Surg.* 28, 321–326.

Grice, E. A., Snitkin, E. S., Yockey, L. J., Bermudez, D. M., Program, N. C. S., Liechty, K. W., et al. (2010). Longitudinal shift in diabetic wound microbiota correlates with prolonged skin defense response. *Proc. Natl. Acad. Sci. U S A* 107, 14799–14804. doi: 10.1073/pnas.1004204107

Habas, K., and Shang, L. (2018). Alterations in intercellular adhesion molecule 1 (ICAM-1) and vascular cell adhesion molecule 1 (VCAM-1) in human endothelial cells. *Tissue Cell* 54, 139–143. doi: 10.1016/j.tice.2018.09.002

Han, X., Wang, Y., Chen, H., Zhang, J., Xu, C., Li, J., et al. (2016). Enhancement of ICAM-1 via the JAK2/STAT3 signaling pathway in a rat model of severe acute pancreatitis-associated lung injury. *Exp. Ther. Med.* 11, 788–796. doi: 10.3892/etm.2016.2988

Herrera, B. S., Hasturk, H., Kantarci, A., Freire, M. O., Nguyen, O., Kansal, S., et al. (2015). Impact of resolvin E1 on murine neutrophil phagocytosis in type 2 diabetes. *Infect. Immun.* 83, 792–801. doi: 10.1128/iai.02444-14

Hurabielle, C., Link, V. M., Bouladoux, N., Han, S. J., Merrill, E. D., Lightfoot, Y. L., et al. (2020). Immunity to commensal skin fungi promotes psoriasiform skin inflammation. *Proc. Natl. Acad. Sci. U S A* 117, 16465–16474. doi: 10.1073/pnas.2003022117

Hurlow, J., Couch, K., Laforet, K., Bolton, L., Metcalf, D., and Bowler, P. (2015). Clinical Biofilms: A Challenging Frontier in Wound Care. *Adv. Wound. Care* 4, 295–301. doi: 10.1089/wound.2014.0567

Kolluru, G. K., Bir, S. C., and Kevil, C. G. (2012). Endothelial dysfunction and diabetes: effects on angiogenesis, vascular remodeling, and wound healing. *Int. J. Vasc. Med.* 2012, 918267.

Lefort, C. T., and Ley, K. (2012). Neutrophil arrest by LFA-1 activation. *Front. Immunol.* 3:157. doi: 10.3389/fimmu.2012.00157

Ley, K., Laudanna, C., Cybulsky, M. I., and Nourshargh, S. (2007). Getting to the site of inflammation: the leukocyte adhesion cascade updated. *Nat. Rev. Immunol.* 7, 678–689. doi: 10.1038/nri2156

Liang, C. C., Park, A. Y., and Guan, J. L. (2007). In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. *Nat. Protoc.* 2, 329–333. doi: 10.1038/nprot.2007.30

Meisel, J. S., Sfyroera, G., Bartow-McKenney, C., Gimblet, C., Bugayev, J., Horwinski, J., et al. (2018). Commensal microbiota modulate gene expression in the skin. *Microbiome* 6:20.

Muller, W. A. (2003). Leukocyte-endothelial-cell interactions in leukocyte transmigration and the inflammatory response. *Trends Immunol.* 24, 327–334.

Nagaoka, T., Kaburagi, Y., Hamaguchi, Y., Hasegawa, M., Takehara, K., Steeber, D. A., et al. (2000). Delayed wound healing in the absence of intercellular adhesion molecule-1 or L-selectin expression. *Am. J. Pathol.* 157, 237–247. doi: 10.1016/s0002-9440(10)64534-8

Prokopowicz, Z., Marcinkiewicz, J., Katz, D. R., and Chain, B. M. (2012). Neutrophil myeloperoxidase: soldier and statesman. *Arch. Immunol. Ther. Exp.* 60, 43–54. doi: 10.1007/s00005-011-0156-8

Qiu, D., Nikita, D., Zhang, L., Deng, J., Xia, Z., Zhan, J., et al. (2020). ICAM-1 deletion delays the repair process in aging diabetic mice. *Metabolism* 114, 154412. doi: 10.1016/j.metabol.2020.154412

Qiu, D., Xia, Z., Deng, J., Jiao, X., Liu, L., and Li, J. (2019). Glucorticoid-induced obesity individuals have distinct signatures of the gut microbiome. *Biofactors* 45, 892–901. doi: 10.1002/biof.1565

Qiu, D., Xia, Z., Jiao, X., Deng, J., Zhang, L., and Li, J. (2018). Altered Gut Microbiota in Myasthenia Gravis. *Front. Microbiol.* 9:2627. doi: 10.3389/fmicb.2018.02627

Rada, B. (2019). Neutrophil Extracellular Traps. *Methods Mol. Biol.* 1982, 517–528. doi: 10.1111/pin.12715

Sigal, A., Bleijs, D. A., Grabovsky, V., van Vliet, S. J., Dwir, O., Figdor, C. G., et al. (2000). The LFA-1 integrin supports rolling adhesions on ICAM-1 under physiological shear flow in a permissive cellular environment. *J. Immunol.* 165, 442–452. doi: 10.4049/jimmunol.165.1.442

Standiford, T. J., Kunkel, S. L., Basha, M. A., Chensue, S. W., Lynch, J. P. III, Toews, G. B., et al. (1990). Interleukin-8 gene expression by a pulmonary epithelial cell line. A model for cytokine networks in the lung. *J. Clin. Invest.* 86, 1945–1953. doi: 10.1172/jci114928

Su, Y., and Richmond, A. (2015). Chemokine Regulation of Neutrophil Infiltration of Skin Wounds. *Adv. Wound Care* 4, 631–640. doi: 10.1089/wound.2014.0559

Sumagin, R., Brazil, J. C., Nava, P., Nishio, H., Alam, A., Luissint, A. C., et al. (2016). Neutrophil interactions with epithelial-expressed ICAM-1 enhances intestinal mucosal wound healing. *Mucosal. Immunol.* 9, 1151–1162. doi: 10.1038/mi.2015.135

Tamanini, A., Rolfini, R., Nicolis, E., Melotti, P., and Cabrini, G. (2003). MAP kinases and NF-kappaB collaborate to induce ICAM-1 gene expression in the early phase of adenovirus infection. *Virology* 307, 228–242. doi: 10.1016/s0042-6822(02)00078-8

van der Zijpp, Y. J., Poot, A. A., and Feijen, J. (2003). ICAM-1 and VCAM-1 expression by endothelial cells grown on fibronectin-coated TCPS and PS. *J. Biomed. Mater Res. A* 65, 51–59. doi: 10.1002/jbm.a.10327

Witkowska, A. M., and Borawska, M. H. (2004). Soluble intercellular adhesion molecule-1 (sICAM-1): an overview. *Eur. Cytokine Netw.* 15, 91–98.

Wolcott, R. D., Rhoads, D. D., and Dowd, S. E. (2008). Biofilms and chronic wound inflammation. *J. Wound Care* 17, 333–341. doi: 10.12968/jowc.2008.17.8.30796

Wolcott, R., Sanford, N., Gabrilska, R., Oates, J. L., Wilkinson, J. E., and Rumbaugh, K. P. (2016). Microbiota is a primary cause of pathogenesis of chronic wounds. *J. Wound Care* 25, S33–S43.

Woodfin, A., Beyrau, M., Voisin, M. B., Ma, B., Whiteford, J. R., Hordijk, P. L., et al. (2016). ICAM-1-expressing neutrophils exhibit enhanced effector functions in murine models of endotoxemia. *Blood* 127, 898–907. doi: 10.1182/blood-2015-08-664995

Xingyuan, M., Wenyun, Z., and Tianwen, W. (2006). Leukocyte function-associated antigen-1: structure, function and application prospects. *Protein Pept. Lett.* 13, 397–400. doi: 10.2174/092986606775974429

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 January 2021
doi: 10.3389/fgene.2020.633455

Dissecting the Invasion-Associated Long Non-coding RNAs Using Single-Cell RNA-Seq Data of Glioblastoma

Bo Pang[†], *Fei Quan*[†], *Yanyan Ping, Jing Hu, Yujia Lan** *and Lin Pang**

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Na Yang,
National Institute on Aging, National Institutes of Health (NIH),
United States
Hongyi Zhang,
The University of Texas Southwestern Medical Center, United States
Shenglan Li,
The University of Texas Health Science Center at Houston,
United States

***Correspondence:**
Lin Pang
panglin@hrbmu.edu.cn
Yujia Lan
lanyujia@hrbmu.edu.cn

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

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

Received: *25 November 2020*
Accepted: *14 December 2020*
Published: *11 January 2021*

Citation:
Pang B, Quan F, Ping Y, Hu J, Lan Y and Pang L (2021) Dissecting the Invasion-Associated Long Non-coding RNAs Using Single-Cell RNA-Seq Data of Glioblastoma. Front. Genet. 11:633455. doi: 10.3389/fgene.2020.633455

Glioblastoma (GBM) is characterized by rapid and lethal infiltration of brain tissue, which is the primary cause of treatment failure and deaths for GBM. Therefore, understanding the molecular mechanisms of tumor cell invasion is crucial for the treatment of GBM. In this study, we dissected the single-cell RNA-seq data of 3345 cells from four patients and identified dysregulated genes including long non-coding RNAs (lncRNAs), which were involved in the development and progression of GBM. Based on co-expression network analysis, we identified a module (M1) that significantly overlapped with the largest number of dysregulated genes and was confirmed to be associated with GBM invasion by integrating EMT signature, experiment-validated invasive marker and pseudotime trajectory analysis. Further, we denoted invasion-associated lncRNAs which showed significant correlations with M1 and revealed their gradually increased expression levels along the tumor cell invasion trajectory, such as VIM-AS1, WWTR1-AS1, and NEAT1. We also observed the contribution of higher expression of these lncRNAs to poorer survival of GBM patients. These results were mostly recaptured in another validation data of 7930 single cells from 28 GBM patients. Our findings identified lncRNAs that played critical roles in regulating or controlling cell invasion and migration of GBM and provided new insights into the molecular mechanisms underlying GBM invasion as well as potential targets for the treatment of GBM.

Keywords: single-cell RNA sequencing, glioblastoma, invasion, long non-coding RNA, survival

INTRODUCTION

Glioblastoma (GBM) is the most common primary malignant brain tumor, comprising 16% of all primary brain and central nervous system neoplasms (Thakkar et al., 2014), with the average age-adjusted incidence rate of 3.2 per 100,000 population (Ostrom et al., 2015). Due to fast and invasive growth of the tumor, the current therapeutic option shows many limitations in its efficacy and almost all patients present the progression of the disease with a mean progression-free survival of 7–10 months (Stupp et al., 2005) and a 5-year survival rate of less than 10% (Yang et al., 2019). Though great endeavors have been performed in the past few decades, survival has not improved significantly (Wolf et al., 2019). Therefore, determining the factors which are associated with the invasion of glioblastoma is of great significance.

Apart from protein-coding genes (PCGs), long non-coding RNAs (lncRNAs), as one kind of important regulators in biological development and disease progression (Batista and Chang, 2013), were frequently reported to control the invasion and metastasis of diverse cancer types, including glioblastoma. For example, epigenetic silencing of LINC00632 could result in the CDR1as depletion, which promoted invasion *in vitro* and metastasis *in vivo* through a miR-7-independent, IGF2BP3-mediated mechanism in melanoma (Hanniford et al., 2020). The lncRNA-ATB was upregulated in hepatocellular carcinoma and further promoted the upregulation of ZEB1 and ZEB2 by competitively binding the miR-200 family, which finally induced epithelial-mesenchymal transition (EMT) and invasion (Yuan et al., 2014). The gain-of-function or loss-of-function experiments also validated the association of lncRNAs SChLAP1 and Zbtb7a with invasive prostate cancer (Prensner et al., 2013; Wang et al., 2013). Although these studies contributed to the understanding of tumor invasion, they mostly focused on few lncRNAs. Besides, utilizing traditional experiment techniques including bulk RNA sequencing also has limitations in revealing the molecular mechanisms underlying GBM invasion.

Instead, single-cell RNA sequencing (scRNA-seq) generates gene expression profiles at single-cell resolution (Tang et al., 2009), which has emerged as a powerful tool to comprehensively determine cellular states in healthy and diseased tissues (Hovestadt et al., 2019). It has been applied to subtly characterize the heterogeneity of diverse cancers and identify rare cell populations as well as key factors associated with tumorigenesis and progression (Chung et al., 2017; Li et al., 2017), which also provides an unprecedented chance to capture the important lncRNAs that participate in GBM invasion and precisely delineate their roles during GBM progression.

In the current study, we took advantage of scRNA-seq data to identify modules that showed significant overlap with differentially expressed genes (DEGs). We integrated multiple resources including EMT signatures, invasive markers and pseudotime analysis to determine the GBM invasion-associated lncRNAs and further validated our findings in an extra scRNA-seq data set. Finally, our results of the present study could provide new insights into pathological mechanism research and new therapeutic target of GBM invasion.

MATERIALS AND METHODS

Quantification and Quality Control

The raw data for most of the analyses in this study were downloaded from the GEO database (GSE84465). This data was published by Darmanis et al. (2017) and included 3589 cells from four primary GBM patients (BT_S1, BT_S2, BT_S4, and BT_S6). The labels of malignant cells and normal cells were provided by the authors. Raw reads were mapped to the human genome (hg19) by Bowtie (version 1.1.1) (Langmead et al., 2009), and the gene expression levels were quantified as transcripts per million (TPM) using RSEM (version 1.2.28) (Li and Dewey, 2011) with the option estimate-rspd and default parameters. Log2 transformed TPM values with an offset of 1 were used to denote expression levels. We excluded low-quality cells with less than 100,000 aligned reads or with less than 2000 detected genes. We further discarded genes with the number of expressed cells less than 50. As a result, we retained 998 GBM cells and 2347 normal cells with 11520 PCGs and 1877 lncRNAs.

The processed data (GSE131928) for validation was downloaded from the GEO database, which contains 6863 GBM cells and 1067 normal cells from 28 patients. This data was published by Neftel et al. (2019). We excluded PCGs with less than 50 expressed cells or lncRNAs with less than 5 expressed cells. Finally, we retained 11441 PCGs and 585 lncRNAs.

Differential Expression Analysis and Functional Annotation

We used the MAST software package (version 1.14.0) (Finak et al., 2015) to identify genes that were differentially expressed in malignant cells compared with normal cells. Briefly, this probabilistic method takes log-transformed TPM values as input and uses the shrinkage variance estimate obtained by the empirical Bayes method. The genes with an absolute logFC > 1 and FDR < 0.05 were considered as significantly DEGs.

Then, the functional annotation and pathway enrichment analysis of genes were implemented by ClueGO (Bindea et al., 2009) with the threshold of FDR < 0.05.

WGCNA Analysis

The co-expression network analysis was performed using Weighted Gene Co-Expression Network Analysis (WGCNA, version 1.69) (Langfelder and Horvath, 2008). The TPM values of PCGs were used as input for module detection. First, the soft threshold for network construction was selected, which was 6 here. The soft threshold made the adjacency matrix to be the continuous value between 1 and 20, so that the constructed network was conformed to be the power-law distribution and was closer to the real biological network state. Second, the scale-free network was constructed using *blockwiseModules* function, followed by the module partition analysis to identify gene co-expression modules, which could group genes with similar patterns of expression. The modules were defined by cutting the clustering tree into branches using a dynamic tree-cutting algorithm and assigned to different colors for visualization. Finally, we obtained three modules containing less than 1000 member genes. The co-expression network of each module was exported using *exportNetworkToCytoscape* function and further visualized by Cytoscape (version 3.6.0) (Shannon et al., 2003).

The Effects of LncRNAs on Clinical Outcomes of GBM Patients

The expression profiles of 165 GBM samples from TCGA were downloaded from https://osf.io/gqrz9/ (Tatlow and Piccolo, 2016), with the clinical information for survival analysis obtained from the public cBio Cancer Genomics Portal[1] (Cerami et al., 2012; Gao et al., 2013). The overall survival and disease-free survival were used as the end points. The Kaplan–Meier method

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

was used for the visualization purposes and the differences between survival curves were calculated by log-rank test. The *P* values less than 0.05 were considered to be statistically significant. All of these statistical analyses were performed using R software[2], version 3.4.4.

Clustering of GBM Cells in Validation Data From Neftel et al. (2019)

Clustering cells was performed using Monocle (version 2.6.4) (Trapnell et al., 2014) with regressing out the patient effect. We used the *reduceDimension* function, which actually used the *lmFit* function in R package limma (Ritchie et al., 2015) to remove the patient effect on gene expression. We selected genes with average expression level more than 0.1 and high dispersion for clustering, which were marked using *setOrderingFilter* function. Then *clusterCells* function was used to cluster cells in an unsupervised manner, with parameters rho_threshold = 2 and delta_threshold = 4. Monocle employs a density-based approach (Rodriguez and Laio, 2014) to automatically cluster cells based on each cell's local density (rho_threshold) and the nearest distance (delta_threshold) to another cell with higher distance. Certain cells with local density and distance more than the thresholds are considered as the density peaks, which are then used to identify the clusters for all cells. We finally identified 15 cell clusters in validation data from Neftel et al. (2019)

Estimation of Activity for Diverse Signatures

The GSVA scores of EMT were calculated using predefined gene sets (**Supplementary Table 1**) extracted from the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011). For invasive scores and cell type scores, we calculated the mean expression levels of GBM invasion-associated genes which were manually extracted from previous studies (**Supplementary Table 1**) and brain cell type-specific markers defined by Darmanis et al. (2015).

RESULTS

The Characterization of the Dysregulated Transcriptome in GBM

Although previous studies have reported the close relationships of PCGs and lncRNAs with cancers using bulk RNA sequencing data (Chen Q. et al., 2018; Tao et al., 2020), few have focused on the roles of lncRNAs in tumorigenesis and progression of GBM at single-cell level. To address this issue, we initially downloaded the single-cell RNA-seq data of 3589 cells from four GBM patients [published by Darmanis et al. (2017)]. After preprocessing and quality control (see section "Materials and Methods"), we retained 998 GBM cells and 2347 normal cells with 11520 PCGs and 1877 lncRNAs. Compared with PCGs, most of lncRNAs showed relatively lower expression levels on average (**Figure 1A**). However, we also observed a small part of lncRNAs had comparably high expression levels with PCGs. And lncRNAs had more variable expression as shown by the high coefficient of variation (CV) for averaged expression than PCGs (CV = 2.98 for lncRNAs and CV = 2.09 for PCGs), suggesting their potential functional relevance. This was supported by the observations that the Spearman rank correlation coefficients calculated between any two cell pairs for lncRNAs were significantly lower than those for PCGs in both GBM cells and normal cells (Wilcoxon rank sum test, $P < 0.001$, **Figure 1B**).

To capture the functional molecules during tumorigenesis, we further utilized MAST (Finak et al., 2015), which was specifically designed for single-cell RNA-seq data to identify the DEGs between GBM and normal cells (see section "Materials and Methods"). We totally identified 2050 upregulated and 385 downregulated PCGs (**Figure 1C** and **Supplementary Table 2**), among which *TNC* (Nie et al., 2015; Xia et al., 2016), *IGFBP2* (Hsieh et al., 2010; Patil et al., 2015), and *EGFR* (Giannini et al., 2005; Beck et al., 2011) ranked in the top 10 DEGs and were all reported to be associated with gliomagenesis and GBM invasion. Functional enrichment analysis revealed that the upregulated PCGs were involved in biological processes like glial cell differentiation, glial cell proliferation and regulation of neurotransmitter transport and the downregulated PCGs mainly participated in defense response and regulation of neurons, such as myeloid leukocyte mediated immunity, regulation of leukocyte apoptotic process, cytokine production involved in immune response and negative regulation of neuron apoptotic process (**Supplementary Figure 1**). Moreover, we obtained 72 upregulated and 9 downregulated lncRNAs (**Figure 1C** and **Supplementary Table 2**). Besides some well-known cancer-associated lncRNAs such as *LINC01158* (Li Y. et al., 2018), *LINC00461* (Dong et al., 2019), *XIST* (Yu et al., 2017), and *HOTAIRM1* (Li Q. et al., 2018), we also identified several potential GBM progression-associated lncRNAs like *POLR2J4*, *WWTR1-AS1*, and *VIM-AS1*.

Identification of GBM-Associated Modules at Single-Cell Level

Since genes usually synergistically play important roles in tumorigenesis, we performed WGCNA (Langfelder and Horvath, 2008) on the PCG expression profiles of GBM cells to identify highly co-expressed clusters of genes (see section "Materials and Methods"). We finally obtained three modules (M1, M2, and M3), which contained 53, 37, and 30 genes, respectively. The genes in each module were highly connected to form a tight network structure (**Figure 2A**), showing strong correlations of expression levels with each other (**Supplementary Figure 2**). To determine the contribution of each module to gliomagenesis and progression, we performed the functional enrichment analysis of module genes. M1 genes were mainly involved in cell-cell adhesion, wound healing and spreading of cells, cell migration and positive regulation of lipid localization (**Figure 2B**). M2 genes were only enriched into one biological process of smooth muscle cell migration and there were no functions enriched by M3 genes. The pathway enrichment analysis on the genes in the three modules revealed that M1 genes were involved in human complement system, zinc homeostasis and senescence and autophagy in cancer (**Supplementary Figure 3**). M2 genes were only enriched into p52 signaling pathway while none pathways were enriched by

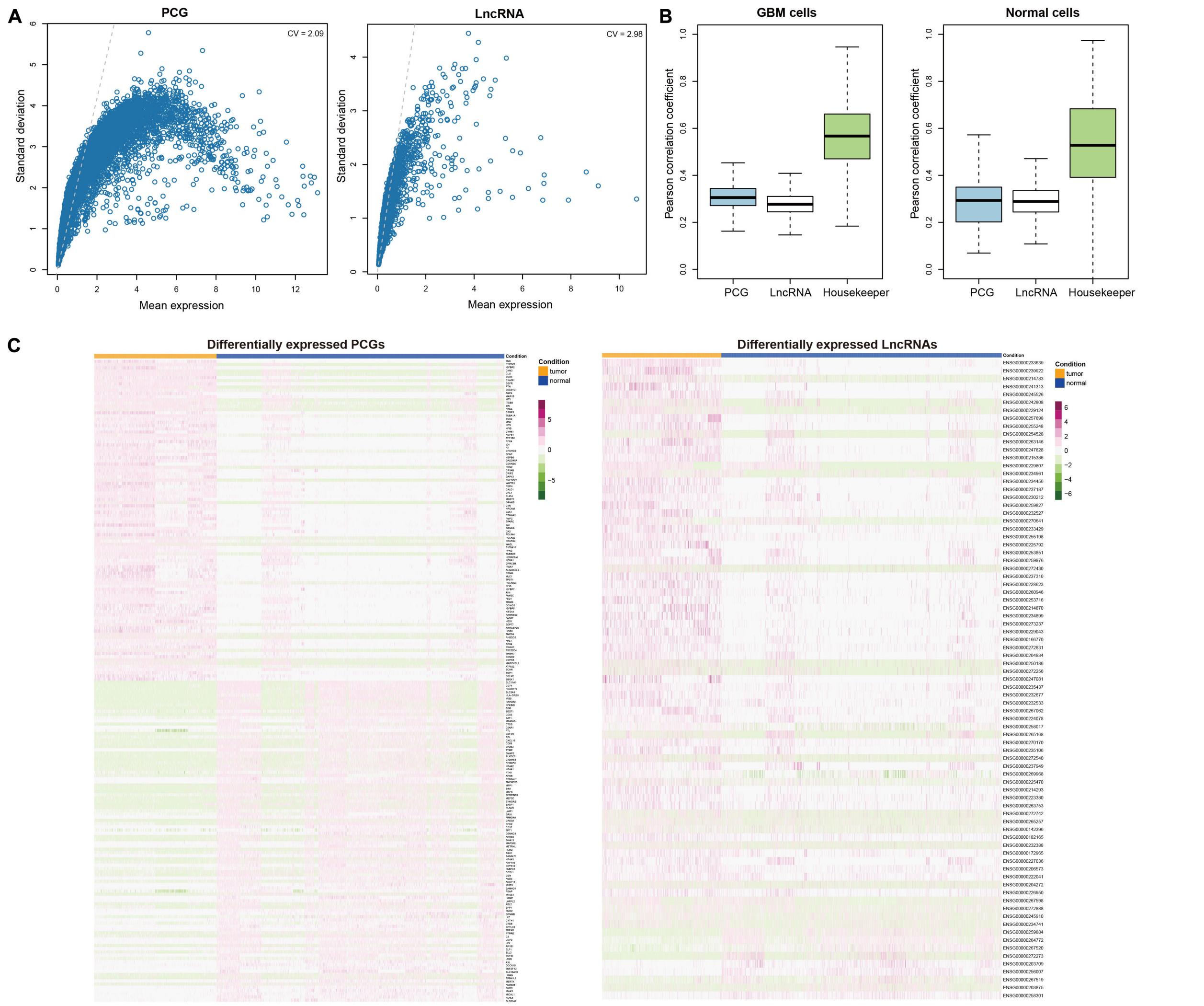


FIGURE 1 | Characterization of dysregulated transcriptome in GBM at single cell level. **(A)** Scatter plots evaluating the average expression levels of PCGs (left) and lncRNAs (right) with their variations across cells, respectively. **(B)** Comparison of correlation coefficients between cells which were calculated based on the expression levels of PCGs, lncRNAs and housekeepers in GBM cells (left) and normal brain cells (right). **(C)** Heatmaps showing the top 100 upregulated PCGs and top 100 downregulated PCGs (left) and all differentially expressed lncRNAs (right). Each row represents one PCG or lncRNA and each column represents a cell. Orange denotes the GBM cells and blue denotes the normal cells.

M3 genes. Moreover, we found that M1 showed a significant overlap with DEGs (hypergeometric test, $P = 8.76 \times 10^{-21}$), accounting for 75.5 percentage (40/53) of module genes (**Figure 2C**). M2 contained 11 DEGs, which accounted for 29.7 percentage (11/37) of modules, while there was no significant overlap between M3 genes and DEGs since M3 contained only one DEG. These results implied the critical roles of these modules in the tumorigenesis and progression of GBM, especially for M1.

Determination of GBM Invasion-Associated Module

Since M1 was the most significant and largest module that enriched for DEGs, we further assessed its contribution to GBM progression. Most M1 genes showed relatively high positive correlations of expression levels with each other, except for *CD99*, *MTRNR2L1*, and *MTRNR2L2* (**Figure 3A**). Notably, many DEGs in M1 have been reported to be associated with migration and invasion. For example, EPAS1 was an important transcription factor (TF) that was validated to promote the invasive potential of GBM cells by our previous work (Pang et al., 2019). Many studies revealed that ANX family proteins (ANXA1 and ANXA2), especially ANXA2, could promote cancer progression including proliferation, invasion and metastasis (Chen C.Y. et al., 2018). The S100 proteins such as S100A11 could promote GBM progression through ANXA2-mediated NF-κB signaling pathway (Tu et al., 2019) and S100A10 could form heterotetramers with ANXA2 to promote the activation of matrix metalloproteases (MMPs) to increase the invasive ability

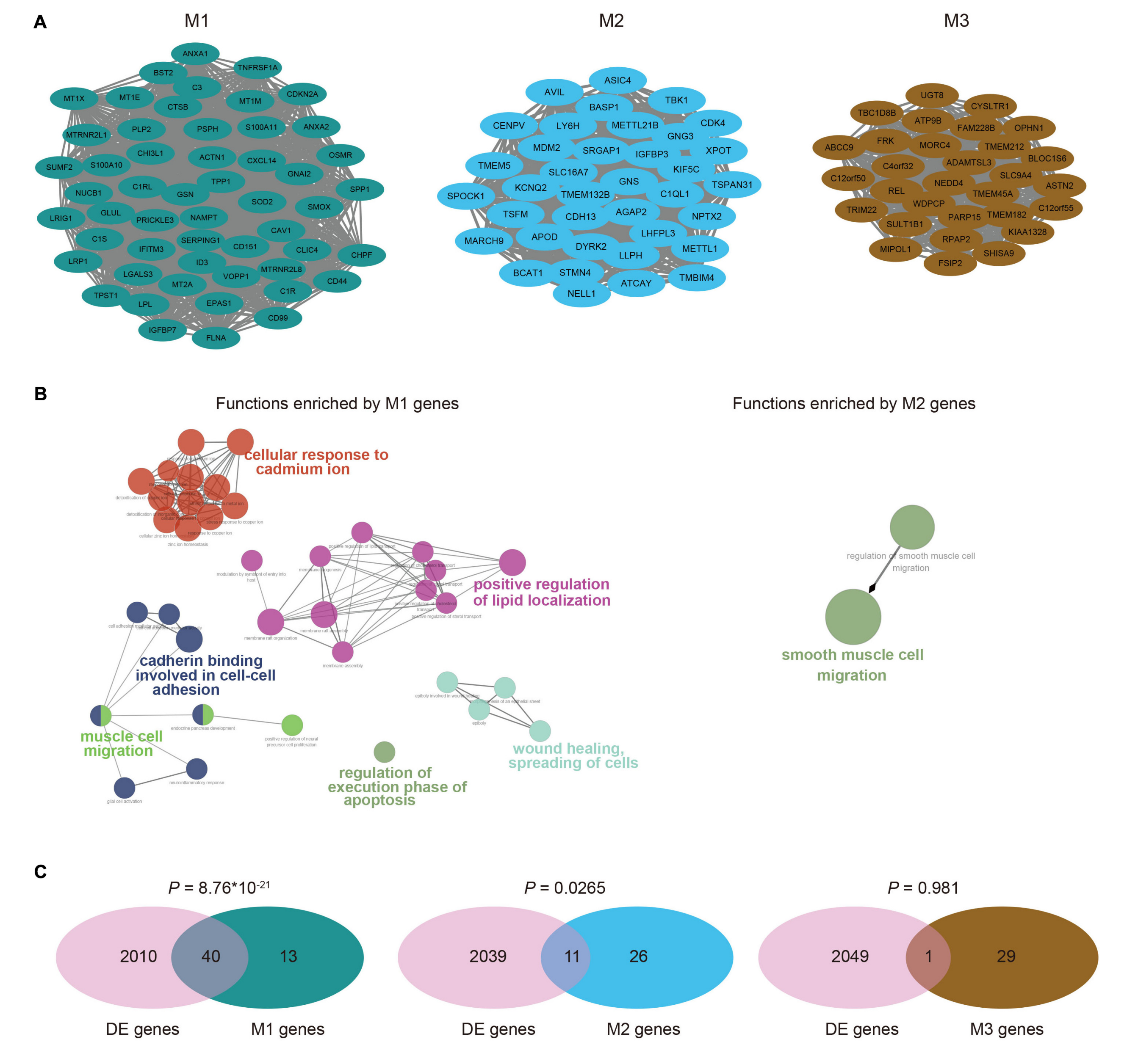

A
M1
ANXA1
BST2
TNFRSF1A
C3
MT1X
MT1E
CTSB
MT1M
CDKN2A
MTRNR2L1
PLP2
PSPH
S100A11
ANXA2
CHI3L1
ACTN1
CXCL14
OSMR
SUMF2
S100A10
GNAI2
TPP1
NUCB1
C1RL
GSN
SPP1
SOD2
SMOX
GLUL
LRIG1
PRICKLE3
NAMPT
CAV1
C1S
IFITM3
SERPING1
CD151
CLIC4
CHPF
LRP1
ID3
MTRNR2L8
LGALS3
VOPP1
MT2A
CD44
C1R
TPST1
LPL
EPAS1
CD99
IGFBP7
FLNA
M2
ASIC4
AVIL
TBK1
BASP1
CENPV
LY6H
METTL21B
CDK4
GNG3
MDM2
SRGAP1
IGFBP3
XPOT
TMEM5
SLC16A7
KIF5C
GNS
TSPAN31
SPOCK1
KCNQ2
TMEM132B
C1QL1
TSFM
CDH13
AGAP2
NPTX2
LHFPL3
MARCH9
APOD
DYRK2
METTL1
LLPH
BCAT1
STMN4
ATCAY
TMBIM4
NELL1
M3
UGT8
CYSLTR1
TBC1D8B
ATP9B
FAM228B
OPHN1
FRK
ABCC9
MORC4
TMEM212
BLOC1S6
C4orf32
ADAMTSL3
C12orf50
SLC9A4
NEDD4
ASTN2
REL
TMEM45A
WDPCP
C12orf55
TRIM22
PARP15
TMEM182
SULT1B1
KIAA1328
RPAP2
MIPOL1
SHISA9
FSIP2
B
Functions enriched by M1 genes
cellular response to cadmium ion
response to cadmium ion
response to zinc ion
cellular response to zinc ion
detoxification of copper ion
stress response to metal ion
detoxification of inorganic compound
cellular response to copper ion
stress response to copper ion
response to copper ion
cellular zinc ion homeostasis
zinc ion homeostasis
positive regulation of lipid transport
modulation by symbiont of entry into host
membrane biogenesis
regulation of cholesterol transport
regulation of sterol transport
positive regulation of lipid localization
positive regulation of cholesterol transport
positive regulation of sterol transport
membrane raft assembly
membrane raft organization
membrane assembly
cell adhesion mediator activity
cell-cell adhesion mediator activity
cadherin binding involved in cell-cell adhesion
epiboly involved in wound healing
morphogenesis of an epithelial sheet
epiboly
endocrine pancreas development
muscle cell migration
positive regulation of neural precursor cell proliferation
wound healing, spreading of cells
regulation of execution phase of apoptosis
neuroinflammatory response
glial cell activation
Functions enriched by M2 genes
regulation of smooth muscle cell migration
smooth muscle cell migration
C
P = 8.76*10^-21
2010
40
13
DE genes
M1 genes
P = 0.0265
2039
11
26
DE genes
M2 genes
P = 0.981
2049
1
29
DE genes
M3 genes

FIGURE 2 | Co-expressed modules identified by WGCNA. **(A)** The co-expression network of module M1, M2, and M3, visualized by Cytoscape. **(B)** Functional annotations for genes in M1 and M2, which were implemented by ClueGO. There were no functions enriched by M3 genes. **(C)** Venn diagrams showed the significant overlaps of genes in each module with differentially expressed genes, except for M3. *P* values were calculated by hypergeometric test.

of tumor cells (Chen C.Y. et al., 2018). Interestingly, *ANXA1*, *ANXA2*, *S100A10*, and *S100A11* were all contained in M1 and represented high correlation, especially for *ANXA2* and *S100A10*. These observations suggested the potential association of M1 with GBM invasion.

To validate the above observations, we combined the results from our previous work (Pang et al., 2019), in which we identified 12 cell clusters using the same data set. And cluster 3, 4, 7, and 9 showed relatively higher expression of EMT-associated genes. Here, we calculated the mean expression levels of M1 genes as the M1 scores in each cell of clusters and found that cluster 3 displayed the highest M1 scores, followed by cluster 7 and 9 (**Figure 3B**), which was consistent with our previous observations. However, we similarly calculated the M2 and M3 scores and found that cluster 5 and 10 showed higher M2 scores and cluster 4 and 11 showed higher M3 scores (**Supplementary Figure 4**). Further, we collected experimentally validated genes that could contribute to the invasive ability of glioblastoma cells (such as *ZEB1*, *HNRNPC*, *WNT5A*, and *DRAM1*) to evaluate the invasive scores for each cell (see section "Materials and Methods"). Similar results were observed that those three cell clusters were the top-ranked ones with high invasive scores

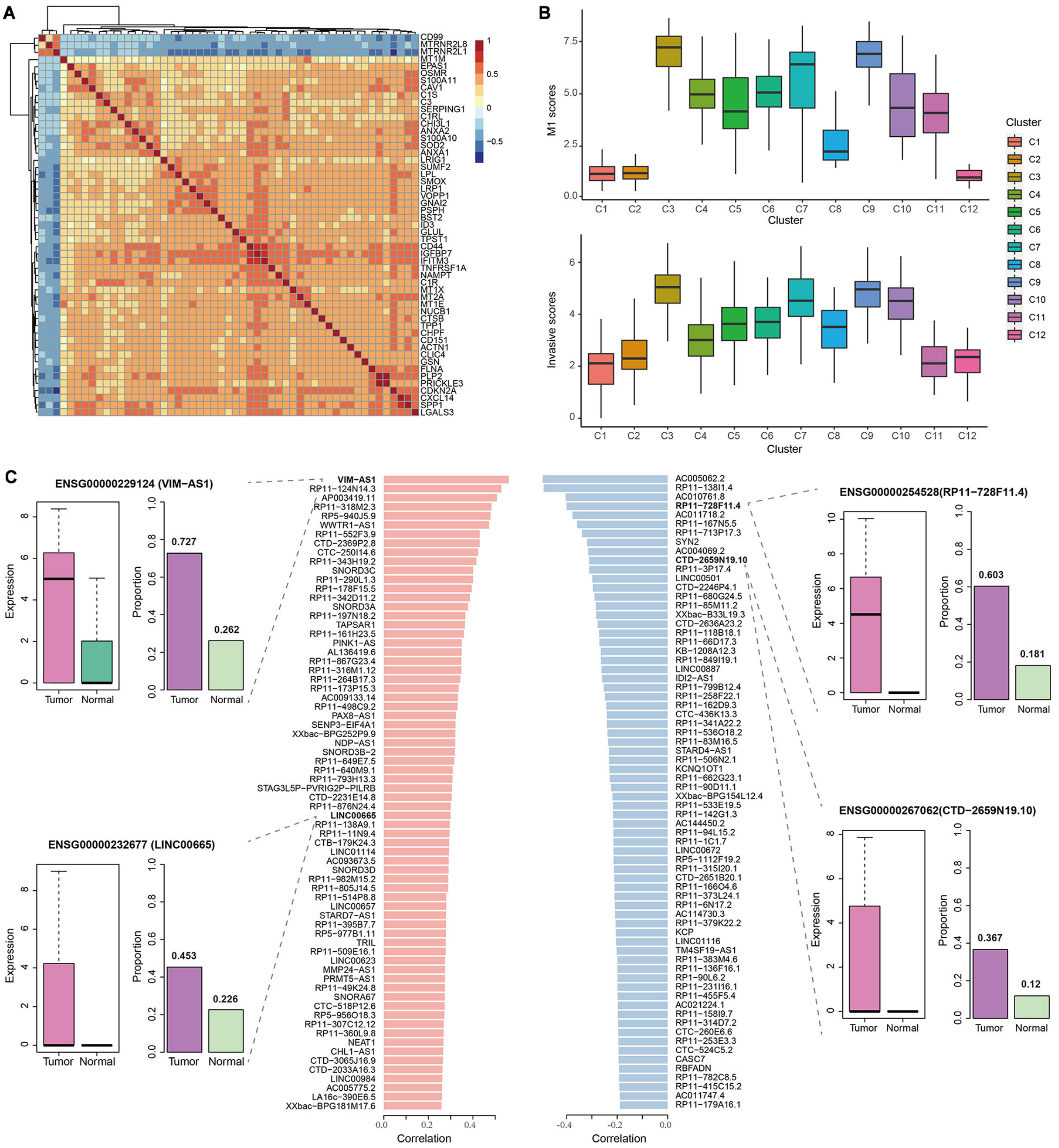


FIGURE 3 | The correlation of M1 with GBM invasion. **(A)** Heatmap showing the Spearman correlation coefficients of expression levels for any gene pair in M1. **(B)** Boxplots showing the M1 scores (top) and invasive scores (bottom) of each cell cluster identified by our previous work using the same data. The GBM invasion-associated markers were manually extracted from previous studies. **(C)** Barplots in the middle showing the significant Spearman correlation coefficients of top 100 positively (left) and negatively (right) lncRNAs between their expression levels and M1 scores. In the examples of lncRNAs, boxplots represent the expression levels of the corresponding lncRNA in tumor cells and normal cells, while barplots represent the proportion of cell with their detected expression.

(**Figure 3B**), which further supported the contribution of M1 to GBM invasion.

Identification of GBM Invasion-Associated LncRNAs

Given the close association of M1 with GBM invasion, we next calculated the Spearman rank correlation coefficients between the expression levels of each lncRNA and M1 scores across all GBM cells and identified 1264 significantly correlated lncRNAs (including 611 positively correlated lncRNAs and 653 negatively correlated lncRNAs, **Supplementary Table 3**), which were considered as GBM invasion-associated lncRNAs. The top 100 positively and negatively correlated lncRNAs were shown in **Figure 3C**. For example, among the positively correlated lncRNAs, *VIM-AS1* ranked among the top one with the correlation coefficient of 0.56, which was upregulated in GBM cells with a higher expressed proportion (72.7%) compared

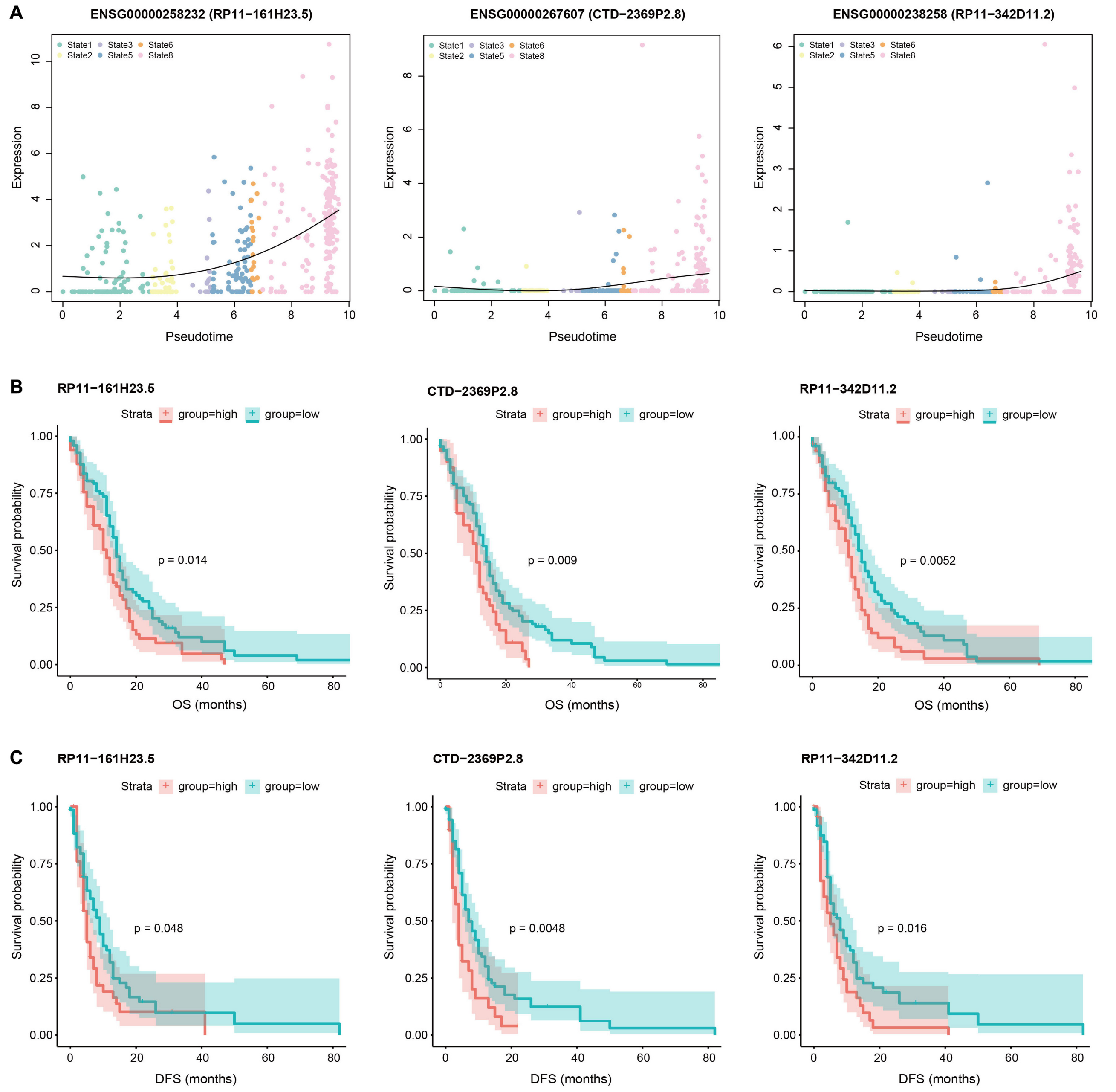


FIGURE 4 | Pseudotime and survival analysis of invasion-associated lncRNAs. **(A)** Scatter plots showing the expression levels of three example lncRNAs (RP11-161H23.5, CTD-2369P2.8, and RP11-342D11.2) increase as a function of pseudotime in "stem-to-invasion" path that identified in our previous work, containing state 1, 2, 3, 5, 6, and 8 cells. A natural spline was used to model gene expression as a smooth, non-linear function over pseudotime. **(B)** Comparison of overall survival among patients with high expression levels of these three lncRNAs (red line) and those with low expression levels of corresponding lncRNAs (green line) by Kaplan–Meier analysis (with log-rank *P* values) in the cohort of 165 GBM patients. The patients were divided into two groups based on the average expression level of corresponding lncRNAs across all patients. **(C)** Comparison of disease-free survival among patients with high expression levels of these three lncRNAs (red line) and those with low expression levels of corresponding lncRNAs (green line) by Kaplan–Meier analysis (with log-rank *P* values) in the cohort of 165 GBM patients. The patients were divided into two groups based on the average expression level of corresponding lncRNAs across all patients.

to normal cells (26.2%). Previous studies also revealed that the high expression of *VIM-AS1* was positively associated with patients' worse prognosis (Suo et al., 2020). Other lncRNAs like *WWTR1-AS1* and *LINC00665* similarly showed significantly higher expression levels and cell proportions in tumor cells. For negatively correlated lncRNAs, *ENSG00000254528* (*RP11-728F11.4*) and *ENSG00000267062* (*CTD-2659N19.10*) ranked among the top four and ten, both of which showed significantly higher expression levels in GBM cells and nearly no expression in normal cells. Notably, *VIM-AS1* and *WWTR1-AS1* were the top two lncRNAs with the highest correlations between their expression levels and pseudotime along the "stem-to-invasion path" in our previous work (Pang et al., 2019). These findings promoted us to explore the dynamic changes of GBM invasion-associated lncRNAs along the "stem-to-invasion path." We found that the expression levels of many lncRNAs such as *ENSG00000258232* (*RP11-161H23.5*), *ENSG00000267607* (*CTD-2369P2.8*), and *ENSG00000238258* (*RP11-342D11.2*), gradually increased as cells transferred from cancer stem cell-like state to invasive state (**Figure 4A**). These consistent results confirmed the potential roles of these lncRNAs on GBM invasion.

Given that cancer-associated mortality is principally attributable to the development of invasion and metastasis, we speculated that these GBM invasion-associated lncRNAs might be of importance in determining patient outcomes. Next, we performed survival analysis using the expression profiles and clinical information of 165 GBM patients (see section "Materials and Methods"). Among invasion-associated lncRNAs, several of them showed significant correlations with prognosis of patients. For example, the overall survival (OS) of patients with high expression levels of *ENSG00000258232* (*RP11-161H23.5*), *ENSG00000267607* (*CTD-2369P2.8*), and *ENSG00000238258* (*RP11-342D11.2*) were significantly shorter than those with low expression levels (P = 0.014, P = 0.009, and P = 0.0052, respectively, **Figure 4B**). Moreover, patients with high expression levels of these three lncRNAs also had worse disease-free survival (DFS) than those with low expression levels (P = 0.048, P = 0.0048, and P = 0.016, respectively, **Figure 4C**). These results suggested potential implication of invasion-associated genes in GBM tumorigenesis, progression and prognosis.

Validation of the Invasion-Associated Module and LncRNAs by Extra Data of GBM

To validate the contribution of M1 genes and lncRNAs to GBM invasion, we downloaded another single-cell RNA-seq data of 28 GBM patients [published by Neftel et al. (2019)]. After quality control, we retained 6863 GBM cells and 1067 normal cells with 11441 PCGs and 585 lncRNAs, in which the numbers of commonly detected PCGs and lncRNAs in both data sets were 11441 and 192, respectively. In this validation data, we identified 1676 DEGs and 13 dysregulated lncRNAs (**Supplementary Table 4**), among which 1066 DEGs and 6 dysregulated lncRNAs were shared by both data sets.

We recaptured the modularity of M1 genes in this validation data as they showed stronger co-expression pattern compared to the other two module genes (**Figure 5A**), suggesting their functional synergy. The similar patterns were observed in data from children and adults with GBM, respectively (**Supplementary Figure 5**). To determine whether M1 genes were involved in GBM invasion, we first used Monocle (Trapnell et al., 2014) to group GBM cells into 15 clusters, excluding patient-specific effects with linear regression (see section "Materials and Methods," **Figure 5B** and **Supplementary Figure 6**). Each cluster consisted of cells from multiple patients (**Supplementary Figure 7**). Then we calculated the EMT, invasive and M1 scores as above for each cell and found that they showed quite similar distribution patterns (**Figures 5B,C**). Cluster 5 and 6 consistently had the highest scores, followed by cluster 4, 14, and 15, which located adjacent to each other in the transcriptome space of **Figure 5B**. These results again confirmed the association of M1 genes with GBM invasion.

Therefore, we calculated the Spearman rank correlation coefficients between the expression levels of each invasion-associated lncRNA identified in data from Darmanis et al. This resulted in 71 significantly correlated lncRNAs (including 49 positively correlated lncRNAs and 22 negatively correlated lncRNAs, **Supplementary Table 5**) among the 192 commonly detected lncRNAs. Notably, NEAT1 was the top one lncRNA with a positive correlation coefficient of 0.54 in validation data (**Figure 5D**), which also ranked among the top 63 in the data from Darmanis et al. Moreover, the high expression level of NEAT1 was significantly correlated with poor OS and DFS of patients (**Figure 5E**), which was accordant with the roles of NEAT1 in promoting malignant phenotypes and progression of GBM (Chen Q. et al., 2018; Zhou et al., 2019). All these results again validated the contributions of the identified lncRNAs to GBM invasion and progression.

DISCUSSION

The fast and invasive growth is the hallmark of GBM, which is a major factor contributing to dismal outcomes (Du et al., 2008). Therefore, understanding the molecular mechanisms underlying tumor cell invasion and migration is crucial for the treatment of GBM. Although previous studies have made massive efforts to identify many PCGs and lncRNAs promoting glioblastoma cell invasion using bulk sequencing data, few have actually achieved successful clinical application. In this study, we utilized single-cell RNA-seq data from multiple GBM patients to dissect invasion-associated factors including lncRNAs, which provided new insights into the development and progression of glioblastoma.

Central to our understanding of glioblastoma biology is the idea that a subpopulation of glioblastoma stem cells drives tumorigenesis and progression (Singh et al., 2004). Lan et al. (2017) analyzed the growth dynamics of GBM clones and revealed that the initiation of human GBM may result from the aberrant reactivation of a normal developmental program. Couturier et al. (2020) compared the lineage hierarchy of the developing human brain to the transcriptome of 53586 adult glioblastoma cells at single-cell level and found that glioblastoma development recapitulates a normal neurodevelopmental hierarchy. These findings suggested the important roles of the development system in tumorigenesis and progression of GBM and were also supported by many other studies (Filbin et al., 2018; Yuan et al., 2018). Consistently, in this work, we identified dysregulated PCGs and lncRNAs and the functional enrichment analyses showed that these PCGs participated in brain development-associated biological processed, such as glial cell differentiation and glial cell

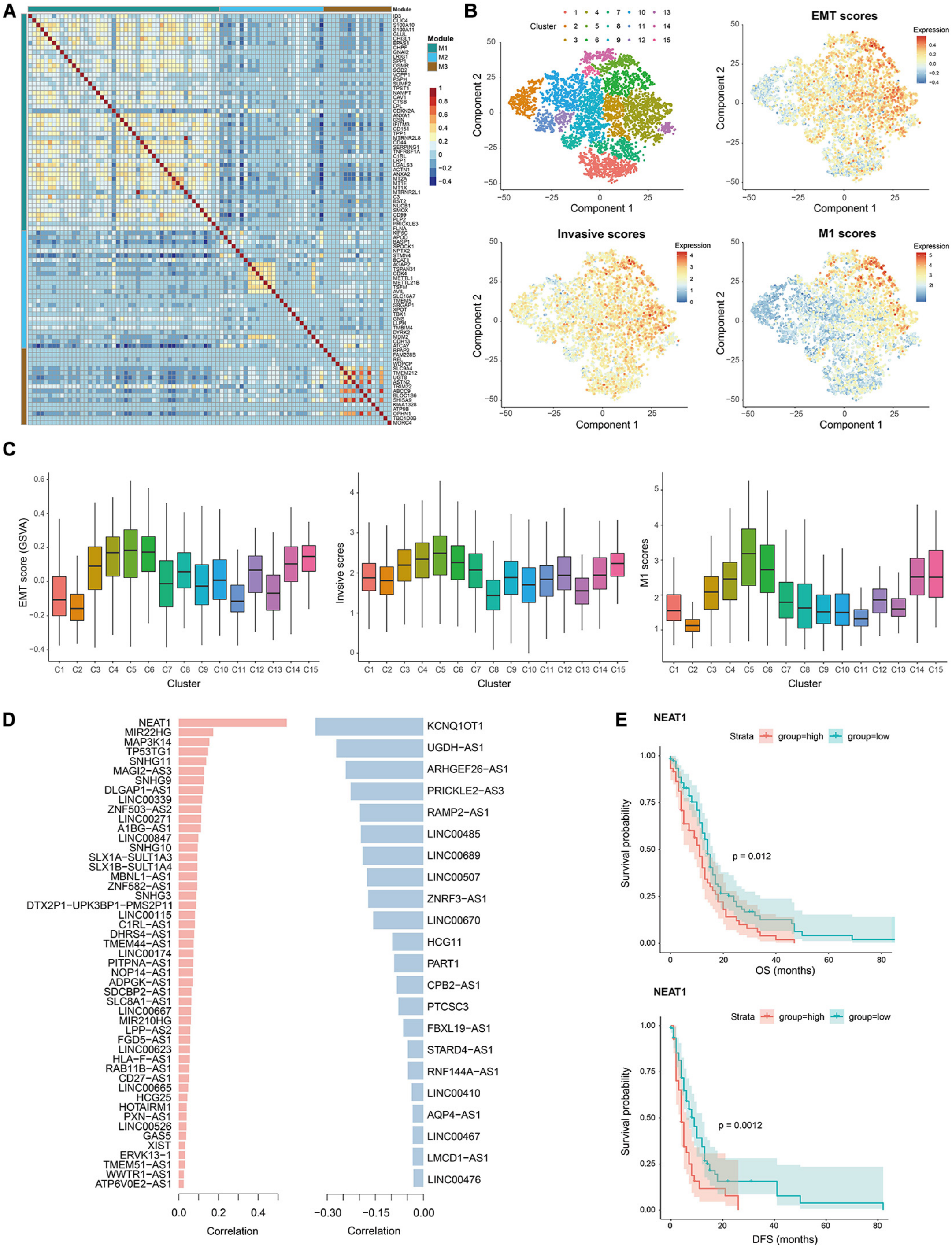


FIGURE 5 | Validation of invasion-associated M1 and lncRNAs using data from Neftel et al. (2019) **(A)** Heatmap showing the Spearman correlation coefficients of expression levels for any gene pair in M1, M2, and M3. **(B)** T-SNE plots of tumor cells showing 15 clusters and the EMT scores, invasive scores and M1 scores in each cell. Red denotes high scores and blue denote low scores. **(C)** Comparison of EMT scores, invasive scores and M1 scores in cells of each cluster, indicating the similar distribution as cluster 5, 6, 14, and 15 display relatively higher scores. **(D)** List of commonly identified positively (left) and negatively (right) lncRNAs as well as their Spearman correlation coefficient with M1 scores in this validation data. **(E)** Comparison of overall (top) and disease-free (bottom) survival among patients with high expression levels (red line) of lncRNA NEAT1 and those with low expression levels (green line) by Kaplan–Meier analysis (with log-rank *P* values) in the cohort of 165 GBM patients. The patients were divided into two groups based on the average expression level of NEAT1 across all patients.

proliferation. This implied that we indeed captured the potential key factors contributing to GBM initiation and progression.

Since invasion and metastasis are the late events during the course of multi-step tumor progression (Lambert et al., 2017), which result in the vast majority of deaths from cancer (Coghlin and Murray, 2010), we seek to identify critical factors, especially lncRNAs, that are involved in the regulation of GBM invasion. Given the lack of functional annotation of lncRNAs, we first identified co-expressed PCG modules by WGCNA to determine the invasion-associated genes. Among the three modules, M1 significantly enriched the largest number of differentially expressed PCGs, many of which have been reported the association with GBM invasion, such as *EPAS1*, *ANXA2* and its target gene *OSMR* (Matsumoto et al., 2020). And ANAX2 was also the target of lncRNA LINC00941, which was one of the invasion-assocaited lncRNAs. Previous studies have revealed that S100A10 could form a heterotetramer with ANXA2 to promote tumor cell invasion (Chen C.Y. et al., 2018) and S100A11 could also interact with ANXA1 which is a Ca 2^+-regulated phospholipid-binding protein (Boudhraa et al., 2016) to form Ca 2^+-dependent heterotetramers. These genes were all contained in M1 with high expression in GBM cells, underlying the functions of cellular response to cadmium ion (**Figure 2B**) enriched by M1 genes, which might be a potential molecular mechanism of GBM invasion. Surprisingly, although most of M1 genes showed positive correlations, *CD99*, *MTRNR2L1*, and *MTRNR2L8* were negatively correlated with others. As it has been widely reported that overexpression of *CD99* could increase the migration and invasiveness of GBM cells (Seol et al., 2012; Cardoso et al., 2019), we deduced that although *CD99* and other invasion-associated PCGs play key roles in regulating tumor cell invasion, their mediated mechanisms were distinct and redundant, resulting in their mutually exclusive expression patterns. Moreover, combining our previous work for characterization of cell clusters and construction of progression trajectory, we further confirmed the contribution of M1 to GBM invasion as M1 genes showed relatively high expression in cell clusters with high EMT and invasive scores. Interestingly, we calculated the average expression levels of cell type-specific markers defined by previous study (Darmanis et al., 2015) as the cell type scores in each cluster and found that cluster 3, 4, 7, and 9 with higher M1 scores consistently showed the highest expression levels of microglia cell markers (**Supplementary Figure 8**), implying the roles of microglia in GBM invasion. These observations were also recaptured in another single-cell RNA-seq data of GBM, suggesting the accuracy and repeatability of our findings.

Based on the determination of the invasion-associated module, we further identified the invasion-associated lncRNAs. In data from Darmanis et al., we found that *VIM-AS1* and *WWTR1-AS1* ranked among the top 1 and 6 in positively correlated lncRNAs with higher expression in GBM cell compared to normal cells. Notably, their expression gradually increased along the "stem-to-invasion path" in our previous work (Pang et al., 2019), confirming their roles in GBM invasion. In validation data from Neftel et al. (2019) *NEAT1* was the top one positively correlated lncRNA and MIAT was the top one negatively correlated lncRNA, consistent with their roles in GBM progression that increased *NEAT1* could promote proliferation, malignant phenotypes and TMZ resistance (Bi et al., 2020) and high expression of *MIAT* is associated with prolonged survival (Bountali et al., 2019). However, we did not recapture the top-ranked lncRNAs like *VIM-AS1* and *WWTR1-AS1* as they were not detected in validation data. This may result from the generally lower expression levels of lncRNAs compared to PCGs and the inherent limitations of scRNA-seq like high dropout rates and data sparsity. Actually, among the 192 commonly detected lncRNAs, 71 were consistently identified as invasion-associated lncRNAs in both data sets, indicating the robustness of our results.

In summary, our work took advantage of scRNA-seq to identify and dissect the GBM invasion-associated lncRNAs and their effect on clinical outcomes at a high resolution, providing new insights into the molecular mechanism of the development and progression of GBM and new potential targets for the treatment of invasive glioblastoma and possibly other solid malignant tumors.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GEO database (GSE84465 and GSE131928).

AUTHOR CONTRIBUTIONS

LP and YL designed the research. FQ, YP, and JH collected and preprocessed the data. BP, FQ, and LP performed the bioinformatics analysis. BP and LP 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 Nos. 31801116, 81802926, and 31900478), the China Postdoctoral Science Foundation (2018M641842 and 2017M621295), the Heilongjiang Postdoctoral Foundation (LBH-Z17160, LBH-Z17141, and LBH-Z19082), Special funds for the construction of higher education in Heilongjiang Province (Grant No. UNPYSCT-2018068), and the Health Department Science Foundation of Heilongjiang Province (Grant No. 2017-167).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.633455/full#supplementary-material

REFERENCES

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

Beck, S., Jin, X., Sohn, Y. W., Kim, J. K., Kim, S. H., Yin, J., et al. (2011). Telomerase activity-independent function of TERT allows glioma cells to attain cancer stem cell characteristics by inducing EGFR expression. *Mol. cells* 31, 9–15. doi: 10.1007/s10059-011-0008-8

Bi, C. L., Liu, J. F., Zhang, M. Y., Lan, S., Yang, Z. Y., and Fang, J. S. (2020). LncRNA NEAT1 promotes malignant phenotypes and TMZ resistance in glioblastoma stem cells by regulating let-7g-5p/MAP3K1 axis. *Biosci. Rep.* 40, BSR20201111.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Boudhraa, Z., Bouchon, B., Viallard, C., D'Incan, M., and Degoul, F. (2016). Annexin A1 localization and its relevance to cancer. *Clin. Sci.* 130, 205–220. doi: 10.1042/cs20150415

Bountali, A., Tonge, D. P., and Mourtada-Maarabouni, M. (2019). RNA sequencing reveals a key role for the long non-coding RNA MIAT in regulating neuroblastoma and glioblastoma cell fate. *Int. J. Biol. Macromol.* 130, 878–891. doi: 10.1016/j.ijbiomac.2019.03.005

Cardoso, L. C., Soares, R. D. S., Laurentino, T. S., Lerario, A. M., Marie, S. K. N., and Oba-Shinjo, S. M. (2019). CD99 expression in glioblastoma molecular subtypes and role in migration and invasion. *Int. J. Mol. Sci.* 20:1137. doi: 10.3390/ijms20051137

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

Chen, C. Y., Lin, Y. S., Chen, C. H., and Chen, Y. J. (2018). Annexin A2-mediated cancer progression and therapeutic resistance in nasopharyngeal carcinoma. *J. Biomed. Sci.* 25:30.

Chen, Q., Cai, J., Wang, Q., Wang, Y., Liu, M., Yang, J., et al. (2018). Long noncoding RNA NEAT1, regulated by the EGFR pathway, contributes to glioblastoma progression through the WNT/beta-catenin pathway by scaffolding EZH2. *Clin. Cancer Res.* 24, 684–695. doi: 10.1158/1078-0432.ccr-17-0605

Chung, W., Eum, H. H., Lee, H. O., Lee, K. M., Lee, H. B., Kim, K. T., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* 8:15081.

Coghlin, C., and Murray, G. I. (2010). Current and emerging concepts in tumour metastasis. *J. Pathol.* 222, 1–15. doi: 10.1002/path.2727

Couturier, C. P., Ayyadhury, S., Le, P. U., Nadaf, J., Monlong, J., Riva, G., et al. (2020). Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat. Commun.* 11, 3406.

Darmanis, S., Sloan, S. A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., et al. (2017). Single-Cell RNA-Seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* 21, 1399–1410. doi: 10.1016/j.celrep.2017.10.030

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., et al. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7285–7290. doi: 10.1073/pnas.1507125112

Dong, L., Qian, J., Chen, F., Fan, Y., and Long, J. (2019). LINC00461 promotes cell migration and invasion in breast cancer through miR-30a-5p/integrin beta3 axis. *J. Cell. Biochem.* 120, 4851–4862. doi: 10.1002/jcb.27435

Du, R., Petritsch, C., Lu, K., Liu, P., Haller, A., Ganss, R., et al. (2008). Matrix metalloproteinase-2 regulates vascular patterning and growth affecting tumor cell survival and invasion in GBM. *Neuro Oncol.* 10, 254–264. doi: 10.1215/15228517-2008-001

Filbin, M. G., Tirosh, I., Hovestadt, V., Shaw, M. L., Escalante, L. E., Mathewson, N. D., et al. (2018). Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* 360, 331–335.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278.

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:l1.

Giannini, C., Sarkaria, J. N., Saito, A., Uhm, J. H., Galanis, E., Carlson, B. L., et al. (2005). Patient tumor EGFR and PDGFRA gene amplifications retained in an invasive intracranial xenograft model of glioblastoma multiforme. *Neuro Oncol.* 7, 164–176. doi: 10.1215/s1152851704000821

Hanniford, D., Ulloa-Morales, A., Karz, A., Berzoti-Coelho, M. G., Moubarak, R. S., Sanchez-Sendra, B., et al. (2020). Epigenetic silencing of CDR1as drives IGF2BP3-mediated melanoma invasion and metastasis. *Cancer cell* 37, 55–70 e15.

Hovestadt, V., Smith, K. S., Bihannic, L., Filbin, M. G., Shaw, M. L., Baumgartner, A., et al. (2019). Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature.* 572, 74–79.

Hsieh, D., Hsieh, A., Stea, B., and Ellsworth, R. (2010). IGFBP2 promotes glioma tumor stem cell expansion and survival. *Biochem. Biophys. Res. Commun.* 397, 367–372. doi: 10.1016/j.bbrc.2010.05.145

Lambert, A. W., Pattabiraman, D. R., and Weinberg, R. A. (2017). Emerging biological principles of metastasis. *Cell* 168, 670–691. doi: 10.1016/j.cell.2016.11.037

Lan, X., Jorg, D. J., Cavalli, F. M. G., Richards, L. M., Nguyen, L. V., Vanner, R. J., et al. (2017). Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* 549, 227–232.

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

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

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., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* 49, 708–718. doi: 10.1038/ng.3818

Li, Q., Dong, C., Cui, J., Wang, Y., and Hong, X. (2018). Over-expressed lncRNA HOTAIRM1 promotes tumor growth and invasion through up-regulating HOXA1 and sequestering G9a/EZH2/Dnmts away from the HOXA1 gene in glioblastoma multiforme. *J. Exp. Clin. Cancer Res.* 37:265.

Li, Y., Li, Y., Wang, D., and Meng, Q. (2018). Linc-POU3F3 is overexpressed in hepatocellular carcinoma and regulates cell proliferation, migration and invasion. *Biomed. Pharmacother.* 105, 683–689. doi: 10.1016/j.biopha.2018.06.006

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Matsumoto, Y., Ichikawa, T., Kurozumi, K., Otani, Y., Fujimura, A., Fujii, K., et al. (2020). Annexin A2-STAT3-Oncostatin M receptor axis drives phenotypic and mesenchymal changes in glioblastoma. *Acta Neuropathol. Commun.* 8:42.

Neftel, C., Laffy, J., Filbin, M. G., Hara, T., Shore, M. E., Rahme, G. J., et al. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* 178, 835–49 e21.

Nie, S., Gurrea, M., Zhu, J., Thakolwiboon, S., Heth, J. A., Muraszko, K. M., et al. (2015). Tenascin-C: a novel candidate marker for cancer stem cells in glioblastoma identified by tissue microarrays. *J. Proteome Res.* 14, 814–822. doi: 10.1021/pr5008653

Ostrom, Q. T., Gittleman, H., Fulop, J., Liu, M., Blanda, R., Kromer, C., et al. (2015). CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2008-2012. *Neuro Oncol.* 17Suppl. 4(Suppl. 4), iv1–iv62.

Pang, B., Xu, J., Hu, J., Guo, F., Wan, L., Cheng, M., et al. (2019). Single-cell RNA-seq reveals the invasive trajectory and molecular cascades underlying glioblastoma progression. *Mol. Oncol.* 13, 2588–2603. doi: 10.1002/1878-0261.12569

Patil, S. S., Railkar, R., Swain, M., Atreya, H. S., Dighe, R. R., and Kondaiah, P. (2015). Novel anti IGFBP2 single chain variable fragment inhibits glioma cell migration and invasion. *J. Neuro Oncol.* 123, 225–235. doi: 10.1007/s11060-015-1800-7

Prensner, J. R., Iyer, M. K., Sahu, A., Asangani, I. A., Cao, Q., Patel, L., et al. (2013). The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* 45, 1392–1398. doi: 10.1038/ng.2771

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

Rodriguez, A., and Laio, A. (2014). Machine learning. Clustering by fast search and find of density peaks. *Science* 344, 1492–1496. doi: 10.1126/science.1242072

Seol, H. J., Chang, J. H., Yamamoto, J., Romagnuolo, R., Suh, Y., Weeks, A., et al. (2012). Overexpression of cd99 increases the migration and invasiveness of human malignant glioma cells. *Genes Cancer* 3, 535–549. doi: 10.1177/1947601912473603

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

Singh, S. K., Hawkins, C., Clarke, I. D., Squire, J. A., Bayani, J., Hide, T., et al. (2004). Identification of human brain tumour initiating cells. *Nature* 432, 396–401.

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* 352, 987–996.

Suo, S. T., Gong, P., Peng, X. J., Niu, D., and Guo, Y. T. (2020). Knockdown of long non-coding RNA VIM-AS1 inhibits glioma cell proliferation and migration, and increases the cell apoptosis via modulation of WEE1 targeted by miR-105-5p. *Eur. Rev. Med. Pharmacol. Sci.* 24, 6834–6847.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Tao, W., Zhang, A., Zhai, K., Huang, Z., Huang, H., Zhou, W., et al. (2020). SATB2 drives glioblastoma growth by recruiting CBP to promote FOXM1 expression in glioma stem cells. *EMBO Mol. Med.* 12:e12291.

Tatlow, P. J., and Piccolo, S. R. (2016). A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. *Sci. Rep.* 6:39259.

Thakkar, J. P., Dolecek, T. A., Horbinski, C., Ostrom, Q. T., Lightner, D. D., Barnholtz-Sloan, J. S., et al. (2014). Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiol. Biomarkers Prev.* 23, 1985–1996. doi: 10.1158/1055-9965.epi-14-0275

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859

Tu, Y., Xie, P., Du, X., Fan, L., Bao, Z., Sun, G., et al. (2019). S100A11 functions as novel oncogene in glioblastoma via S100A11/ANXA2/NF-kappaB positive feedback loop. *J. Cell. Mol. Med.* 23, 6907–6918. doi: 10.1111/jcmm.14574

Wang, G., Lunardi, A., Zhang, J., Chen, Z., Ala, U., Webster, K. A., et al. (2013). Zbtb7a suppresses prostate cancer through repression of a Sox9-dependent pathway for cellular senescence bypass and tumor invasion. *Nat. Genet.* 45, 739–746. doi: 10.1038/ng.2654

Wolf, K. J., Chen, J., Coombes, J., Aghi, M. K., and Kumar, S. (2019). Dissecting and rebuilding the glioblastoma microenvironment with engineered materials. *Nat. Rev. Mater.* 4, 651–668. doi: 10.1038/s41578-019-0135-y

Xia, S., Lal, B., Tung, B., Wang, S., Goodwin, C. R., and Laterra, J. (2016). Tumor microenvironment tenascin-C promotes glioblastoma invasion and negatively regulates tumor proliferation. *Neuro Oncol.* 18, 507–517. doi: 10.1093/neuonc/nov171

Yang, D., Sun, B., Dai, H., Li, W., Shi, L., Zhang, P., et al. (2019). T cells expressing NKG2D chimeric antigen receptors efficiently eliminate glioblastoma and cancer stem cells. *J. Immunother. Cancer* 7:171.

Yu, H., Xue, Y., Wang, P., Liu, X., Ma, J., Zheng, J., et al. (2017). Knockdown of long non-coding RNA XIST increases blood-tumor barrier permeability and inhibits glioma angiogenesis by targeting miR-137. *Oncogenesis* 6:e303. doi: 10.1038/oncsis.2017.7

Yuan, J., Levitin, H. M., Frattini, V., Bush, E. C., Boyett, D. M., Samanamud, J., et al. (2018). Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med.* 10:57.

Yuan, J. H., Yang, F., Wang, F., Ma, J. Z., Guo, Y. J., Tao, Q. F., et al. (2014). A long noncoding RNA activated by TGF-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer cell* 25, 666–681. doi: 10.1016/j.ccr.2014.03.010

Zhou, X., Li, X., Yu, L., Wang, R., Hua, D., Shi, C., et al. (2019). The RNA-binding protein SRSF1 is a key cell cycle regulator via stabilizing NEAT1 in glioma. *Int. J. Biochem. Cell Biol.* 113, 75–86. doi: 10.1016/j.biocel.2019.06.003

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

ORIGINAL RESEARCH
published: 26 January 2021
doi: 10.3389/fgene.2021.636419

Five Circular RNAs in Metabolism Pathways Related to Prostate Cancer

Lili Zhang[1†], Wei Zhang[2†], Hexin Li[1], Xiaokun Tang[1], Siyuan Xu[1], Meng Wu[3], Li Wan[4], Fei Su[1] and Yaqun Zhang[3]**

[1] Clinical Biobank, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China, [2] Department of Pathology, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China, [3] Department of Urology, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China, [4] The Key Laboratory of Geriatrics, Beijing Institute of Geriatrics, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

OPEN ACCESS

Edited by:
*Chunjie Jiang,
University of Pennsylvania,
United States*

Reviewed by:
*Yingjie Zhang,
Wistar Institute, United States
Shunmin He,
Institute of Biophysics (CAS), China
Chong Jin,
University of Pennsylvania,
United States*

***Correspondence:**
*Fei Su
sufei4818@bjhmoh.cn
Yaqun Zhang
zhangyaqun3365@bjhmoh.cn*

†These authors have contributed equally to this work

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

Received: *01 December 2020*
Accepted: *06 January 2021*
Published: *26 January 2021*

Citation:
Zhang L, Zhang W, Li H, Tang X, Xu S, Wu M, Wan L, Su F and Zhang Y (2021) Five Circular RNAs in Metabolism Pathways Related to Prostate Cancer. Front. Genet. 12:636419. doi: 10.3389/fgene.2021.636419

Prostate cancer (PCa) is the most common malignant tumor in men, and its incidence increases with age. Serum prostate-specific antigen and tissue biopsy remain the standard for diagnosis of suspected PCa. However, these clinical indicators may lead to aggressive overtreatment in patients who have been treated sufficiently with active surveillance. Circular RNAs (circRNAs) have been recently recognized as a new type of regulatory RNA that is not easily degraded by RNases and other exonucleases because of their covalent closed cyclic structure. Thus, we utilized high-throughput sequencing data and bioinformatics analysis to identify specifically expressed circRNAs in PCa and filtered out five specific circRNAs for further analysis—hsa_circ_0006410, hsa_circ_0003970, hsa_circ_0006754, hsa_circ_0005848, and a novel circRNA, hsa_circ_AKAP7. We constructed a circRNA-miRNA regulatory network and used miRNA and differentially expressed mRNA interactions to predict the function of the selected circRNAs. Furthermore, survival analysis of their cognate genes and PCR verification of these five circRNAs revealed that they are closely related to well-known PCa pathways such as the MAPK signaling pathway, P53 pathway, androgen receptor signaling pathway, cell cycle, hormone-mediated signaling pathway, and cellular lipid metabolic process. By understanding the related metabolism of circRNAs, these circRNAs could act as metabolic biomarkers, and monitoring their levels could help diagnose PCa. Meanwhile, the exact regulatory mechanism for AR-related regulation in PCa is still unclear. The circRNAs we found can provide new solutions for research in this field.

Keywords: prostate cancer, circular RNA, miRNA-mRNA, pathways, bioinformatics

INTRODUCTION

Prostate cancer (PCa) is a slow-growing malignant tumor, the incidence of which increases with age (Carroll and Mohler, 2018; Etzioni and Nyame, 2020). At the beginning of diagnosis, most patients are asymptomatic; however, it is still among the top three causes of cancer-related deaths in men (Siegel et al., 2017). Patients with a high risk of PCa must undergo periodic testing for serum prostate-specific antigen (PSA). Tissue biopsy remains the care standard for diagnosis for suspected PCa (Litwin and Tan, 2017). After the tumor is confirmed by biopsy, the next step is to determine the invasiveness of tumor cells. The Gleason score is the most commonly used scale to assess the grade of PCa. When the grade is high, tumors tend to spread. Most Gleason scores used to evaluate prostate biopsy samples range from 6 to 10. A score of 6 indicates low-risk PCa; a score of 7 indicates intermediate-risk PCa; a score of 8 to 10 indicates high-risk PCa (2019). However, researchers have found that the "normal" PSA level of 0–4 ng/mL does not guarantee cancer-free status; in approximately 25% of men with a PSA below 4 ng/mL, a biopsy still reveals PCa (Kitagawa et al., 2014). Thus, these clinical indexes cannot guarantee the reliability of diagnosis. New assistant biomarkers need to be developed for the diagnosis of PCa.

With the development of sequencing technology, circRNAs are recognized as a new type of regulatory RNA. They were first identified by analysis using next generation RNA sequencing (RNA-seq) in a study of pediatric acute lymphoblastic leukemia (Salzman et al., 2012). Most circRNAs are composed of protein-coding exons; thus, the expression of these circRNAs competes with the production of pre-mRNAs. These events also lead to the expression of circRNAs being higher than that of their cognate linear RNAs under certain conditions (Jeck et al., 2013; Jeck and Sharpless, 2014). CircRNAs can function by directly regulating gene expression or by acting as miRNA sponges (Tay et al., 2014). They are similar to competitive endogenous RNAs (ceRNAs) and contain miRNA response elements (MREs). Therefore, they can function by competing with mRNAs to bind miRNAs (Hansen et al., 2013). For example, dysregulation of circRNA-0001946 contributes to tumor cell proliferation and metastasis in colorectal cancer by targeting microRNA-135a-5p (Deng et al., 2020). CircRNAs are not easily degraded by RNases and other exonucleases due to a covalent closed cyclic structure without free 5′ or 3′ends. They have a longer half-life (>48 h) than linear RNAs (Suzuki et al., 2006; Jeck and Sharpless, 2014). CircRNAs have been known to be rich in tumors (Salzman et al., 2012). Therefore, compared with other RNAs, circRNAs have more advantages as novel biomarkers of cancer and other diseases (Arnaiz et al., 2019).

Studies have shown that circRNAs are functional in PCa. Overexpression of circ0005276 and its host gene X-linked inhibitor of apoptosis protein (XIAP) can promote cell proliferation, migration, and epithelial–mesenchymal transition in PCa tissues compared with that in normal tissues (Feng et al., 2019). CircRNAs can act as oncogenes in the progression of PCa and are differentially expressed between cancer tissues and normal tissues (Feng et al., 2019). It is also reported that circRNAs can act as therapeutic targets. For example, the overexpression of circRNA cir-ITCH significantly inhibits the proliferation, migration, and invasion of PCa cells. By targeting miR-17 in PC-3 and LNCaP cell lines, circRNAs could act as therapeutic targets in PCa, especially in castration-resistant prostate cancer (CRPC) (Li et al., 2020). CircRNAs also affect carbohydrate, lipid, and amino acid metabolism in cancer. By regulating transcription factors, circRNAs can modulate glycolysis (Yu et al., 2019). Thus, it is important to identify differentially expressed circRNAs in PCa and explore their potential as diagnostic and therapeutic targets in cancer.

In this study, we compared circRNAs between four PCa tissues and two adjacent normal tissues of two PCa patients by sequencing six sets of RNA-seq. We selected five circRNAs that were highly expressed in tumor tissues and found that the fold change in expression of these five circRNAs was significantly higher than that of their cognate linear RNAs. We verified these circRNAs by PCR in PCa cell lines and used the circRNA-miRNA-mRNA method to predict biological pathways regulated by these circRNAs. Some well-known pathways in PCa were enriched, such as the p53 signaling pathway, MAPK signaling pathway, hormone-mediated signaling pathway, and cellular lipid metabolic process. These pathways also confirmed the high reliability of the five circRNAs that participated in the regulation of PCa.

MATERIALS AND METHODS

Patients and Samples

For sequencing samples, two pairs of PCa tissues and adjacent tissues were derived from surgical samples. Sections from normal and malignant tissues were examined after staining with hematoxylin and eosin. The tumor specimen comprised >80% malignant cells, and the benign specimen comprised an approximately equal admixture of normal epithelial and stromal cells. The pathology of the prostate tumor was checked by a pathologist and established as a combined Gleason Score of 6 (3 + 3), stage T2a, with focal involvement of the surgical margin. RNA was purified from minced frozen tissue using Trizol reagent (Life Technologies, Inc., Rockville, MD, United States). Total RNA was briefly treated with DNase I. For each sample, an RNA library was constructed using 3 μg total RNA. Ribo-Zero Gold Kits were used to remove rRNA. According to the instructions of the NEB-Next Ultra Directional RNA Library Prep Kit for Illumina (NEB, Ispawich, United States), different index tags were selected. The constructed libraries were sequenced using Illumina; the sequencing strategy used was PE150.

For qRT-PCR, 20 pairs of PCa tissues and adjacent normal tissues were collected from the Department of Pathology of Beijing Hospital with Gleason scores of 6 (14 cases) and 7 (6 cases). All tissues were fixed in phosphate-buffered formalin, dehydrated with ethanol, and embedded in paraffin. The malignant status and Gleason score were obtained for these samples by histological analysis. The work was approved by the Beijing Hospital Ethics Committee.

None of these patients had undergone hormonal therapy prior to surgery.

Quality Control and Mapping of Sequencing Data

The quality of the fastq data of RNA-seq was evaluated using fastQC (Andrews, 2010). We found some reads mixed with adapters, and then Trimmomatic (Bolger et al., 2014) was used to filter these sequences. The reads with lengths of less than 28 were dropped. The read quality was filtered through a four-base sliding window with an average quality threshold of 15. After the reads passed the sequence quality tests, the filtered reads were mapped to the human hg38 genome (Lander et al., 2001) using the aligner software STAR (Dobin et al., 2013) with parameter "–chimSegmentMin 10."

Differential Expressed circRNAs Filtered

CIRCexplorer2 (Zhang et al., 2016) and CLEAR/CIRCexplorer3 (Ma et al., 2019) were mainly used to obtain circRNAs in our research. For CIRCexplorer2, we used the parse module to analyze circRNA fusion junction reads and annotate modules for circRNA gene information. The expression of circRNAs was quantified by CIRCscore in CLEAR/CIRCexplorer3, which indicates the circRNA expression level by linear RNA expression level adjustment. The expression of fold change between tumor circRNAs and normal adjacent prostatic tissue circRNAs was calculated to filter tumor-specific expressed circRNAs.

Bioinformatics Analysis of CircRNAs

All of the interaction binding sites between circRNAs and miRNAs were downloaded from circBank (Liu et al., 2019). For the novel circRNAs, we used miRNDB (Chen and Wang, 2020) to predict the related miRNAs. We identified the function of the predicted miRNAs by manual literature mining. Then, Cytoscape (Shannon et al., 2003) was used to build a network between circRNAs and miRNAs. For differentially expressed mRNAs, we chose featureCounts (Liao et al., 2014) to quantify read counts for each gene. Based on paired-end data, "requireBothEndsMapped = TRUE" and "isPairedEnd = TRUE" were set additionally. Then, we calculated the normalized expression levels in fragments per kilobase per million mapped reads (FPKM) by using the DGEList and rpkm function from edgeR (Robinson et al., 2010).

To predict circRNA-related pathways, we regarded miRNAs as a middleman to find circRNA-related mRNAs. The interactions between miRNAs and mRNAs were obtained from miRDB (Chen and Wang, 2020). Meanwhile, it provided interaction scores to assess accuracy; only scores higher than 90 and differentially expressed mRNAs were considered for further analysis.

Functional Enrichment Analysis and Survival Analysis

The circRNA-related mRNA list was analyzed using the functional enrichment tool GOseq (Young et al., 2010). Compared with other tools, GOseq can alleviate selection bias more effectively. The pathways were drawn using ggplot2 (Wickham, 2016) in the R language. For survival analysis, TCGA PCa (PRAD) data were selected in UCSC Xena (Goldman et al., 2020) with progression free intervals to draw Kaplan-Meier plots of circRNA cognate genes. The expression level of the gene was used for survival analysis.

RNA Extraction and Real-Time Quantitative PCR (qPCR)

A total of 20 PCa tissues and 20 adjacent normal tissue samples were prepared. RNA was extracted from three 10-μm FPE sections per sample. Paraffin was removed by xylene extraction followed by washing with ethanol. RNA was isolated from the sectioned tissue blocks using the purification kit, total RNA was extracted, and RNA was subjected to DNase I (Invitrogen, AM2222) treatment. The qRT-PCR was performed using the TransScript II Green One-Step qRT-PCR SuperMix kit (TransGen Biotech, AQ311-01) with 100 ng RNA as template in a 20 μL reaction volume on an ABI 7,500 real-time cycler (Qiagen).

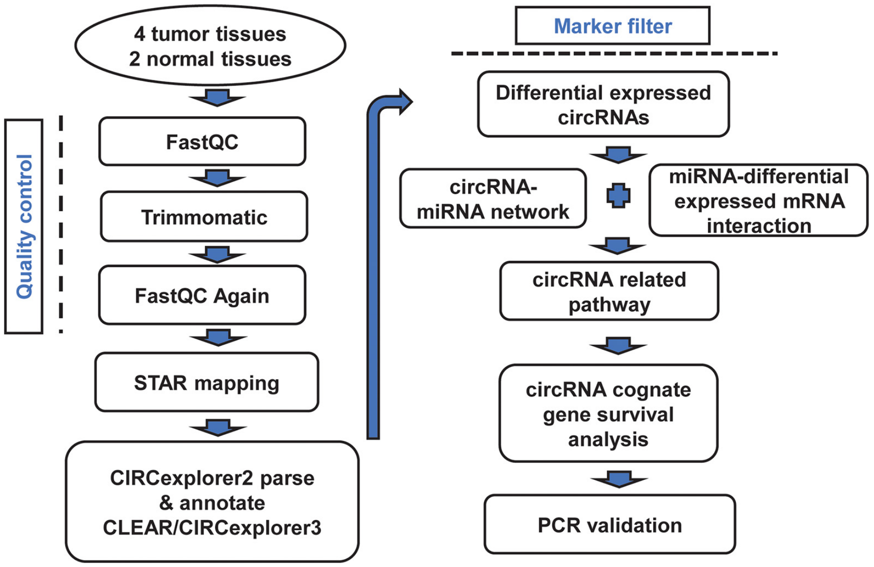


FIGURE 1 | Bioinformatics analysis pipeline of our study.

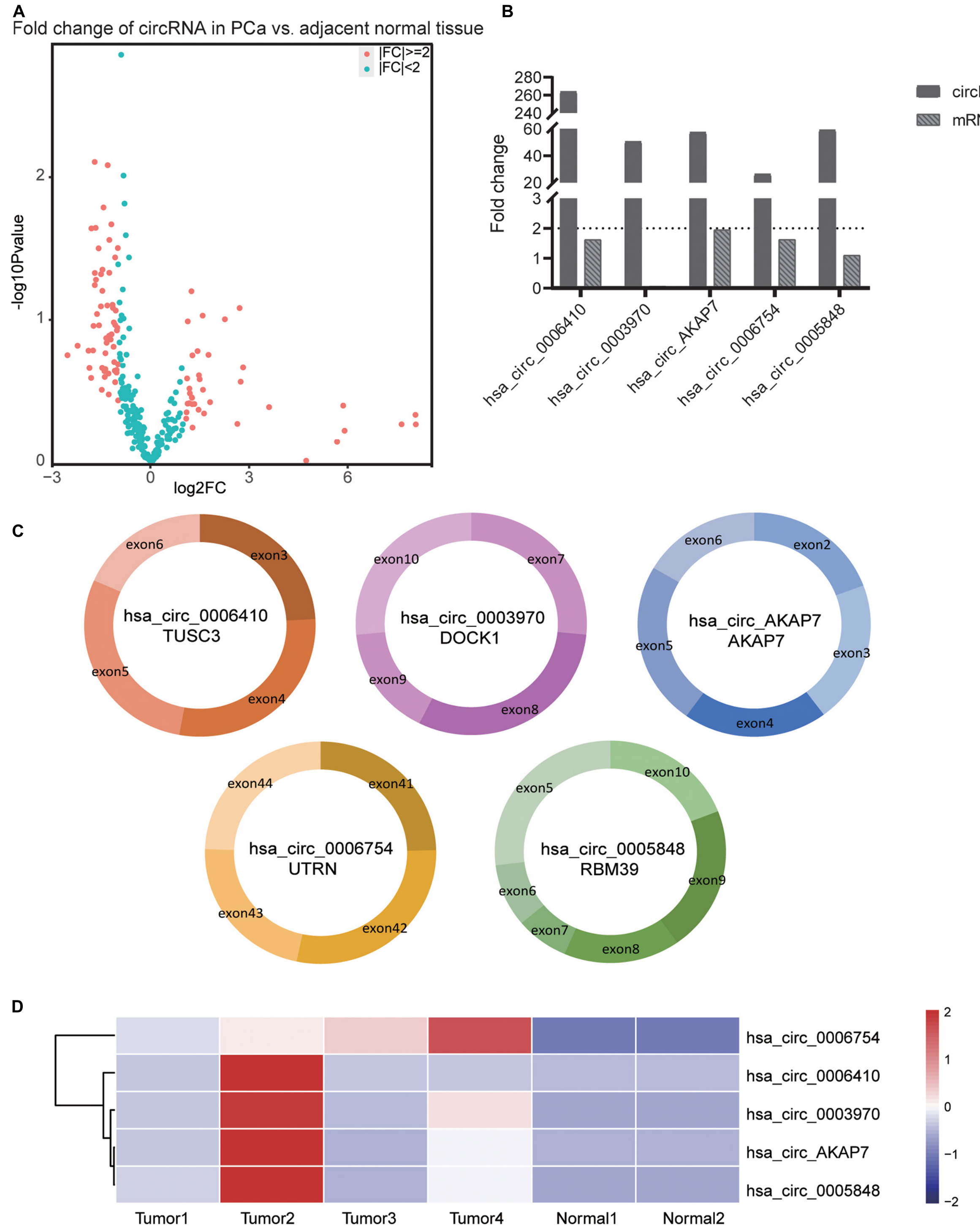


FIGURE 2 | Information of the five PCa specific circRNAs. **(A)** Volcano plot of circRNAs in PCa. The red points represent differentially expressed circRNAs in PCa with fold-change >2 or fold-change <0.5. The paired *t*-test was used to obtain *p*-value. **(B)** The fold-change of circRNAs and its cognate mRNA expression between tumor tissue and adjacent normal tissue. The fold-change of circRNAs was much higher than its mRNAs especially in hsa_circ_0006410. In both tumor as well as normal samples, the expression of hsa_circ_0003970 cognate mRNA was 0. Thus, the mRNA fold change of hsa_circ_0003970 was 0. **(C)** The exon composition of hsa_circ_0006410, hsa_circ_0003970, hsa_circ_0006754, hsa_circ_0005848, and hsa_circ_AKAP7. **(D)** Heatmap of the expression of five specific circRNAs. These five circRNAs are highly expressed in tumor tissues than in normal tissues. The value used in this figure is the expression of circRNAs, and we used "scale = row" to make the graph more consecutive.

TABLE 1 | The detailed information of circRNAs.

circRNA ID	Chrom	Start	End	Strand	Cognate mRNA	Cognate gene	Pearson correlation
hsa_circ_0006410	chr8	15650696	15673836	+	ENST00000382020.8	TUSC3	0.997861
hsa_circ_0003970	chr10	126996747	127000307	+	ENST00000280333.9	DOCK1	0.962136
hsa_circ_AKAP7	chr6	131145284	131199573	+	ENST00000431975.7	AKAP7	−0.31984
hsa_circ_0006754	chr6	144531051	144539443	+	ENST00000367545.7	UTRN	0.841363
hsa_circ_0005848	chr20	35721739	35732135	−	ENST00000639702.1	RBM39	0.97713

The genome version is hg38.

PCR cycling was performed as follows: one cycle at 95°C for 10 min, 95°C for 20 s, and 40 cycles at 60°C for 45 s. The threshold cycle for a given amplification curve during RT-PCR occurs at the point where the fluorescent signal grows beyond a specified fluorescence threshold setting.

The results were normalized with beta actin, and the relative RNA expression was calculated by the 2-ΔΔCt method. To evaluate the statistical significance of PCR data, the paired sample *t*-test was used. The hsa_circ_0003970 primer sequences were as follows: left primer 5′-AGCTGAGGGACAACAACACC-3′; right primer 5′-CCTCTTGTAACCTTTCCTCCA-3′. The hsa_circ_0006410 primer sequences were as follows: left primer 5′-GTGGA ACCATATCCGTGGAC-3′; right primer 5′-GAAAAACGTCT GTCCCCTCA-3′. The hsa_circ_0006754 primer sequences were as follows: left primer 5′-CTGAATTGGAGATGCTTTCAGA-3′; right primer 5′-TGGAGCACAGGTATCAACCA-3′. The hsa_circ_0005848 primer sequences were as follows: left primer 5′-GGGAAGTGCTGGACCTATGA-3′; right primer 5′-TCACGGCTTTTGCTCTTTTT-3′. The hsa_circ_AKAP7 primer sequences were as follows: left primer 5′-AGGCA TCCTGGTAGGAGAGAG-3′; right primer 5′-AGCAAATGG CATGTCTACCA-3′.

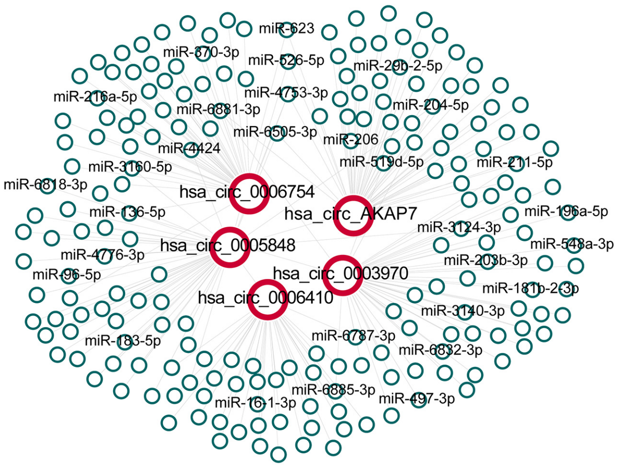


FIGURE 3 | The network between five specific circRNAs and their predicted interactions of miRNAs. All interactions between circRNAs and miRNAs were obtained and extracted for hsa_circ_0006410, hsa_circ_0003970, hsa_circ_0006754, hsa_circ_0005848 and hsa_circ_AKAP7 to build the network.

RESULTS

Differentially Expressed CircRNAs in PCa

Analysis of RNA-seq data (**Figure 1**) showed 89 differentially expressed circRNAs with fold-change >2 or fold-change <0.5. Due to the small sample size in our sequencing control set, the *p*-value was not accurate for initial screening. The differentially expressed circRNAs included 32 upregulated circRNAs and 57 downregulated circRNAs (**Figure 2A**). According to this result, we selected the top five differentially expressed circRNAs for further analysis—included hsa_circ_0006410, hsa_circ_0003970, hsa_circ_0006754, hsa_circ_0005848, and a novel circRNA that we named hsa_circ_AKAP7 (**Figure 2C** and **Table 1**). Because of the circRNA back-splicing feature, an mRNA may correspond to multiple circRNAs. **Figure 2C** shows the precise corresponding exon, enabling the identification of its component. These five circRNAs were highly expressed in tumor tissues compared with their cognate mRNAs (**Figure 2B**). The Pearson correlation coefficients of circRNAs and mRNAs were consistent with those found in previous studies; circRNAs and mRNAs tend to have a high correlation (Yang et al., 2018). The produced circRNAs and their cognate mRNAs usually inhibit each other but **Figure 2B** did not show an opposite trend of circRNA and mRNA expression. This indicated that these five circRNAs have key functions in tumor tissue instead of their cognate mRNAs. The expression of these five circRNAs is depicted in **Figure 2D**. The prominent higher expression of these five specific circRNAs led us to further research. Therefore, we regarded circRNAs as miRNA sponges to explore the functions of these five circRNAs.

CircRNA-miRNA Network and miRNA-Related mRNAs

After obtaining circRNA and miRNA interactions from circBank and miRNDB, we built five circRNA and miRNA interaction networks (**Figure 3**). We found 215 circRNA-miRNA interactions in the network, and each circRNA had an average of 43 miRNA interactions. Many of these miRNAs have been reported to be associated with PCa (**Table 2**). Hsa_circ_0003970 and hsa_circ_0005848 interacted with miRNA-204-5p, which is a tumor suppressor that promotes apoptosis by targeting BCL2 in PCa cells (Lin et al., 2017). Hsa_circ_0005848 and hsa_circ_0006754 related miR-3160-5p is a PCa cell proliferation suppressor that targets the F-box protein (Lin et al., 2018). MiR-548 acts as an anti-oncogenic factor that inhibits the phosphoinositide three-kinase (PI3K)/AKT signaling pathway in lung cancer and is associated with high-risk Gleason scores in PCa (Shi et al., 2015). The PI3K/AKT pathway is involved in tumor immunological surveillance and immune suppression (Dituri et al., 2011). Hsa-miR-181b-2-3p and hsa-miR-96-5p were associated with the androgen receptor and Gleason score (Mekhail et al., 2014). These results further highlight the contribution of this study.

The miRNA target prediction based on the short seed sequence provided many false positive results. Thus, we filtered differentially expressed mRNAs in PCa to analyze miRNA and mRNA interactions. Only scores >90 interactions were selected to predict circRNAs function. Then, we used functional enrichment analysis to explore the function of the five circRNAs.

CircRNA-Related Pathways in Metabolism Pathways

Analysis of these five circRNA-related mRNAs showed that they were all enriched in many well-known PCa pathways (**Figure 4A**). Hsa_circ_0006410, hsa_circ_0003970, hsa_circ_AKAP7, hsa_circ_0006754, and hsa_circ_0005848 were all related to the MAPK signaling pathway. MAPK signaling is an important regulator of cancer, especially PCa. It includes three cross-signaling pathways: p38, JNK, and ERK (Dhillon et al., 2007). Each pathway comprises several levels of kinases. The p38-MAPK pathway is important for the production of inflammatory cytokines and IFN-γ. It can also positively regulate Th1 differentiation instead of Th2 (Martinez et al., 2009). The JNK–MAPK pathway plays pro-inflammatory roles in macrophages, inducing M1 differentiation. Activation of the ERK–MAPK pathway favors cell differentiation into CD4 lineage and is critical for CD4 T cell polarization of Th2 because it is required for IL-4 receptor function (Alessandro et al., 2019). These regulators are significant in PCa.

Other significant pathways were the hormone-mediated signaling pathway and cellular lipid-related process, which were associated with hsa_circ_0006410, hsa_circ_0003970, hsa_circ_AKAP7, hsa_circ_0006754, and hsa_circ_0005848 (**Figures 4B–F**). Steroid androgen hormones play key roles in the progression and treatment of PCa. Androgen deprivation therapy (ADT) is the first-line treatment used to control cancer growth (Munkley et al., 2016). It functions by inhibiting the production of male hormone testosterone and preventing it from reaching PCa cells. ADT can cause apoptosis of PCa cells and can make them grow slowly. Studies have indicated that dietary fat intake is related to PCa development, suggesting that lipid metabolism plays a role in the carcinogenesis and progression of PCa (Tamura et al., 2009). Dysregulation of metabolism of lipids, especially sphingolipid, is a hallmark of the malignant phenotype. Increased lipid accumulation leading to changes in levels of lipid metabolic enzymes has been verified in various tumors, including PCa (Wu et al., 2014). Castration-resistant PCa (CRPC) is considered to utilize *de novo* lipid synthesis to produce fatty acids to obtain energy (Eidelman et al., 2017). The five circRNAs were related to both the hormone-mediated signaling pathway and the lipid-related process, indicating that they are involved in PCa regulation. Several pathways found to be closely related to PCa include the chemokine pathway, cell cycle, p53 signaling

TABLE 2 | The literature mining of circRNAs related miRNAs.

circRNA ID	miRNA ID	miRNA description in PCa or other tumors	References
hsa_circ_0003970	hsa-miR-181b-2-3p	AR signaling in PCa; cancer stem cell (CSC) formation in PCa	Mekhail et al., 2014
hsa_circ_0003970	hsa-miR-196a-5p	Associated with SNPs that can be useful in screening for cancer risk	Mekhail et al., 2014
hsa_circ_0003970	hsa-miR-203b-3p	Anti-metastatic in PCa; epithelial to mesenchymal transition (EMT) in PCa	Mekhail et al., 2014
hsa_circ_0003970	hsa-miR-211-5p	Tumor suppressor by targeting ACSL4 in Hepatocellular Carcinoma	Qin et al., 2020
hsa_circ_0003970	hsa-miR-497-3p	Down-regulated in PCa	Mekhail et al., 2014
hsa_circ_0003970	hsa-miR-548a-3p	Anti-oncogenic factor inhibiting the PI3K/AKT signaling pathway in lung cancer and associated with high-risk Gleason scores in prostate cancer	Li et al., 2013
hsa_circ_0003970; hsa_circ_0005848	hsa-miR-204-5p	Tumor suppressor miRNA-204-5p promotes apoptosis by targeting BCL2 in PCa	Mekhail et al., 2014
hsa_circ_0006754	hsa-miR-216a-5p	Inhibits malignant progression in small cell lung cancer: involvement of the Bcl-2 family proteins	Sun et al., 2018
hsa_circ_0006754	hsa-miR-370-3p	Up-regulated in PCa	Mekhail et al., 2014
hsa_circ_0006754; hsa_circ_AKAP7	hsa-miR-526b-5p	hsa_circ_0085539 promotes osteosarcoma progression by regulating miR-526b-5p and SERP1	Mekhail et al., 2014
hsa_circ_0006410	hsa-miR-16-1-3p	Biochemical failure in PCa	Mekhail et al., 2014
hsa_circ_0005848	hsa-miR-183-5p	Up-regulated in PCa	Mekhail et al., 2014
hsa_circ_0005848	hsa-miR-3160-5p	Suppressed prostate cancer cell proliferation	Lin et al., 2018
hsa_circ_0005848	hsa-miR-96-5p	Biochemical failure in PCa;Gleason score in PCa	Mekhail et al., 2014
hsa_circ_0005848; hsa_circ_AKAP7	hsa-miR-623	Suppressed tumor progression in human lung adenocarcinoma	Wei et al., 2016
has_circ_AKAP7	hsa-miR-206	Anti-metastatic in PCa	Mekhail et al., 2014
has_circ_AKAP7	hsa-miR-29b-2-5p	Anti-metastatic in PCa	Mekhail et al., 2014

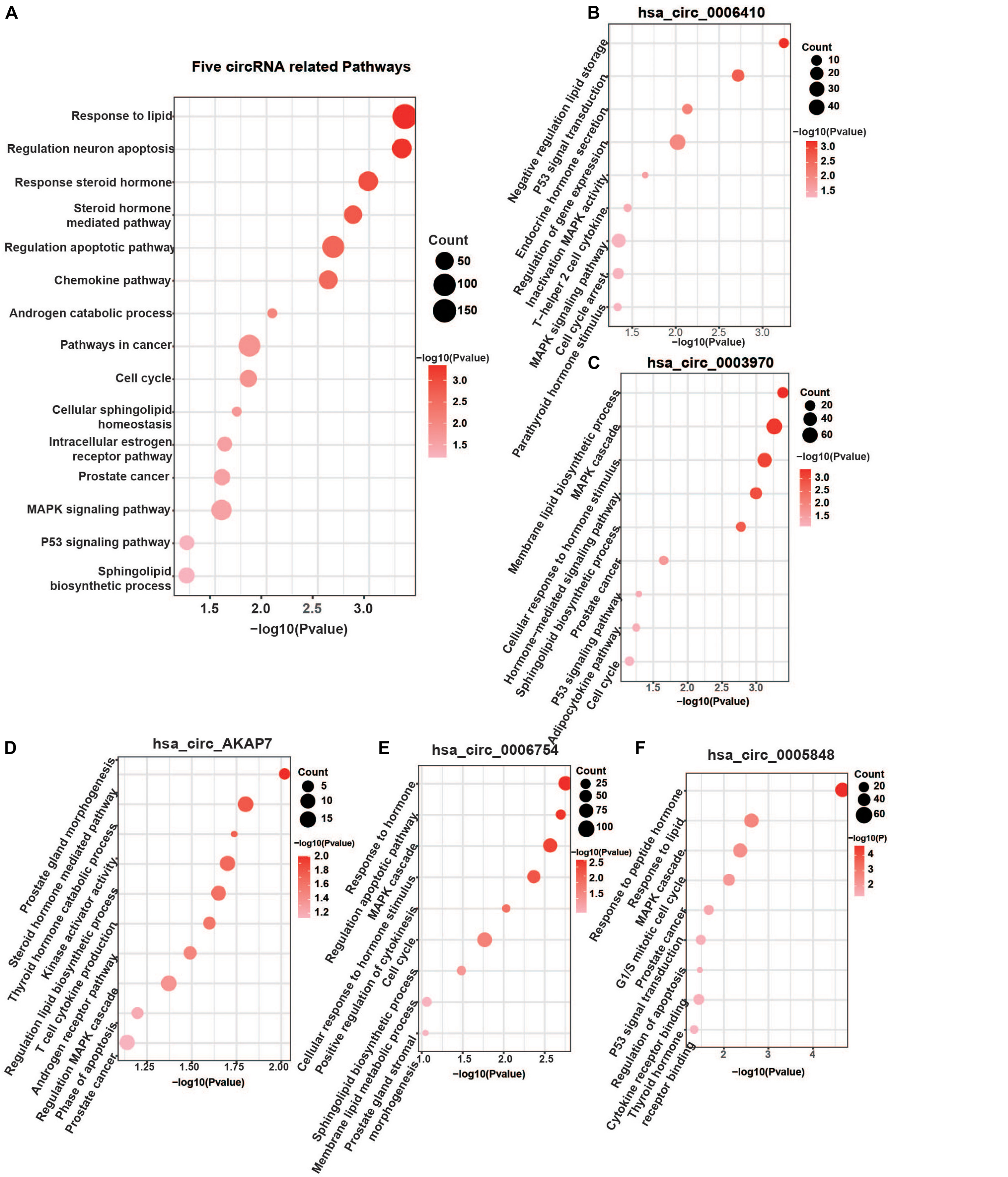


FIGURE 4 | Five specific circRNA-related pathways. **(A)** The five circRNA-related pathways. Only differentially expressed mRNAs with prediction score higher than 90 were considered. Well-known PCa-related pathways, such as the MAPK signaling pathway, P53 pathway, AR pathway, cell cycle, steroid hormone-mediated signaling pathway, and lipid-related process, were all found. **(B)** hsa_circ_0006410 related pathways. **(C)** hsa_circ_0003970 related pathways. **(D)** hsa_circ_AKAP7 related pathways. **(E)** hsa_circ_0006754 related pathways. **(F)** hsa_circ_0005848 related pathways. "Count" represents the number of genes in the relevant categories.

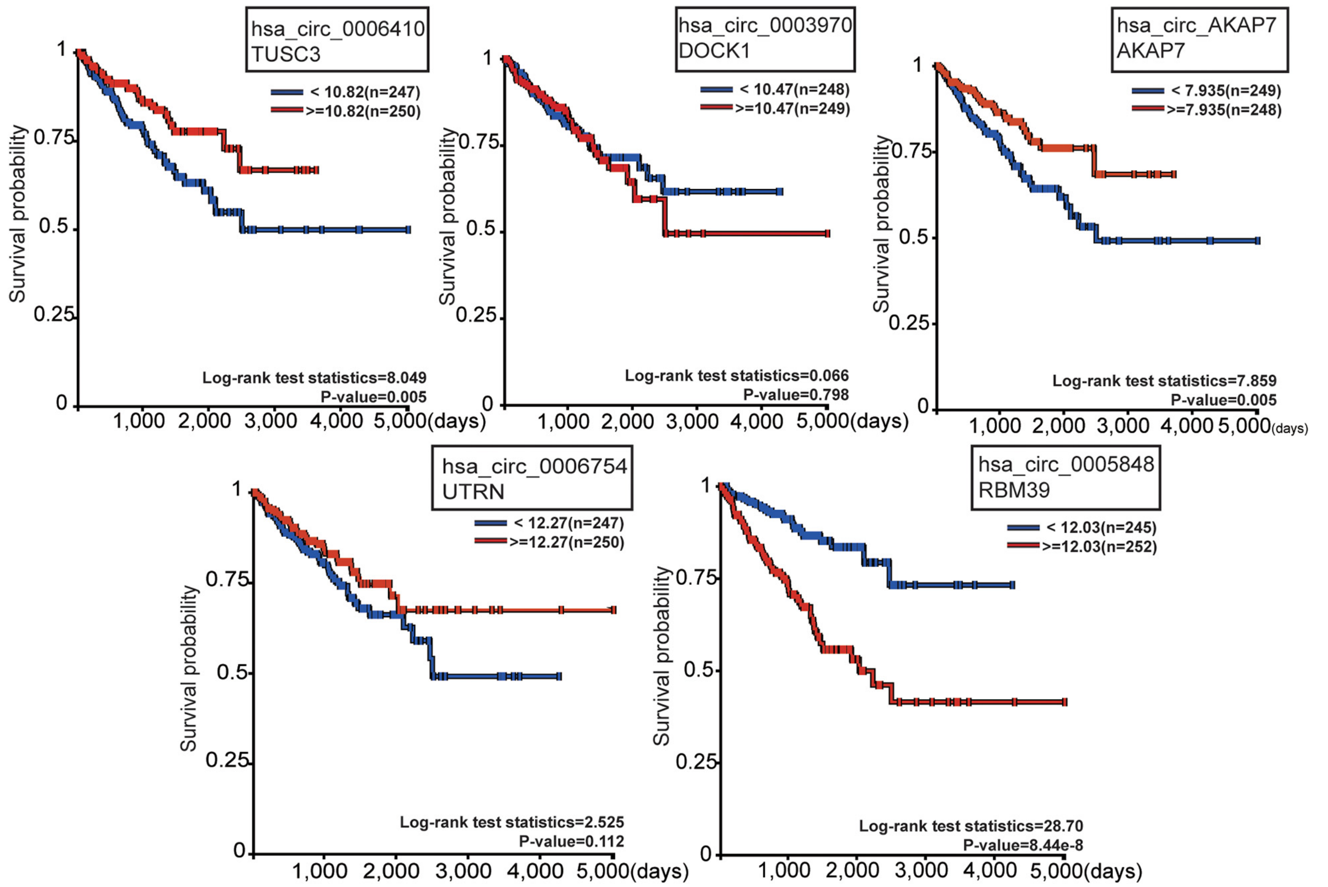


FIGURE 5 | Survival analysis of the five circRNA cognate genes in PCa. Among these, TUSC3, AKAP7, and RBM39 were significantly related with survival probability. The red line represents high gene expression, and the blue line represents low gene expression.

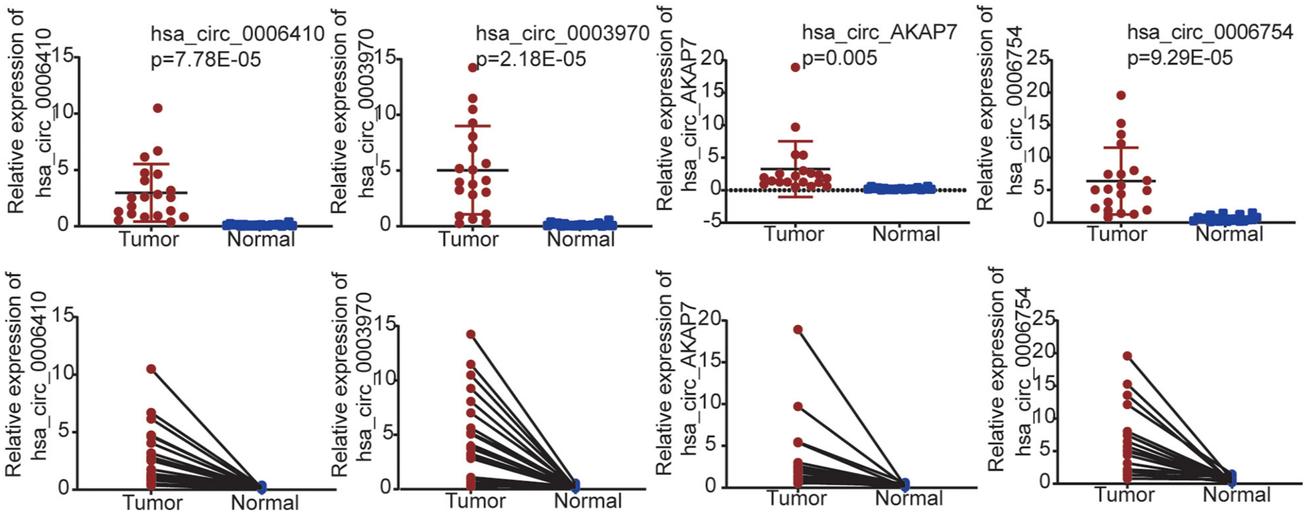


FIGURE 6 | qRT-PCR validation for PCa tissues and adjacent normal tissues in 20 samples from patient samples diagnosed with PCa. Only four circRNAs were significantly validated.

pathway, apoptosis and transcriptional misregulation in cancer (**Figure 4A**).

Survival Analysis of CircRNA Cognate Genes and qRT-PCR Validation of Five CircRNAs

To observe the effect of the five circRNAs in PCa patients, survival analysis of circRNA cognate genes in TCGA data was performed **(Figure 5**). TUSC3 (hsa_circ_0006410 cognate gene), AKAP7 (hsa_circ_AKAP7 cognate gene), and RBM39 (hsa_circ_0005848 cognate gene) were all significantly associated with progression-free survival of PCa patients, as shown by the Kaplan-Meier plot (*P*-value < 0.05) **(Figure 5A**). Patients with high expression of TUSC3 and AKAP7 showed better overall survival. This is consistent with the fact that TUSC3 is a tumor suppressor gene (Yu et al., 2017). Meanwhile, low expression of RBM39 was found to be associated with low overall survival. RBM39 is associated with precursor messenger RNA (pre-mRNA) splicing factors, and inactivation of RBM39 causes aberrant pre-mRNA splicing. Previous studies have shown that several single amino acid substitutions in RBM39 confer resistance to the toxic effects of indisulam in cultured cancer cells and in mice with tumor xenografts (Han et al., 2017). Since the direction of differential expression varied among the five mRNAs, we think that circRNAs might act as oncogenes or tumor suppressor genes in PCa. The direction of different functional circRNAs is different in its cognate mRNAs (**Figure 5**). However, further experiments are required for confirming this.

We used qRT-PCR to validate the five circRNAs in 20 PCa and normal samples **(Figure 6**). In our results, four circRNAs were significantly validated—hsa_circ_0006410, hsa_circ_0003970, hsa_circ_AKAP7, and hsa_circ_0006754. The relative expression of circRNAs indicated that these circRNAs were highly expressed in tumor tissues compared to normal tissues and validated our analysis.

DISCUSSION

Based on sequencing data of PCa tissues and adjacent normal tissues, we identified differentially expressed circRNAs in PCa. We filtered five specific highly expressed circRNAs that had never been studied in PCa before for further analysis. We found that the miRNAs and mRNA pathways related to these circRNAs were related to known metabolic pathways, such as PI3K-Akt signaling pathway, MAPK signaling pathway, and lipid metabolic process. This also confirmed the reliability of our findings. Through bioinformatics analysis, we analyzed expression levels of circRNA through linear RNA expression level adjustment using CIRCexplorer3. For these five circRNAs, the expression of circRNAs and mRNAs in tumor tissue was highly correlated, which is consistent with the results of previous studies. However, the fold changes of circRNAs expression were notably larger than those of their cognate mRNAs, suggesting that circRNAs play a role in tumor tissues. The underlying mechanism, however, is still unknown and requires further research.

We predicted circRNAs function by using circRNA-miRNA-mRNA interactions and showed that they were all significantly enriched in the lipid metabolism pathway. The link between PCa development and lipid metabolism is well established, with AR intimately involved in a number of lipogenic processes. Altered lipid signatures may offer insights into metabolic reprogramming. Lipid pathway deregulation in advanced PCa is a hot research field to identify a therapeutic pathway. Several therapeutic agents, such as warfarin, atostatin, and orlistat, are known to block key processes in lipid metabolism and negatively influence PCa progression. Lipid metabolism is also activated by the PI3K-Akt signaling pathway by sterol regulatory element-binding protein 1 (SREBP1) (Edlind and Hsieh, 2014). The hsa_circ_0006410, hsa_circ_0003970, hsa_circ_0006754, hsa_circ_0005848, and hsa_circ_AKAP7 were all enriched in lipid related pathways, which shows their potential as targets.

The PI3K-Akt signaling pathway is deregulated in 42% of localized disease and 100% of advanced-stage disease in PCa. This implies that the alteration of this pathway is another factor in the development of CRPC. Gene amplifications, mutations, and changes in mRNA expression of PI3K signaling pathway are highly correlated with PCa patients (Edlind and Hsieh, 2014). Hsa_circ_0006410, hsa_circ_0003970, hsa_circ_0005848, and hsa_circ_AKAP7 were all related with this pathway. CircRNAs have been reported to activate the PI3K/Akt signaling pathway by regulating gene expression in PCa (Wang et al., 2020). Although the exact mechanism affecting PI3K/Akt signaling pathway is unclear, the circRNAs identified in this study also provided support for this field.

We used survival analysis and qRT-PCR to validate our findings. Survival analysis is a good indicator to assess the function of genes. Three cognate genes of these five circRNAs were significantly identified in survival analysis, alluding that their cognate genes were key genes in regulating tumor progression. Meanwhile, four circRNAs were well verified by qRT-PCR, except for hsa_circ_0005848. We inferred that this may be due to the space structure or the false positive expression of hsa_circ_0005848.

Our research was based on the bioinformatics analysis of RNA-seq between prostate tumor tissues and adjacent normal tissues. We found five specific circRNAs that were highly related to the AR signaling pathway, MAPK signaling pathway, hormone-mediated signaling pathway, and cellular lipid metabolic process. Furthermore, survival analysis and qRT-PCR validation also verified that the circRNAs were closely related to tumor progression of PCa. These five circRNAs can provide new solutions for research in this field.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Genome Sequence Archive for Human (https://bigd.big.ac.cn/bioproject/browse/PRJCA003890) (accession: PRJCA003890).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Beijing Hospital 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

LZ and WZ conceived the study and wrote the manuscript. LZ, FS, and YZ designed the detail analysis pipeline. LZ and FS did the bioinformatics analysis. WZ, HL, XT, and SX performed the experiments. MW and LW participated in revising the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (Grant 81902618), CAMS Innovation Fund for Medical Sciences (2018-I2M-1-002), Beijing Hospital Nova Project (No. BJ-2020-083), and Beijing Hospital Project (No. BJ-2019-153). Beijing Dongcheng District Outstanding Talent Funding Project (BJ-2020-047 and 2019WJGW-10-01).

ACKNOWLEDGMENTS

The authors would like to thank National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences for the data transmission. We would like to thank editage (www.editage.cn) for English language editing.

REFERENCES

Alessandro, M. S., Golombek, D. A., and Chiesa, J. J. (2019). Protein kinases in the photic signaling of the mammalian circadian clock. *Yale J. Biol. Med.* 92, 241–250.

Andrews, S. (2010). *Fastqc: A Quality Control Tool for High Throughput Sequence Data*. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed August 10, 2020).

Arnaiz, E., Sole, C., Manterola, L., Iparraguirre, L., Otaegui, D., and Lawrie, C. H. (2019). CircRNAs and cancer: biomarkers and master regulators. *Semin. Cancer Biol.* 58, 90–99. doi: 10.1016/j.semcancer.2018.12.002

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

Carroll, P. H., and Mohler, J. L. (2018). NCCN guidelines updates: prostate cancer and prostate cancer early detection. *J. Natl. Compr. Canc. Netw.* 16, 620–623. doi: 10.6004/jnccn.2018.0036

Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* 48, D127–D131. doi: 10.1093/nar/gkz757

Deng, Z., Li, X., Wang, H., Geng, Y., Cai, Y., Tang, Y., et al. (2020). Dysregulation of circRNA_0001946 contributes to the proliferation and metastasis of colorectal cancer cells by targeting microRNA-135a-5p. *Front. Genet.* 11:357. doi: 10.3389/fgene.2020.00357

Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene* 26, 3279–3290. doi: 10.1038/sj.onc.1210421

Dituri, F., Mazzocca, A., Giannelli, G., and Antonaci, S. (2011). PI3K functions in cancer progression, anticancer immunity and immune evasion by tumors. *Clin. Dev. Immunol.* 2011:947858. doi: 10.1155/2011/947858

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Edlind, M. P., and Hsieh, A. C. (2014). PI3K-AKT-mTOR signaling in prostate cancer progression and androgen deprivation therapy resistance. *Asian J. Androl.* 16, 378–386. doi: 10.4103/1008-682X.122876

Eidelman, E., Twum-Ampofo, J., Ansari, J., and Siddiqui, M. M. (2017). The metabolic phenotype of prostate cancer. *Front. Oncol.* 7:131. doi: 10.3389/fonc.2017.00131

Etzioni, R., and Nyame, Y. A. (2020). Prostate cancer screening guidelines for Black men: spotlight on an empty stage. *J. Natl. Cancer Inst.* djaa172. doi: 10.1093/jnci/djaa172

Feng, Y., Yang, Y., Zhao, X., Fan, Y., Zhou, L., Rong, J., et al. (2019). Circular RNA circ0005276 promotes the proliferation and migration of prostate cancer cells by interacting with FUS to transcriptionally activate XIAP. *Cell Death Dis.* 10:792. doi: 10.1038/s41419-019-2028-9

Goldman, M. J., Craft, B., Hastie, M., Repecka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* 38, 675–678. doi: 10.1038/s41587-020-0546-8

Han, T., Goralski, M., Gaskill, N., Capota, E., Kim, J., Ting, T. C., et al. (2017). Anticancer sulfonamides target splicing by inducing RBM39 degradation via recruitment to DCAF15. *Science* 356:eaal3755. doi: 10.1126/science.aal3755

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

Jeck, W. R., and Sharpless, N. E. (2014). Detecting and characterizing circular RNAs. *Nat. Biotechnol.* 32, 453–461. doi: 10.1038/nbt.2890

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157. doi: 10.1261/rna.035667.112

Kitagawa, Y., Ueno, S., Izumi, K., Kadono, Y., Konaka, H., Mizokami, A., et al. (2014). Cumulative probability of prostate cancer detection in biopsy according to free/total PSA ratio in men with total PSA levels of 2.1-10.0 ng/ml at population screening. *J. Cancer Res. Clin. Oncol.* 140, 53–59. doi: 10.1007/s00432-013-1543-9

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Li, S., Yu, C., Zhang, Y., Liu, J., Jia, Y., Sun, F., et al. (2020). Circular RNA cir-ITCH is a potential therapeutic target for the treatment of castration-resistant prostate cancer. *Biomed. Res. Int.* 2020:7586521. doi: 10.1155/2020/7586521

Li, Y., Xie, J., Xu, X., Wang, J., Ao, F., Wan, Y., et al. (2013). MicroRNA-548 down-regulates host antiviral response via direct targeting of IFN-lambda1. *Protein Cell* 4, 130–141. doi: 10.1007/s13238-012-2081-y

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Lin, P., Zhu, L., Sun, W., Yang, Z., Sun, H., Li, D., et al. (2018). Prostate cancer cell proliferation is suppressed by microRNA-3160-5p via targeting of F-box and WD repeat domain containing 8. *Oncol. Lett.* 15, 9436–9442. doi: 10.3892/ol.2018.8505

Lin, Y. C., Lin, J. F., Tsai, T. F., Chou, K. Y., Chen, H. E., and Hwang, T. I. (2017). Tumor suppressor miRNA-204-5p promotes apoptosis by targeting BCL2 in prostate cancer cells. *Asian J. Surg.* 40, 396–406. doi: 10.1016/j.asjsur.2016.07.001

Litwin, M. S., and Tan, H. J. (2017). The diagnosis and treatment of prostate cancer: a review. *JAMA* 317, 2532–2542. doi: 10.1001/jama.2017.7248

Liu, M., Wang, Q., Shen, J., Yang, B. B., and Ding, X. (2019). Circbank: a comprehensive database for circRNA with standard nomenclature. *RNA Biol.* 16, 899–905. doi: 10.1080/15476286.2019.1600395

Ma, X. K., Wang, M. R., Liu, C. X., Dong, R., Carmichael, G. G., Chen, L. L., et al. (2019). CIRCexplorer3: a Clear pipeline for direct comparison of circular and linear RNA expression. *Genom. Proteom. Bioinform.* 17, 511–521. doi: 10.1016/j.gpb.2019.11.004

Martinez, F. O., Helming, L., and Gordon, S. (2009). Alternative activation of macrophages: an immunologic functional perspective. *Annu. Rev. Immunol.* 27, 451–483. doi: 10.1146/annurev.immunol.021908.132532

Mekhail, S. M., Yousef, P. G., Jackinsky, S. W., Pasic, M., and Yousef, G. M. (2014). miRNA in prostate cancer: new prospects for old challenges. *EJIFCC* 25, 79–98.

Munkley, J., Vodak, D., Livermore, K. E., James, K., Wilson, B. T., Knight, B., et al. (2016). Glycosylation is an androgen-regulated process essential for prostate cancer cell viability. *EBioMedicine* 8, 103–116. doi: 10.1016/j.ebiom.2016.04.018

PCa (2019). NICE guidance – prostate cancer: diagnosis and management: (c) NICE (2019) Prostate cancer: diagnosis and management. *BJU Int.* 124, 9–26. doi: 10.1111/bju.14809

Qin, X., Zhang, J., Lin, Y., Sun, X. M., Zhang, J. N., and Cheng, Z. Q. (2020). Identification of MiR-211-5p as a tumor suppressor by targeting ACSL4 in *Hepatocellular carcinoma*. *J. Transl. Med.* 18:326. doi: 10.1186/s12967-020-02494-7

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

Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., and Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* 7:e30733. doi: 10.1371/journal.pone.0030733

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shi, Y., Qiu, M., Wu, Y., and Hai, L. (2015). MiR-548-3p functions as an anti-oncogenic regulator in breast cancer. *Biomed. Pharmacother.* 75, 111–116. doi: 10.1016/j.biopha.2015.07.027

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *CA Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387

Sun, Y., Hu, B., Wang, Y., Li, Z., Wu, J., Yang, Y., et al. (2018). miR-216a-5p inhibits malignant progression in small cell lung cancer: involvement of the Bcl-2 family proteins. *Cancer Manag. Res.* 10, 4735–4745. doi: 10.2147/CMAR.S178380

Suzuki, H., Zuo, Y., Wang, J., Zhang, M. Q., Malhotra, A., and Mayeda, A. (2006). Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. *Nucleic Acids Res.* 34:e63. doi: 10.1093/nar/gkl151

Tamura, K., Makino, A., Hullin-Matsuda, F., Kobayashi, T., Furihata, M., Chung, S., et al. (2009). Novel lipogenic enzyme ELOVL7 is involved in prostate cancer growth through saturated long-chain fatty acid metabolism. *Cancer Res.* 69, 8133–8140. doi: 10.1158/0008-5472.CAN-09-0775

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

Wang, Y., Yin, L., and Sun, X. (2020). CircRNA hsa_circ_0002577 accelerates endometrial cancer progression through activating IGF1R/PI3K/Akt pathway. *J. Exp. Clin. Cancer Res.* 39:169. doi: 10.1186/s13046-020-01679-8

Wei, S., Zhang, Z. Y., Fu, S. L., Xie, J. G., Liu, X. S., Xu, Y. J., et al. (2016). Hsa-miR-623 suppresses tumor progression in human lung adenocarcinoma. *Cell Death Dis.* 7:e2388. doi: 10.1038/cddis.2016.260

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wu, X., Daniels, G., Lee, P., and Monaco, M. E. (2014). Lipid metabolism in prostate cancer. *Am. J. Clin. Exp. Urol.* 2, 111–120.

Yang, Q., Wu, J., Zhao, J., Xu, T., Zhao, Z., Song, X., et al. (2018). Circular RNA expression profiles during the differentiation of mouse neural stem cells. *BMC Syst. Biol.* 12(Suppl. 8):128. doi: 10.1186/s12918-018-0651-1

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14. doi: 10.1186/gb-2010-11-2-r14

Yu, T., Wang, Y., Fan, Y., Fang, N., Wang, T., Xu, T., et al. (2019). CircRNAs in cancer metabolism: a review. *J. Hematol. Oncol.* 12:90. doi: 10.1186/s13045-019-0776-8

Yu, X., Zhai, C., Fan, Y., Zhang, J., Liang, N., Liu, F., et al. (2017). TUSC3: a novel tumour suppressor gene and its functional implications. *J. Cell. Mol. Med.* 21, 1711–1718. doi: 10.1111/jcmm.13128

Zhang, X. O., Dong, R., Zhang, Y., Zhang, J. L., Luo, Z., Zhang, J., et al. (2016). Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* 26, 1277–1287. doi: 10.1101/gr.202895.115

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.

ORIGINAL RESEARCH
published: 11 February 2021
doi: 10.3389/fgene.2021.625236

Identification of Prognostic Stromal-Immune Score–Based Genes in Hepatocellular Carcinoma Microenvironment

Shanshan Liu[1,2], Guangchuang Yu[3], Li Liu[1,4]*, Xuejing Zou[1,2], Lang Zhou[3], Erqiang Hu[3] and Yang Song[1,5]*

[1] Country Guangdong Provincial Key Laboratory of Viral Hepatitis Research, Hepatology Unit and Department of Infectious Diseases, Nanfang Hospital, Southern Medical University, Guangzhou, China, [2] State Key Laboratory of Organ Failure Research, Nanfang Hospital, Southern Medical University, Guangzhou, China, [3] Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, China, [4] Department of Medical Quality Management, Nanfang Hospital, Southern Medical University, Guangzhou, China, [5] Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Yang Xie,
Weill Cornell Medicine, United States
Mengqi Zhang,
University of Pennsylvania,
United States
Kuixi Zhu,
University of Arizona, United States

****Correspondence:***
Guangchuang Yu
gcyu1@smu.edu.cn
Li Liu
liuli@i.smu.edu.cn

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

Received: *02 November 2020*
Accepted: *06 January 2021*
Published: *11 February 2021*

Citation:
Liu S, Yu G, Liu L, Zou X, Zhou L, Hu E and Song Y (2021) Identification of Prognostic Stromal-Immune Score–Based Genes in Hepatocellular Carcinoma Microenvironment. Front. Genet. 12:625236. doi: 10.3389/fgene.2021.625236

A growing amount of evidence has suggested the clinical importance of stromal and immune cells in the liver cancer microenvironment. However, reliable prognostic signatures based on assessments of stromal and immune components have not been well-established. This study aimed to identify stromal-immune score–based potential prognostic biomarkers for hepatocellular carcinoma. Stromal and immune scores were estimated from transcriptomic profiles of a liver cancer cohort from The Cancer Genome Atlas using the ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumors using Expression data) algorithm. Least absolute shrinkage and selection operator (LASSO) algorithm was applied to select prognostic genes. Favorable overall survivals and progression-free interval were found in patients with high stromal score and immune score, and 828 differentially expressed genes were identified. Functional enrichment analysis and protein–protein interaction networks further showed that these genes mainly participated in immune response, extracellular matrix, and cell adhesion. *MMP9* (matrix metallopeptidase 9) was identified as a prognostic tumor microenvironment–associated gene by using LASSO and TIMER (Tumor IMmune Estimation Resource) algorithms and was found to be positively correlated with immunosuppressive molecules and drug response.

Keywords: liver cancer, ESTIMATE, bioinformatics analysis, biomarker, tumor-microenvironment

INTRODUCTION

Hepatocellular carcinoma (HCC) is the third leading cause of cancer death worldwide. The median survival of HCC patients in China is about 23 months, and ≥ 60% of patients present with intermediate-stage or advanced-stage HCC (Kanwal and Singal, 2019; Yang et al., 2019). Currently, the main treatment for HCC patients in early stages is surgery, combination with transarterial chemoembolization, ablation, and liver transplantation. For others in advanced stages, the effective approaches involve molecular targeting agents (sorafenib, lenvatinib, and regorafenib). Although these methods have improved the prognosis of HCC patients, the overall survival (OS) of HCC remains challenging for the heterogeneity of HCC. And also, there is still a lack of molecular markers used in determination of prognosis and treatment for patients (Bruix et al., 2016).

The liver cancer microenvironment consists of not only tumor cells but also stromal cells, including distinct immune cell subsets. Tumor-infiltrating immune cells and stromal cells are associated with angiogenesis, immune suppression, chemotherapeutic resistance, and tumor cell migration (Affo et al., 2017; Barry et al., 2020; Jin and Jin, 2020; Son et al., 2020; Zhang et al., 2020). An increasing amount of evidence has suggested the clinical importance of stromal cells and immune cells in the microenvironment of liver cancer tissues, tumor microenvironment (TME)–associated genes also have potential as novel biomarkers for a range of cancers (Yang et al., 2020).

In the present study, the Estimation of STromal and Immune cells in MAlignant Tumors using Expression data (ESTIMATE) algorithm (Yoshihara et al., 2013) was applied to estimate the stromal and immune scores of a series of cancer tissues based on their transcriptional profiles, to perform a comprehensive analysis of immune and stromal cells, and to correlate the data to clinical outcomes of patients.

The least absolute shrinkage and selection operator (LASSO) method is a compressed estimation used to obtain a refined model by constructing a penalty function (Korenberg, 2006). It can help with the selection of variables at the time of parameter estimation so as to better solve the multicollinearity problem of regression analysis. A growing body of research confirms that LASSO is an effective method for gene selection of tumors (Wang et al., 2020; Xu et al., 2020).

Tumor IMmune Estimation Resource (TIMER) integrates multiple state-of-the-art algorithms for immune infiltration estimation, which can explore various associations between immune infiltrates and genetic features in The Cancer Genome Atlas (TCGA) cohorts (Li et al., 2017, 2020). Computational Analysis of REsistance (CARE) is a computational method focused on targeted therapies, to infer genome-wide transcriptomic signatures of drug efficacy from cell line compound screens (Jiang et al., 2018). Previous studies have confirmed that the efficacy of immunotherapy is strongly influenced by the composition and abundance of immune cells in the TME (Boyero et al., 2020).

Thus, we combined LASSO, TIMER algorithms, and CARE to preliminarily demonstrate that the expression of TME-associated genes could be new prognostic and reliable drug response biomarkers for HCC patients.

MATERIALS AND METHODS

Database

In total, data from 365 HCC patients and 18,161 RNAs extracted from RNA-seq data according to ENSEMBL Genomes (hg38) were analyzed in this study. All RNA expression data and the corresponding clinical data were obtained from TCGA (data version, July 19, 2019)[1]. The clinicopathological characteristics of the analyzed patients are listed in **Supplementary Table 1**. The progression-free interval (PFI) is characterized as a time without a new tumor occurrence or a death from cancer. The Estimation of STromal and Immune cells in MAlignant Tumors using Expression data (ESTIMATE) algorithm was applied to the normalized expression matrix for estimating the stromal and immune scores by using "estimate" R package in R software (version: 3.6.3) for each HCC sample.

Correlations Between Prognoses and Stromal/Immune Scores

OS and PFI was used as the primary prognosis endpoint and was estimated by the GraphPad Prism 8.0. **Supplementary Figure 3B** is realized by R package "Survival" (Therneau, 2020), "Survminer" (Kassambara et al., 2019), and "timeROC" (Paul Blanche and Jacqmin-Gadda, 2013). Based on the stromal and immune scores estimated from each sample, patients were classified into two groups by using X-tile, and prognoses for each group were examined. The bioinformatics tool, X-tile (Camp et al., 2004), was used to determine the optimum cutoff point according to the minimum *P*-value defined by the Kaplan–Meier analysis and log-rank test. The principle of X-tile is "enumeration method that different values are grouped as truncation values to conduct statistical tests, and the test result with the lowest *P*-value can be considered as the best truncation value. The survival outcomes of the two groups were compared by log-rank tests. $P < 0.05$ was considered as statistically significant.

Identification of Differentially Expressed Genes

Data analysis was performed using an open-source web tool NetworkAnalyst[2] (Xia et al., 2013a,b; Zhou et al., 2019). Log2 fold change > 1 and adjusted $P < 0.05$ were set as the cutoffs to screen for differentially expressed genes (DEGs). A website Venn diagrams tool (Bardou et al., 2014)[3]was used to identify the commonly upregulated or downregulated DEGs in the immune and stromal groups. Heatmaps and clustering were generated using the R package "ggplot2" (Wickham, 2016), "ggtree" (Yu, 2020b), and "aplot" (Yu, 2020a).

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Pathway Enrichment Analyses

GO (Gene Ontology) enrichment analyses were performed by the "Goseq" (Young et al., 2010) R package, and visualization of bubble diagrams used Hiplot[4]. KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analyses and visualization of intersection genes were performed by the "clusterProfiler"

[1] https://xenabrowser.net

[2] https://www.networkanalyst.ca/NetworkAnalyst/home.xhtml

[3] http://www.ehbio.com/test/venn/#/

[4] https://hiplot.com.cn

(Yu et al., 2012) R package and "enrichplot" (Yu, 2019) R package with $P < 0.05$ as the cutoff value.

Protein–Protein Interaction Network Construction

The protein–protein interaction (PPI) network was retrieved from Search Tool for the Retrieval of Interaction Gene/Proteins (STRING) (Szklarczyk et al., 2019) database with high confidence (0.7) and reconstructed via the Cytoscape software (Shannon et al., 2003). In Cytoscape, we used Molecular COmplex DEtection (MCODE) (Bader and Hogue, 2003) to select two clusters that contained the largest number of nodes. ClueGo (Bindea et al., 2009) App was used to perform enrichment analysis of each cluster selected by MCODE.

Identification of TME-Associated Prognostic Genes

LASSO algorithm was used to identify candidate genes by "glmnet" (Friedman et al., 2010) R package with the number of lambda = 1,000. Clinical outcomes and gene expression profiles were analyzed by LASSO. Lambda.min is the cutoff point that brings minimum mean cross-validated error. Genes with the highest lambda values were selected for further analysis.

Identification of TME-Associated Prognostic Genes

The TIMER algorithm was used to calculate the tumor abundance of six infiltrating immune cells ($CD4^+$ T cells, $CD8^+$ T cells, B cells, neutrophils, macrophages, and dendritic cells) based on RNA-Seq expression profiles data. The correlation between the selected prognostic genes and immune cells was calculated by Spearman correlation analysis by TIMER. The estimation results were calculated by TIMER2.0, CIBERSORT, quanTIseq, xCell, MCP-counter, and EPIC methods. Relations between immunoinhibitors and expression of matrix metallopeptidase 9 (*MMP9*) were calculated by Spearman correlation analysis by a web tool TISIDB[5] (Ru et al., 2019). The correlation coefficient value <0.3 indicates that the correlation is negligible, whereas the correlation coefficient ≥ 0.3 indicates a positive/negative correlation. The CARE software[6] was used to identify genome-scale biomarkers of targeted therapy response using compound screen data. For each gene, the CARE score indicates the association between its molecular alteration and drug efficacy. A positive score indicates a higher expression value (or presence of mutation) to be associated with drug response, whereas a negative score indicates drug resistance.

Statistical Analysis

Unpaired *t*-test was used to compare two groups of continuously distributed variables. Jonckheere–Terpstra test was used to compare three or more groups of continuously distributed variables. The FDR correction was performed in multiple tests. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

RESULTS

Association of Stromal and Immune Scores With HCC Pathology and Prognosis

A cohort containing 365 liver hepatocellular carcinoma patients with available expression data and clinical information in TCGA database was analyzed. The general pipeline of the data analysis protocol is shown in **Figure 1**, and the links of tools are listed in **Supplementary Table 6**. The clinicopathological characteristics of the analyzed patients are listed in **Supplementary Table 1**. Based on the gene expression data, immune and stromal scores were calculated using the ESTIMATE algorithm (**Supplementary Table 2**). The associations of stromal and immune scores with HCC patient pathological characteristics were examined by comparing the score distributions among different tumor stages and differentiation grades.

Significant associations were observed between stromal scores and tumor differentiation grades; tumors with poorer differentiation yielded higher stromal scores than those differentiated well (Jonckheere–Terpstra test, $P = 0.002$) (**Figures 2A,B**).

As previously described, serum α-fetoprotein (AFP) values are not only of diagnostic value but also of prognostic significance in patients with HCC (Galle et al., 2019). Thus, we compared changes in immune and stromal scores between AFP low (AFP ≤ 400 ng/mL) and high (AFP > 400 ng/mL) samples. The AFP high cases had the lowest stromal scores (unpaired *t*-test, $P = 0.0204$) (**Figure 2D**). Evidence suggests that AFP plays an immune-suppressing role (Yang et al., 2018), but we found that there is no significant difference in the immune score as shown in **Figure 2C**. We further used the TIMER algorithm to evaluate the effect of AFP on the immune infiltration of HCC, and results showed that the expression of AFP was weakly correlated with the infiltration abundance of the six immune cells (**Supplementary Figure 1A**). AFP is dynamic in the occurrence and development of HCC, whereas TCGA patients were only tested for AFP at the time of initial diagnosis, which may lead to the bias of the results in our study.

Also, when we compared the immune and stromal scores between patients with a new tumor event and without new tumor event after initial treatment, patients without a new tumor event had higher immune and stromal scores (unpaired *t*-test, $P = 0.0461$ for stromal score and $p = 0.1966$ for immune score) (**Figures 2E,F**).

We also analyzed the correlation between other clinical factors and the immune profile, but found no statistically significant difference (**Supplementary Figures 1B,C**).

The association of stromal and immune scores with HCC prognosis was evaluated by dividing patients optimally into two groups based on their scores by using X-tile (see section "Materials and Methods" for details). We found that the high immune score and stromal score positively correlated with both OS and PFI (**Figures 3A–D**).

[5] http://cis.hku.hk/TISIDB/

[6] http://care.dfci.harvard.edu/

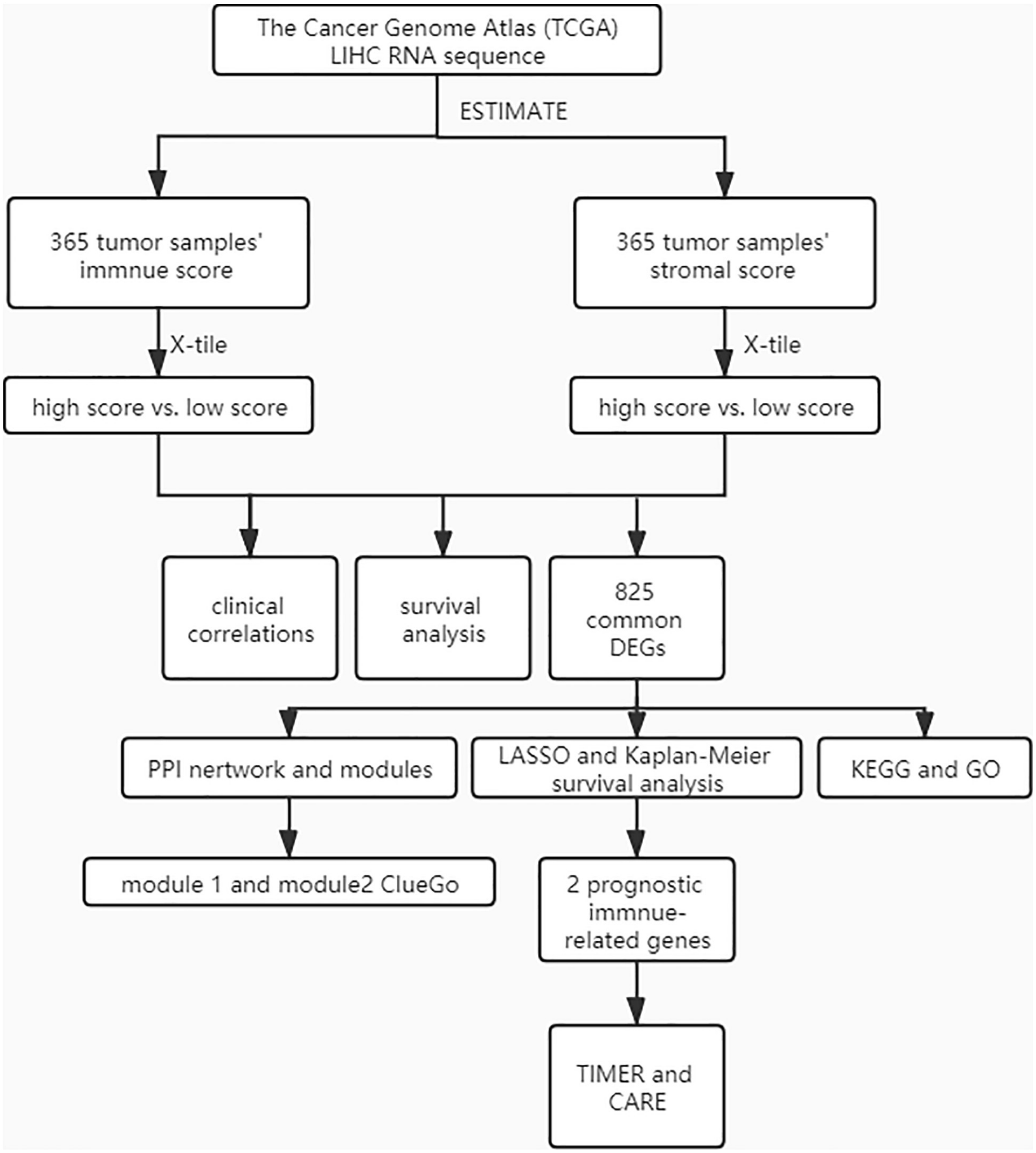

The Cancer Genome Atlas (TCGA) LIHC RNA sequence
ESTIMATE
365 tumor samples' immnue score
365 tumor samples' stromal score
X-tile
X-tile
high score vs. low score
high score vs. low score
clinical correlations
survival analysis
825 common DEGs
PPI nertwork and modules
LASSO and Kaplan-Meier survival analysis
KEGG and GO
module 1 and module2 ClueGo
2 prognostic immnue-related genes
TIMER and CARE

FIGURE 1 | The general pipeline of the data analysis protocol.

Comparison of Gene Expression Profile With Immune Scores and Stromal Scores in HCC

To identify the immune-related and stromal-related genes, differential analysis by using NetworkAnalyst was performed (**Supplementary Table 3**). The expression profiles of stromal and immune score–related DEGs are visualized, respectively, on the heatmaps (**Figures 4A,B**).

There were 797 shared DEGs overexpressed in both the stromal score and immune score groups (**Figure 4C**), and a total of 28 common DEGs were found to be underexpressed in both the stromal score and immune score groups (**Figure 4D**). Eight hundred twenty-five intersection genes were selected for further analysis (overlap zone in **Figures 4C,D**).

Using the "Goseq" and "clusterProfiler" R packages, 1,371 GO terms and 73 KEGG terms were indicated (**Supplementary Table 4**).

The results showed the top 10 biological processes GO terms, cellular component GO terms, and molecular function GO terms (**Figure 4E**). The correlation between the intersection genes and the top five biological processes is shown in **Supplementary Figure 2A**. The top 20 KEGG analysis showed that the intersection genes were associated with immune responses (**Figure 4F**).

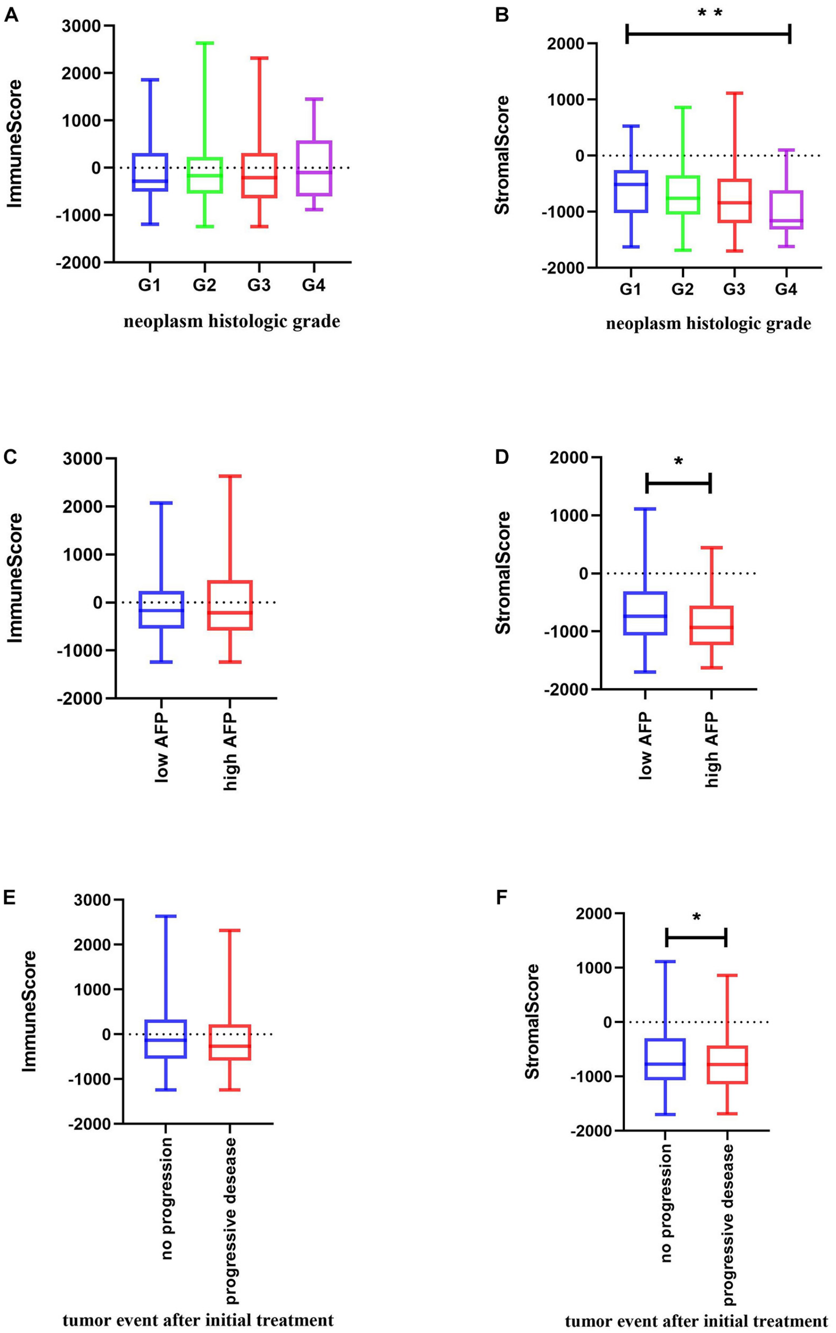


FIGURE 2 | Relationship between immune and stromal scores and HCC clinical and pathological data. **(A,B)** Distribution of immune and stromal scores of HCC grades. **(C,D)** Distribution of immune and stromal scores of AFP value of HCC. AFP is divided into high and low groups at the limit of 400 ng/mL. **(E,F)** Distribution of immune and stromal scores of new tumor event after initial treatment of HCC. Unpaired *t*-test was used to compare two groups of continuously distributed variables. Jonckheere–Terpstra test was used to compare three or more groups of continuously distributed variables. $^{*}P <$ 0.05 and $^{**}P < 0.01$.

Protein–Protein Interactions Among Intersection Genes

To better understand the interplay among the identified DEGs, we obtained PPI networks using the STRING tool. Using the MCODE software, we found modules in the network; the network was made up of eight modules, which included 408 nodes and 2,702 edges. We selected the top two significant modules for further analysis (**Figure 5A** and **Supplementary Figure 2B**).

GO analyses of module 1 (**Figure 5A**) by ClueGo are shown in **Figure 5B**. Likewise, GO analyses of module 2 (**Supplementary Figure 2B**) by ClueGo are shown in

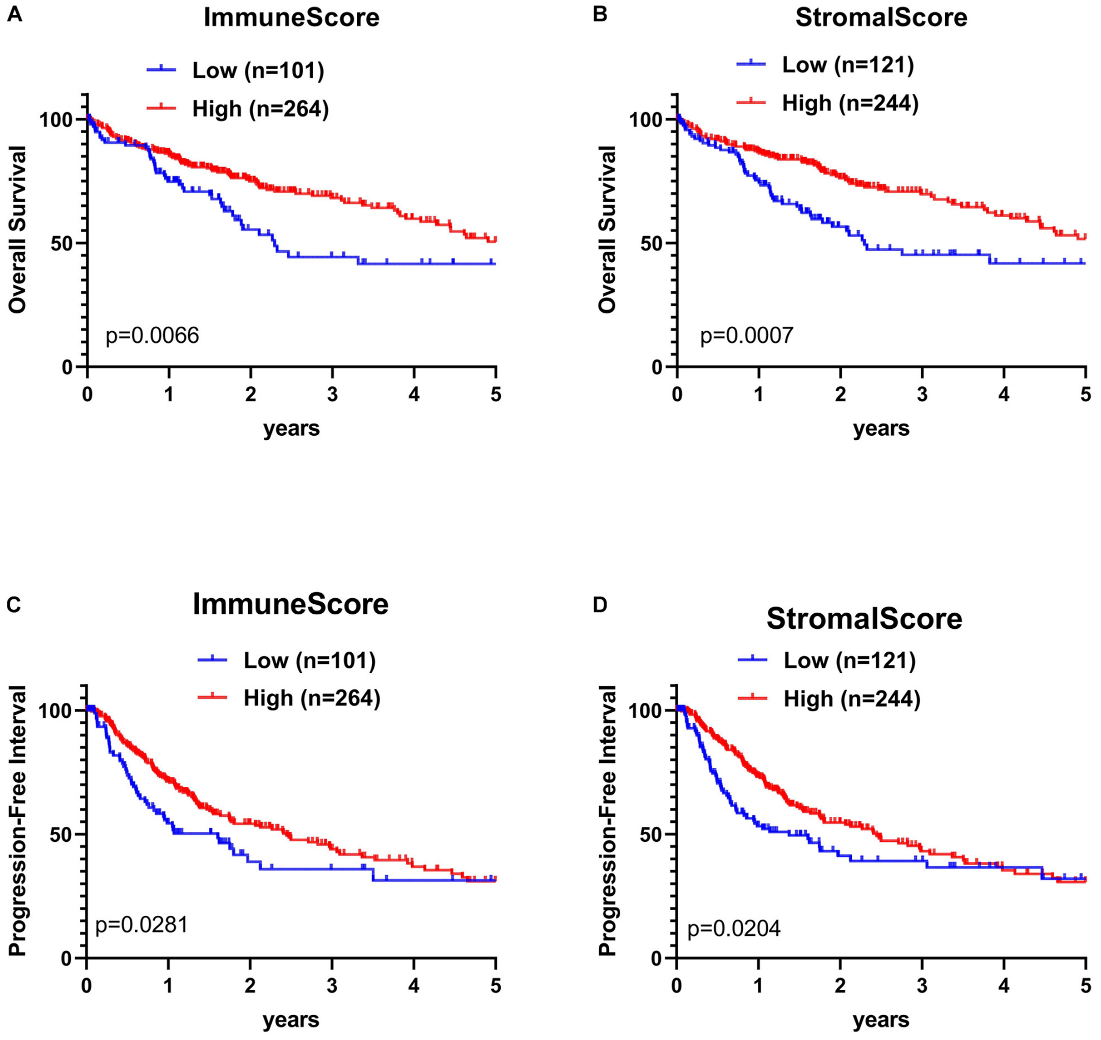


FIGURE 3 | Kaplan–Meier (KM) survival curve of HCC patients based on their immune-stromal scores. Patients were classified into high immune-stromal scores groups and low immune/stromal scores groups by using X-tile. **(A)** The KM curve of overall survival (OS) time of high and low immune score group. **(B)** The KM curve of OS time of high and low stromal score group. **(C)** The KM curve of PFI time according to immune scores. **(D)** The KM curve of progression-free interval (PFI) time according to stromal scores. The survival outcomes of the two groups were compared by log-rank tests. $P < 0.05$ was statistically significant.

Supplementary Figure 2C. The results demonstrated that module 1 was mainly enriched in regulation of dendritic cell apoptotic process, regulation of dendritic cell dendrite assembly, and positive regulation of T cell migration. Module 2 was mainly enriched in the regulation of phospholipase C activity, cellular response to interferon-γ (IFN-γ) and IFN-γ–mediated signaling pathway. Obviously, the top two modules were enriched for functional terms related to immune response processes, especially T cell responses.

Identification of Prognostic DEGs in HCC

To enrich for genes with the greatest prognostic values, we performed LASSO algorithm, and seven genes were identified (**Supplementary Figure 3A**). We also analyzed the association between the seven genes and OS using the Kaplan–Meier survival analysis. We found that the high levels of *GDF10* (P = 0.0484) and *MMP9* (P = 0.0143) negatively correlated with OS (**Figure 6A**).

Immune Cell Infiltration Analysis

To determine whether there is a correlation between tumor infiltration with immune cells and immune-related gene expression, the tumor infiltration with multiple immune cells was analyzed by TIMER 2.0 and other methods (**Supplementary Table 5**). **Figure 6B** shows the strong correlation between six types of immune cell infiltration and the expression of *MMP9*. The expression of *MMP9* positively correlated with the infiltrating levels of B cells (partial correlation = 0.529, P = 3.05e-26), CD8$^+$ T cells (partial correlation = 0.421, P = 4.13e-16), CD4$^+$ T cells (partial correlation = 0.356, P = 9.68e-12), macrophages (partial correlation = 0.473, P = 2.12e-20), neutrophils

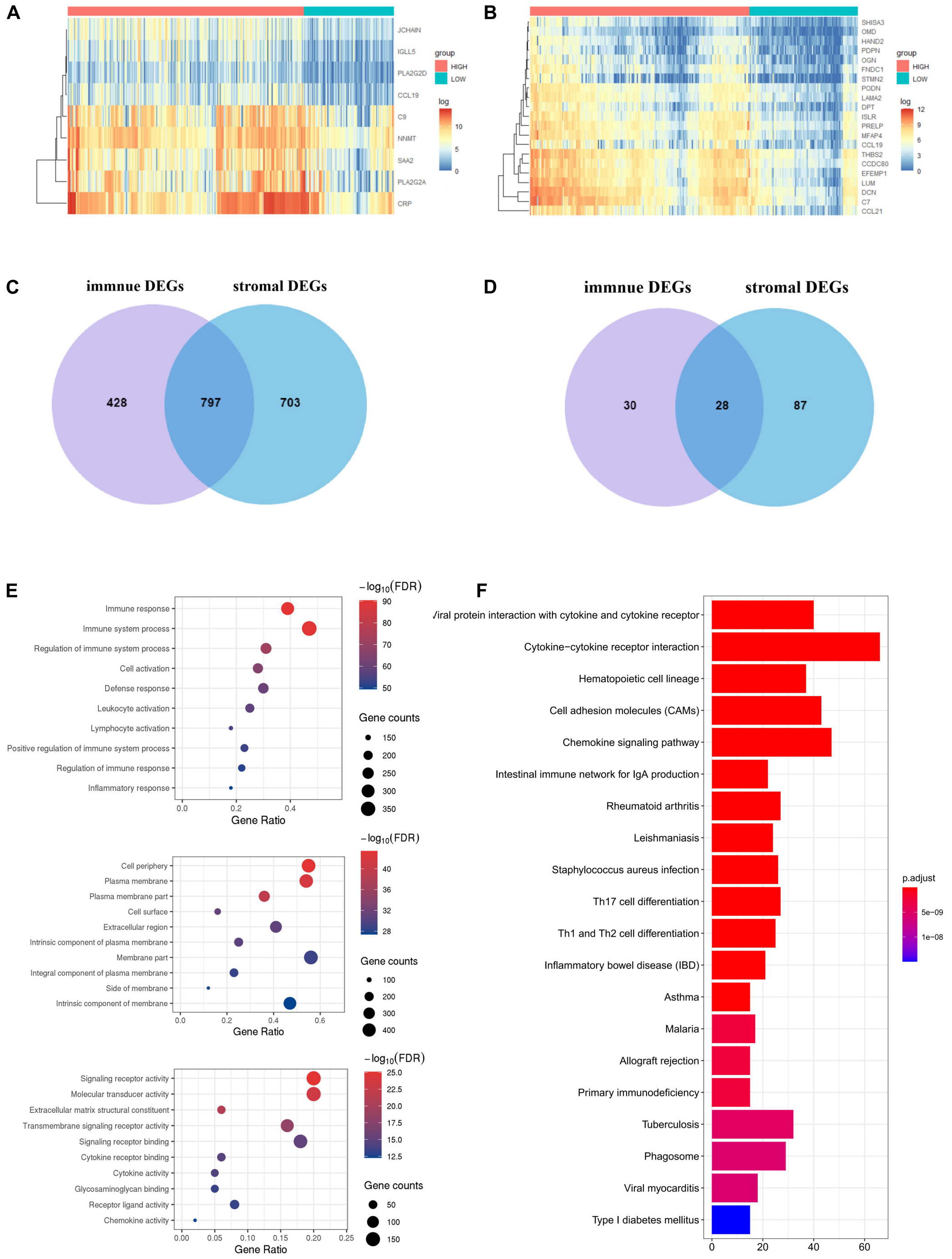


FIGURE 4 | Expression profiles and biological functions of stromal and immune score–related DEGs. **(A,B)** Heatmaps showing expression profiles for selected stromal score (right) and immune score (left)–related DEGs (Log2 fold change ≥ 3 and adjusted $P <$ 0.05) with unsupervised hierarchical clustering analyses, using the complete linkage method to measure distances between clusters. **(C)** Shows the commonly upregulated DEGs, and **(D)** shows the commonly downregulated DEGs. **(E)** The top 10 of biological processes GO terms (top), cellular component GO terms (middle), and molecular function GO terms (bottom); **(F)** KEGG (Kyoto Encyclopedia of Genes and Genomes) analysis of microenvironment-related DEGs.

(partial correlation = 0.34, P = 8.96e-11), and dendritic cells (partial correlation = 0.584, P = 1.72e-32). GDF10 expression was weakly associated with different immune cell infiltrates (**Figure 6B**).

We analyzed the correlation between *MMP9* and immune checkpoints in liver cancer. MMP9 was found to be correlated with the expression of a series of immune checkpoints. Particularly, MMP9 was significantly correlated with

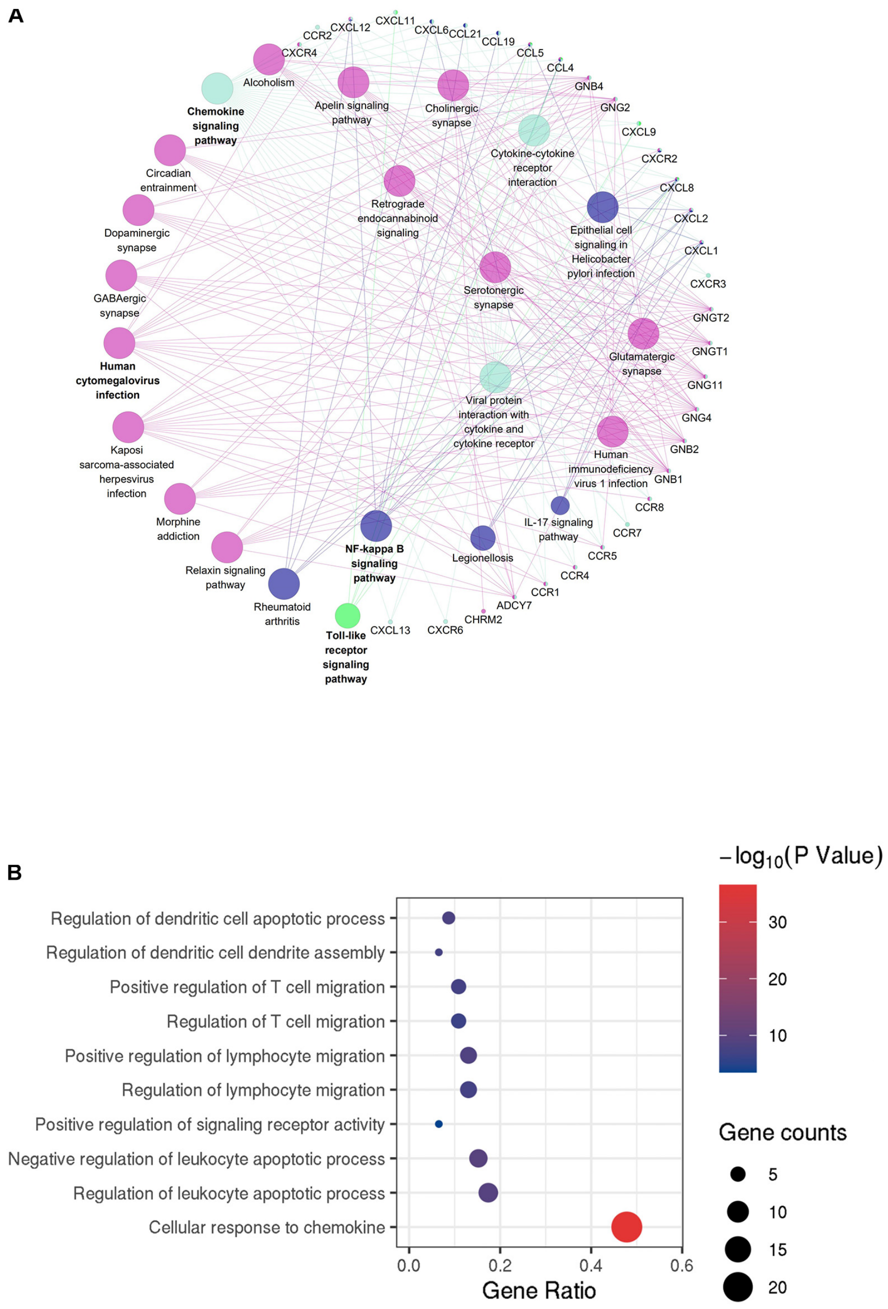


FIGURE 5 | Protein–protein interaction (PPI) network of microenvironment-related genes. **(A)** Module 1 is the top module in the PPI network. **(B)** GO analyses of module 1 (top 10 of biological processes GO terms). The color and thickness of edges reflect the combined score.

PDCD1 (ρ = 0.576), PDCD1LG2 (ρ = 0.372), and CTLA4 (ρ = 0.672) (**Figure 6C**).

Besides, identifying reliable drug response biomarkers is a significant challenge in cancer research. We present CARE, a computational method that enables large-scale inference of response biomarkers and drug combinations for targeted therapies using compound screen data. High expression of *MMP9* has been associated with better response to immunotherapies on CTRP dataset (**Figure 6D**).

DISCUSSION

Prognosis prediction for liver cancer patients remains challenging for clinicians and investigators. Through a specific view of the microenvironment, this study provides a stromal-immune score–based gene signature to help answer this important clinical question.

Using the ESTIMATE algorithm, we revealed the correlation between the immune-stromal scores and the clinical HCC

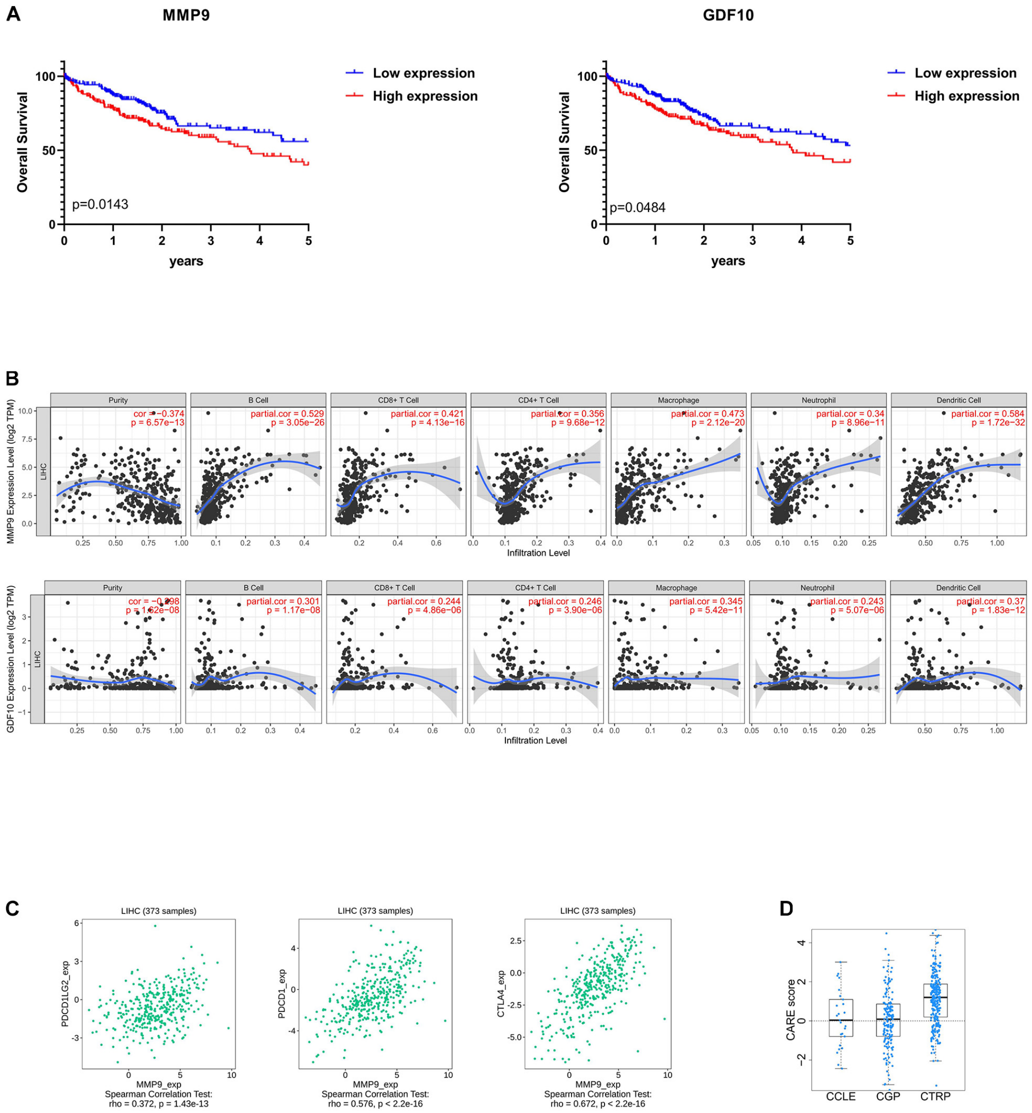


FIGURE 6 | Selection of microenvironment-related prognostic genes and the analysis of immune cell infiltration and immunoinhibitor. **(A)** Kaplan–Meier (KM) survival curve of *GDF10* and *MMP9*. Patients were divided into two groups based on the median of gene expression. The survival outcomes of the two groups were compared by log-rank tests. $P < 0.05$ was statistically significant. **(B)** Correlation of microenvironment-related prognostic genes' expression with immune infiltration level. **(C)** Relations between three kinds of immunoinhibitors and expression of *MMP9*. $P < 0.05$ was statistically significant, and partial correlation ≥0.3 indicates strong correlation. **(D)** The CARE score of *MMP9* on CCLE, CGP, CTRP dataset. A positive score indicates a higher expression value to be associated with drug response.

characteristics obtained from TCGA-CDR. The stromal and immune scores for tumor tissue were found to be positively associated with the clinicopathologic characteristics of the tumor and the patient's prognosis. By analyzing the correlation between the immune scores and tumor recurrence, our data show that high-immune-score patients have a longer PFI and OS rates, indicating that the TME composition affects the clinical outcomes of HCC patients, which is consistent with previous studies (Haider et al., 2020).

Next, we analyzed 825 DEGs yielded from a comparison of high- versus low-immune-score (or stromal scores) groups and found that many of them were involved in the TME, specifically regulate T cell functions (**Figure 4E**). This is consistent with previous reports that the functions of immune cells and extracellular matrix molecules are interrelated in building TME in HCC (Lu et al., 2019; Yin et al., 2019). Moreover, we were able to construct two PPI modules (**Figure 5** and **Supplementary Figures 2B,C**), the major of which were related to IFN-γ. We infer that these TME-associated genes might affect the development of HCC by affecting the T cell functions.

Finally, by using the LASSO algorithm (**Supplementary Figure 3A**), we identified seven TME-related genes. Of the seven genes identified, high levels of *GDF10* and *MMP9* showed a negative correlation to OS, which has been reported to be involved in carcinogenesis and the development of various cancers (Chang et al., 2017; Reggiani et al., 2017; Tekin et al., 2020a). We further correlated the degree of infiltration of six immune cell types with the expression of GDF10 and MMP9 by using TIMER algorithm. The expression of *MMP9* was positively associated with the abundance of six immune in tumor tissues. It is worth reminding that our results did not contradict previous findings that high infiltration of $CD8^+$ T cells indicated beneficial prognosis, but extended and enriched this conclusion. In the recent literature, tumor with higher $CD8^+$ T cell infiltration, but T cell dysfunction and increased immune escape result in a poor prognosis (Hossain et al., 2020; Saka et al., 2020).

Prior studies have largely focused on *MMP*s' ability to promote the invasion and metastasis of cancer cells (Nart et al., 2010; Chen et al., 2012), while evidence is mounting that *MMP*s are highly associated with the microenvironment of tumors and immune cells (Kessenbrock et al., 2010; Li et al., 2016). For example, *MMP9*-cleaved osteopontin fragments contribute to tumor immune escape by inducing the expansion of myeloid-derived suppressor cells (Shao et al., 2017). Macrophages secrete *MMP9* to induce mesenchymal transition, which supports the tumor-promoting role of macrophage influx (Tekin et al., 2020b). Besides, *MMP9* is associated with neutrophil migration (Koymans et al., 2016). Our study confirms the above conclusions and has found that *MMP9* might associate with T cell dysfunction, despite high $CD8^+$ cytotoxic T lymphocyte infiltration.

In addition, we also observed that high expression of *MMP9* indicated higher levels of immune inhibitors (immune checkpoints), better response to immunotherapies, and poor survival in partial HCC patients, which was in line with our above analysis that some HCC patients with high $CD8^+$ T cell infiltration but with dysfunction were immunosuppressed. And previously, inhibition of *MMP9* could modulate immunosuppression in tumor (Melani et al., 2007). We also compared the prediction effect between the other factors, such as AFP (**Supplementary Figure 1A**) and programmed cell death protein 1 (PDCD1) (**Supplementary Figure 3B**), whereas AFP is not a good predictor of the abundance of immune invasion in HCC tissues, and PDCD1 is weakly correlated with the prognosis of HCC. Hence, *MMP9* may be an effective biomarker to evaluate the immune status of patients and predict the effectiveness of immunotherapy before treatment. However, this conclusion will need to be confirmed by clinical trials in the future.

In summary, from comprehensively analyzing the correlation between microenvironmental and genetic factors of TCGA database applied by ESTIMATE algorithm-based immune and stromal scores, we identified *MMP9* as a potential TME-related biomarker of prognostic and immunotherapy response. However, because of the lack of large sequenced HCC cohort and prospective clinical trials that have received immunotherapy, the effect of *MMP9* expression on the efficiency of immunotherapy in HCC patients remains concerned.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SL, GY, and LL: conception and design of study. SL: acquisition of data and drafting the manuscript. SL, XZ, LZ, EH, and YS: analysis and interpretation of data. XZ, GY, and LL: revising the manuscript critically for important intellectual content. SL, GY, LL, XZ, LZ, EH, and YS: approval of the version of the manuscript to be published. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 81773008 and 81672756), Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2015), Guangdong Natural Science Foundation (Grant No. 2017A030311023), Guangzhou Science and Technology Program projects (Grant No. 201804010044), Local Innovative and Research Teams Project of Guangdong Pearl River Talents Program (Grant No. 2017BT01S131), and Postdoctoral Research Foundation of China (Grant No. 2019M660208).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.625236/full#supplementary-material

REFERENCES

Affo, S., Yu, L. X., and Schwabe, R. F. (2017). The role of cancer-associated fibroblasts and fibrosis in liver cancer. *Annu. Rev. Pathol.* 12, 153–186. doi: 10.1146/annurev-pathol-052016-100322

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Bardou, P., Mariette, J., Escudie, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* 15:293. doi: 10.1186/1471-2105-15-293

Barry, A. E., Baldeosingh, R., Lamm, R., Patel, K., Zhang, K., Dominguez, D. A., et al. (2020). Hepatic stellate cells and Hepatocarcinogenesis. *Front. Cell. Dev. Biol.* 8:709. doi: 10.3389/fcell.2020.00709

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Boyero, L., Sanchez-Gastaldo, A., Alonso, M., Noguera-Ucles, J. F., Molina-Pinelo, S., and Bernabe-Caro, R. (2020). Primary and acquired resistance to immunotherapy in lung cancer: unveiling the mechanisms underlying of immune checkpoint blockade therapy. *Cancers (Basel)* 12:3729. doi: 10.3390/cancers12123729

Bruix, J., Reig, M., and Sherman, M. (2016). Evidence-based diagnosis, staging, and treatment of patients with hepatocellular carcinoma. *Gastroenterology* 150, 835–853. doi: 10.1053/j.gastro.2015.12.041

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, Y. C., Chan, Y. C., Chang, W. M., Lin, Y. F., Yang, C. J., Su, C. Y., et al. (2017). Feedback regulation of ALDOA activates the HIF-1alpha/MMP9 axis to promote lung cancer progression. *Cancer Lett.* 403, 28–36. doi: 10.1016/j.canlet.2017.06.001

Chen, R., Cui, J., Xu, C., Xue, T., Guo, K., Gao, D., et al. (2012). The significance of MMP-9 over MMP-2 in HCC invasiveness and recurrence of hepatocellular carcinoma after curative resection. *Ann. Surg. Oncol.* 19 Suppl 3, S375–S384. doi: 10.1245/s10434-011-1836-7

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

Galle, P. R., Foerster, F., Kudo, M., Chan, S. L., Llovet, J. M., Qin, S., et al. (2019). Biology and significance of alpha-fetoprotein in hepatocellular carcinoma. *Liver Int.* 39, 2214–2229. doi: 10.1111/liv.14223

Haider, T., Sandha, K. K., Soni, V., and Gupta, P. N. (2020). Recent advances in tumor microenvironment associated therapeutic strategies and evaluation models. *Mater. Sci. Eng. C. Mater. Biol. Appl.* 116:111229. doi: 10.1016/j.msec.2020.111229

Hossain, M. A., Liu, G., Dai, B., Si, Y., Yang, Q., Wazir, J., et al. (2020). Reinvigorating exhausted CD8(+) cytotoxic T lymphocytes in the tumor microenvironment and current strategies in cancer immunotherapy. *Med. Res. Rev.* 41, 156–201. doi: 10.1002/med.21727

Jiang, P., Lee, W., Li, X., Johnson, C., Liu, J. S., Brown, M., et al. (2018). Genome-scale signatures of gene interaction from compound screens predict clinical efficacy of targeted cancer therapies. *Cell Syst.* 6, 343–354 e5. doi: 10.1016/j.cels.2018.01.009

Jin, M. Z., and Jin, W. L. (2020). The updated landscape of tumor microenvironment and drug repurposing. *Signal Transduct. Target Ther.* 5:166. doi: 10.1038/s41392-020-00280-x

Kanwal, F., and Singal, A. G. (2019). Surveillance for hepatocellular carcinoma: current best practice and future direction. *Gastroenterology* 157, 54–64. doi: 10.1053/j.gastro.2019.02.049

Kassambara, A., Kosinski, M., Biecek, P., and Fabian, S. (2019). *survminer: Drawing Survival Curves Using 'ggplot2'.* R package version 0.4.6. Available online at: https://CRAN.R-project.org/package=survminer (accessed July 25, 2020).

Kessenbrock, K., Plaks, V., and Werb, Z. (2010). Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell* 141, 52–67. doi: 10.1016/j.cell.2010.03.015

Korenberg, M. J. (2006). Applications of nonlinear system identification in molecular biology. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2006, 256–259. doi: 10.1109/IEMBS.2006.259455

Koymans, K. J., Bisschop, A., Vughs, M. M., van Kessel, K. P., de Haas, C. J., and van Strijp, J. A. (2016). Staphylococcal superantigen-like protein 1 and 5 (SSL1 & SSL5) limit neutrophil chemotaxis and migration through MMP-inhibition. *Int. J. Mol. Sci.* 17:1072. doi: 10.3390/ijms17071072

Li, M., Xing, S., Zhang, H., Shang, S., Li, X., Ren, B., et al. (2016). A matrix metalloproteinase inhibitor enhances anti-cytotoxic T lymphocyte antigen-4 antibody immunotherapy in breast cancer by reprogramming the tumor microenvironment. *Oncol. Rep.* 35, 1329–1339. doi: 10.3892/or.2016.4547

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 77, e108–e110. doi: 10.1158/0008-5472.CAN-17-0307

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 48, W509–W514. doi: 10.1093/nar/gkaa407

Lu, C., Rong, D., Zhang, B., Zheng, W., Wang, X., Chen, Z., et al. (2019). Current perspectives on the immunosuppressive tumor microenvironment in hepatocellular carcinoma: challenges and opportunities. *Mol. Cancer* 18:130. doi: 10.1186/s12943-019-1047-6

Melani, C., Sangaletti, S., Barazzetta, F. M., Werb, Z., and Colombo, M. P. (2007). Amino-biphosphonate-mediated MMP-9 inhibition breaks the tumor-bone marrow axis responsible for myeloid-derived suppressor cell expansion and macrophage infiltration in tumor stroma. *Cancer Res.* 67, 11438–11446. doi: 10.1158/0008-5472.CAN-07-1882

Nart, D., Yaman, B., Yilmaz, F., Zeytunlu, M., Karasu, Z., and Kilic, M. (2010). Expression of matrix metalloproteinase-9 in predicting prognosis of hepatocellular carcinoma after liver transplantation. *Liver Transpl.* 16, 621–630. doi: 10.1002/lt.22028

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

Reggiani, F., Labanca, V., Mancuso, P., Rabascio, C., Talarico, G., Orecchioni, S., et al. (2017). Adipose progenitor cell secretion of GM-CSF and MMP9 promotes a stromal and immunological microenvironment that supports breast cancer progression. *Cancer Res.* 77, 5169–5182. doi: 10.1158/0008-5472.CAN-17-0914

Ru, B., Wong, C. N., Tong, Y., Zhong, J. Y., Zhong, S. S. W., Wu, W. C., et al. (2019). TISIDB: an integrated repository portal for tumor-immune system interactions. *Bioinformatics* 35, 4200–4202. doi: 10.1093/bioinformatics/btz210

Saka, D., Gokalp, M., Piyade, B., Cevik, N. C., Arik Sever, E., Unutmaz, D., et al. (2020). Mechanisms of T-Cell exhaustion in pancreatic cancer. *Cancers (Basel)* 12:2274. doi: 10.3390/cancers12082274

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

Shao, L., Zhang, B., Wang, L., Wu, L., Kan, Q., and Fan, K. (2017). MMP-9-cleaved osteopontin isoform mediates tumor immune escape by inducing expansion of myeloid-derived suppressor cells. *Biochem. Biophys. Res. Commun.* 493, 1478–1484. doi: 10.1016/j.bbrc.2017.10.009

Son, J., Cho, J. W., Park, H. J., Moon, J., Park, S., Lee, H., et al. (2020). Tumor-infiltrating regulatory T cell accumulation in the tumor microenvironment is mediated by IL33/ST2 signaling. *Cancer Immunol. Res.* 8, 1393–1406. doi: 10.1158/2326-6066.CIR-19-0828

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

Tekin, C., Aberson, H. L., Waasdorp, C., Hooijer, G. K. J., de Boer, O. J., Dijk, F., et al. (2020a). Macrophage-secreted MMP9 induces mesenchymal transition in pancreatic cancer cells via PAR1 activation. *Cell. Oncol. (Dordr.)* 43, 1161–1174. doi: 10.1007/s13402-020-00549-x

Tekin, C., Aberson, H. L., Waasdorp, C., Hooijer, G. K. J., de Boer, O. J., Dijk, F., et al. (2020b). Macrophage-secreted MMP9 induces mesenchymal transition in pancreatic cancer cells via PAR1 activation. *Cell. Oncol. (Dordr)* 43, 1161–1174. doi: 10.1007/s13402-020-00549-x

Therneau, T. M. (2020). *A Package for Survival Analysis in R.* R Package Version 3.1-11. Available online at: https://CRAN.R-project.org/package=survival (accessed September 9, 2020).

Wang, J. B., Li, P., Liu, X. L., Zheng, Q. L., Ma, Y. B., Zhao, Y. J., et al. (2020). An immune checkpoint score system for prognostic evaluation and adjuvant chemotherapy selection in gastric cancer. *Nat. Commun.* 11:6352. doi: 10.1038/s41467-020-20260-7

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Xia, J., Fjell, C. D., Mayer, M. L., Pena, O. M., Wishart, D. S., and Hancock, R. E. (2013a). INMEX–a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res.* 41, W63–W70. doi: 10.1093/nar/gkt338

Xia, J., Lyle, N. H., Mayer, M. L., Pena, O. M., and Hancock, R. E. (2013b). INVEX–a web-based tool for integrative visualization of expression data. *Bioinformatics* 29, 3232–3234. doi: 10.1093/bioinformatics/btt562

Xu, D., Wang, Y., Liu, X., Zhou, K., Wu, J., Chen, J., et al. (2020). Development and clinical validation of a novel 9-gene prognostic model based on multi-omics in pancreatic adenocarcinoma. *Pharmacol. Res.* 164:105370. doi: 10.1016/j.phrs.2020.105370

Yang, J. D., Hainaut, P., Gores, G. J., Amadou, A., Plymoth, A., and Roberts, L. R. (2019). A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat. Rev. Gastroenterol. Hepatol.* 16, 589–604. doi: 10.1038/s41575-019-0186-y

Yang, X., Chen, L., Liang, Y., Si, R., Jiang, Z., Ma, B., et al. (2018). Knockdown of alpha-fetoprotein expression inhibits HepG2 cell growth and induces apoptosis. *J. Cancer Res. Ther.* 14(Supplement), S634–S643. doi: 10.4103/0973-1482.180681

Yang, Y., Yang, Y., Yang, J., Zhao, X., and Wei, X. (2020). Tumor microenvironment in ovarian cancer: function and therapeutic strategy. *Front. Cell. Dev. Biol.* 8:758. doi: 10.3389/fcell.2020.00758

Yin, Z., Dong, C., Jiang, K., Xu, Z., Li, R., Guo, K., et al. (2019). Heterogeneity of cancer-associated fibroblasts and roles in the progression, prognosis, and therapy of hepatocellular carcinoma. *J. Hematol. Oncol.* 12:101. doi: 10.1186/s13045-019-0782-x

Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612. doi: 10.1038/ncomms3612

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14. doi: 10.1186/gb-2010-11-2-r14

Yu, G. (2019). *enrichplot: Visualization of Functional Enrichment Result.* R package version 1.6.1. Available online at: https://github.com/GuangchuangYu/enrichplot (accessed December 30, 2020).

Yu, G. (2020a). *aplot: Decorate a 'ggplot' with Associated Information.* R package Version 0.0.6. Available online at: https://CRAN.R-project.org/package=aplot (accessed September 3, 2020).

Yu, G. (2020b). Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* 69:e96. doi: 10.1002/cpbi.96

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., Yang, M., Ng, D. M., Haleem, M., Yi, T., Hu, S., et al. (2020). Multi-omics data analyses construct TME and identify the immune-related prognosis signatures in human LUAD. *Mol. Ther. Nucleic Acids* 21, 860–873. doi: 10.1016/j.omtn.2020.07.024

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., and Xia, J. (2019). NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.* 47, W234–W241. doi: 10.1093/nar/gkz240

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.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 22 February 2021
doi: 10.3389/fgene.2021.615655

Macrophage M2 Co-expression Factors Correlate With the Immune Microenvironment and Predict Outcome of Renal Clear Cell Carcinoma

Yutao Wang[1†], Kexin Yan[2†], Jiaxing Lin[1†], Jun Li[1] and Jianbin Bi[1*]*

[1] Department of Urology, The First Hospital of China Medical University, China Medical University, Shenyang, China,
[2] Department of Dermatology, The First Hospital of China Medical University, China Medical University, Shenyang, China

Purpose: In the tumor microenvironment, the functional differences among various tumor-associated macrophages (TAM) are not completely clear. Tumor-associated macrophages are thought to promote the progression of cancer. This article focuses on exploring M2 macrophage-related factors and behaviors of renal clear cell carcinoma.

Method: We obtained renal clear cell carcinoma data from TCGA-KIRC-FPKM, GSE8050, GSE12606, GSE14762, and GSE3689. We used the "Cibersort" algorithm to calculate type M2 macrophage proportions among 22 types of immune cells. M2 macrophage-related co-expression module genes were selected using weighted gene co-expression network analysis (WGCNA). A renal clear cell carcinoma prognosis risk score was built based on M2 macrophage-related factors. The ROC curve and Kaplan–Meier analysis were performed to evacuate the risk score in various subgroups. The Pearson test was used to calculate correlations among M2 macrophage-related genes, clinical phenotype, immune phenotype, and tumor mutation burden (TMB). We measured differences in co-expression of genes at the protein level in clear renal cell carcinoma tissues.

Results: There were six M2 macrophage co-expressed genes (F13A1, FUCA1, SDCBP, VSIG4, HLA-E, TAP2) related to infiltration of M2 macrophages; these were enriched in neutrophil activation and involved in immune responses, antigen processing, and presentation of exogenous peptide antigen via MHC class I. M2-related factor frequencies were robust biomarkers for predicting the renal clear cell carcinoma patient clinical phenotype and immune microenvironment. The Cox regression model, built based on M2 macrophage-related factors, showed a close prognostic correlation (AUC = 0.78). The M2 macrophage-related prognosis model also performed well in various subgroups. Using western blotting, we found that VSIG4 protein expression levels were higher in clear renal cell carcinoma tissues than in normal tissues.

OPEN ACCESS

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Jingting Yu,
Salk Institute for Biological Studies,
United States
Yi Zhang,
Dana-Farber Cancer Institute,
United States
Kuixi Zhu,
University of Arizona, United States

***Correspondence:**
Jianbin Bi
bijianbin@hotmail.com
Jun Li
cmu_lijun@163.com

[†]These authors have contributed equally to this work

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

Received: *09 October 2020*
Accepted: *07 January 2021*
Published: *22 February 2021*

Citation:
Wang Y, Yan K, Lin J, Li J and Bi J (2021) Macrophage M2 Co-expression Factors Correlate With the Immune Microenvironment and Predict Outcome of Renal Clear Cell Carcinoma. Front. Genet. 12:615655. doi: 10.3389/fgene.2021.615655

Conclusion: These co-expressed genes were most related to the M2 macrophage phenotype. They correlated with the immune microenvironment and predicted outcomes of renal clear cell carcinoma. These co-expressed genes and the biological processes associated with them might provide the basis for new strategies to intervene via chemotaxis of M2 macrophages.

Keywords: M2 macrophage, weighted gene co-expression network analysis, tumor-associated macrophage, immune phenotype, risk model

INTRODUCTION

Renal clear cell carcinoma (RCC) accounts for 80–90% of all renal cell carcinomas; clear cell carcinoma is not sensitive to chemotherapy and radiotherapy (Hsieh et al., 2017). For this reason, radical surgery has become the main treatment method. In clinical practice, although radical nephrectomy can benefit mostly patients, 30% of patients experience distant metastases after surgery (Motzer et al., 2013). Although we have adopted various treatment strategies for these patients with poor status, the long-term outcomes are not ideal (Linehan and Ricketts, 2019). With the development of immunotherapy in recent years, there have been studies showing that immunotherapy can benefit patients with renal clear cell cancer (Chowdhury and Drake, 2020; Díaz-Montero et al., 2020; Wang C. et al., 2020).

Renal clear cell carcinoma is characterized by many new tumor antigen peptides and high mutation burden; it is relatively sensitive to immunotherapies such as targeting PD1 and PD-L1 (Wang C. et al., 2020). Immune regulation plays a crucial role in the renal clear cell carcinoma microenvironment. This process includes immune checkpoints [mainly programmed cell death 1 (PD-1) and programmed cell death 1 ligand 1 (PD-L1)], as well as regulatory T cells, the original source of suppressor cell tumor-associated macrophages, and type 2 innate and adaptive lymphocytes (Xu W. et al., 2020). Macrophages in the primary or secondary tumor tissues are called tumor-associated macrophages (TAMs); these are the largest number of macrophages in the tumor stroma (Herberman et al., 1979). In recent years, clinical and experimental evidence has shown that macrophages promote the progression and metastasis of solid tumors, and this is somewhat different from our previous understanding (Pollard, 2004; Karnevi et al., 2014). Tumor-associated macrophages are divided into two types, M1 and M2 (Herberman et al., 1979; DeNardo and Ruffell, 2019). The biological effects of the two types are exact opposites. As tumors progress, increasing numbers of M2 macrophages appear, resulting in a weaker antigen presentation effect. For this reason, targeting macrophages has become a new therapeutic strategy (DeNardo and Ruffell, 2019). M1 type macrophages, namely, classically activated macrophages, highly express IL-12 and IL-23 that enhance antitumor effects (Lawrence and Natoli, 2011). By contrast, M2 type macrophages, namely, alternatively activated macrophages, promote tumor formation and development (Cervantes-Villagrana et al., 2020). The mechanism of this polarization of macrophages is not clear. This article focuses on exploring the M2 macrophage-related genes in renal clear cell cancer, and constructing co-expression networks of M2 macrophages using the WGCNA method. The results of this paper revealed the underlying interaction mechanisms of M2 macrophage co-expressing factors and explained the role of M2 macrophages in the immune microenvironment from the perspective of bioinformatics.

METHODS

Macrophage M2, Tumor Purity, and Tumor Mutation Burden Evaluation

We downloaded The Cancer Genome Atlas TCGA—KIRC FPKM data (http://cancergenome.nih.gov/) containing 539 renal clear cell cancer tissue samples and 72 normal tissues. GSE8050 (Weinzierl et al., 2008), GSE12606 (Stickel et al., 2009), GSE14762 (Wang et al., 2009), and GSE36895 (Peña-Llopis et al., 2012) were also downloaded from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database. The Robust Multi-Array Average (RMA) algorithm of the "sva" (Leek et al., 2012) package was used to remove batch effects among the four GEO cohorts. The TCGA cohort was used to select M2-related genes. Four GEO cohorts were combined using "sva" packages and to verify the results. The Cell type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORT) is a deconvolution algorithm based on a gene expression profile that characterizes the cell composition of complex tissues, quantifies immune cells, and accurately estimates the immune components of tumor samples. It expands the potential of the genomic database, showing the pattern of Renal Clear Cell Carcinoma with comprehensive immune cells. We calculated macrophage M2 cell proportions based on the LM22 matrix using the CIBERSORT (Chen et al., 2018) algorithm, Cibersort was used as an obvious method to evaluate the significance of infiltration of immune cells in the samples. The assessment results of some samples were not statistically significant, and we used $P < 0.05$ to screen the samples. The Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data (ESTIMATE) is a method that infers the fraction of stromal and immune cells using gene expression signatures (Yoshihara et al., 2013). Using the ESTIMATE package, we calculated tumor purity in each renal

Abbreviations: TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; BLCA, Bladder urothelial carcinoma; ROC, Receiver operating characteristic; AUC, Area under the curve; HR, Hazard ratio; TME, Tumor microenvironment.

clear cell cancer sample. TMB (tumor mutation burden) per megabyte is calculated by dividing the total number of mutations by the size of the target coding region (Li et al., 2020; Yang et al., 2020).

Macrophage M2 Co-expression Network Conduction

Weighted gene co-expression network analysis (WGCNA) is a system biology approach that converts co-expression correlations

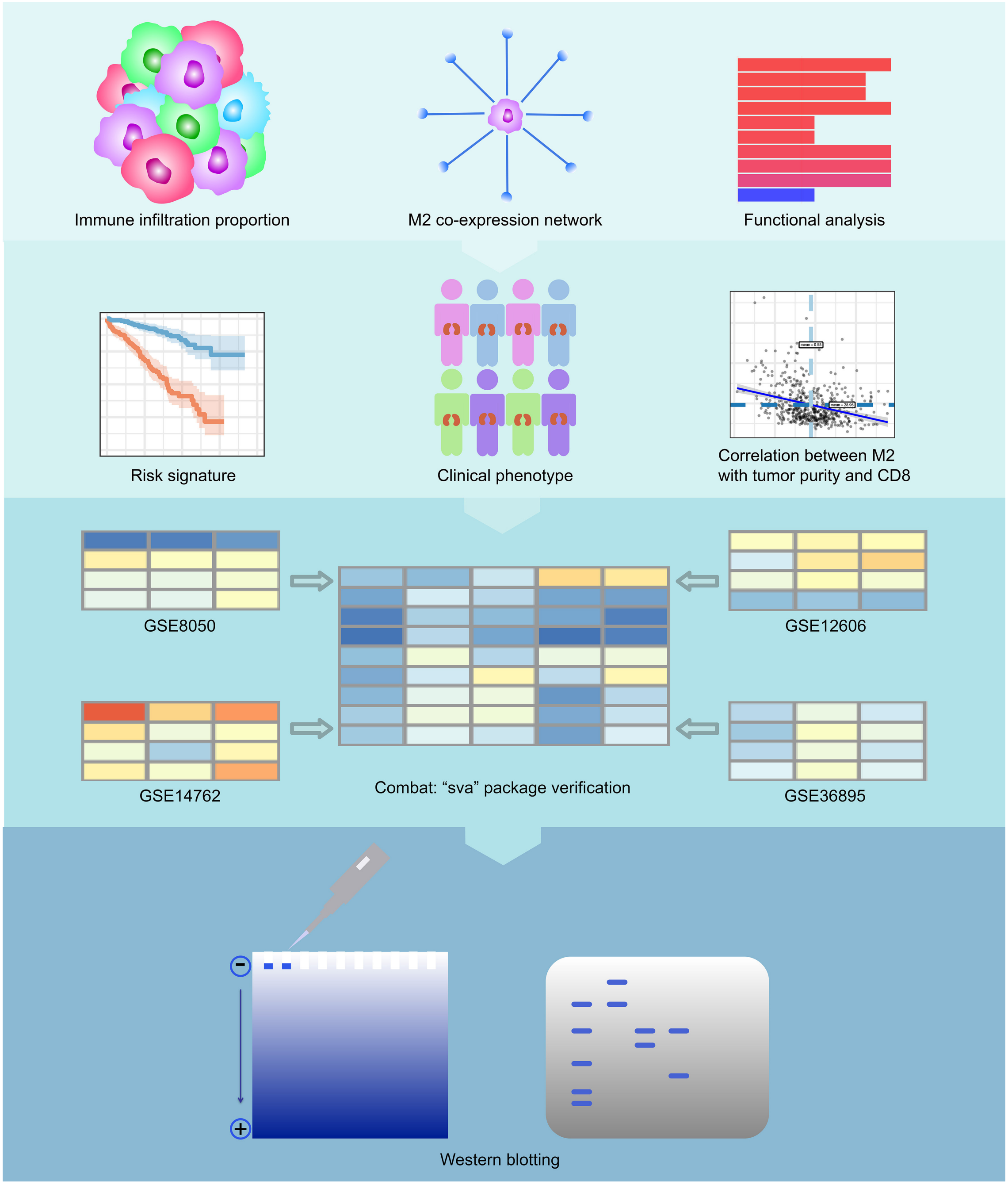


FIGURE 1 | Flowchart of the experimental design. We first calculated immune infiltration to determine the content of M2 macrophages in the immune microenvironment of RCC. Then, we constructed a co-expression network related to M2 macrophages of RCC and analyzed the enriched pathways in this network. We then calculated the survival analysis of these co-expressed genes. We constructed a COX regression prognostic model associated with the co-expression genes of M2 macrophages in RCC and performed a subgroup analysis of this model. We analyzed the relationship between key genes in the model and tumor purity and CD8^{+} T cells. Finally, we also verified the feasibility of the model with 4 GEO datasets and conducted western blotting experiments on VSIG4.

into connection weights or topology overlap values (Langfelder and Horvath, 2008). We used this method to determine proportions of co-expressed genes in the M2 macrophage. The expression patterns are similar for genes with the same biological process and biological function (Jiang et al., 2017). We built a scale-free topology network, set the soft threshold at 5, R square = 0.89, and set the number of genes in the minimum module at 30. The M2 macrophage cell proportion was considered for phenotype files in WGCNA. In this manner, a cluster of M2 macrophage cell proportion-related genes with similar function were identified in the same module. The factors with M2 macrophage correlation >0.4 in the most relevant modules were determined.

M2 Macrophage-Related Module Analysis

The genes were selected using |correlation coefficient| > 0.4. The Database for Annotation, Visualization and Integrated Discovery (DAVID, v6.8) is an open-source database that performs function enrichment (Huang et al., 2007). We used the Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/) (Kanehisa et al., 2017) and Gene Ontology (GO) (http://geneontology.org/) analysis (Ashburner et al., 2000) to identify the biological function in each co-expression module. In this way, we identified the biological processes associated with M2-type macrophage proportion.

M2 Macrophage-Related Genes Analysis

To verify the correlation between these factors and the clinical phenotype, we measured the overall survival from clear cell carcinoma as the prognostic indicator. Survival analysis was performed to evaluate the prognostic value of these co-expressed factors in M2 macrophages. Subsequently, a Cox regression hazard model was built based on the M2 macrophage-related genes. Next, we generated a model validation of clinical subgroups, which was based on age, gender, tumor metastasis, tumor stage, tumor purity, and degree of tumor mutation

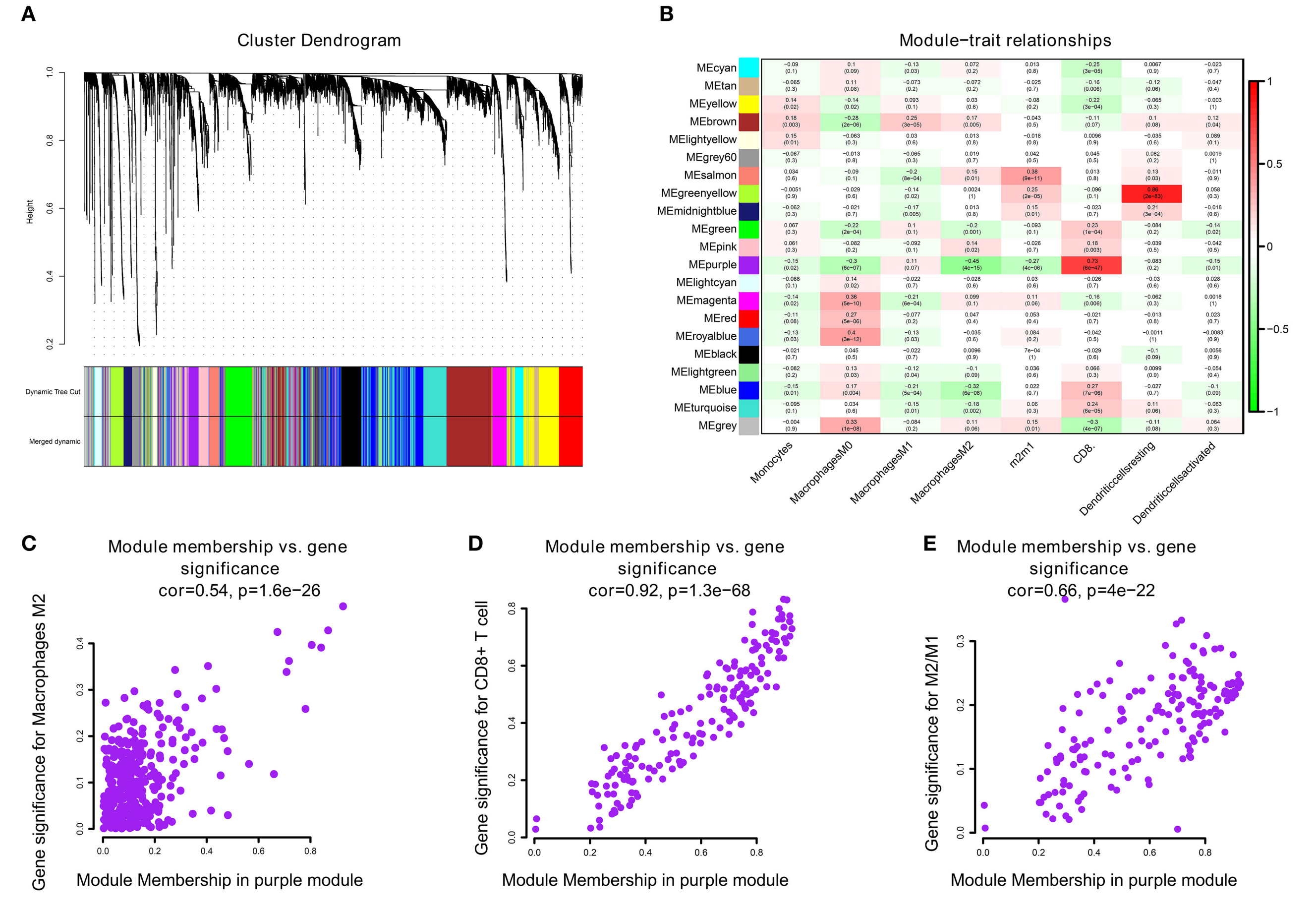


FIGURE 2 | (A) A hierarchical clustering tree was built using the dynamic hybrid cutting method, where each leaf on the tree represents a gene, and each branch represents a co-expression module; 21 co-expression models were generated. **(B)** The correlation coefficients between each phenotype and co-expression module of TCGA. The purple module had the strongest correlation with M2 macrophage cell proportions in the TCGA–KIRC cohort (Cor = −0.45; $P = 4e^{-15}$) and had the strongest correlation with $CD8^+$ T cell proportions in the TCGA–KIRC cohort (Cor = 0.73; $P = 6e^{-47}$). **(C)** The relationship between the purple module membership degree and the gene significance of M2 macrophages (cor = 0.54; P = 1.6e−26). **(D)** The relationship between the purple module membership degree and the gene significance of CD8+ T Cells (cor = 0.92; P = 1.3e−68). **(E)** The relationship between the purple module membership degree and the gene significance of M2/M1 ratio (cor = 0.66; P = 4e−22).

burden. In different subgroups, we evaluated the predictive abilities of M2 macrophage-related prognostic models. Finally, we calculated tumor purity in TCGA samples and explored the correlations between macrophage-related factors and tumor purity.

HPA

To verify the protein expression levels of candidate genes in melanoma and normal tissues, the human protein atlas (HPA, https://www.proteinatlas.org/) database was used to demonstrate differences in co-expressed genes at the protein level (Uhlén et al., 2015).

Western Blotting

Thirty clear renal cell carcinoma tissue samples were obtained from patients who underwent Nephrectomy at the First Affiliated Hospital of China Medical University. This study was authorized by the Ethics Committee of the First Affiliated Hospital of China Medical University. All patients signed informed consent. Protein exaction and western blotting were conducted as described previously (Pripp, 2018). An antibody against VSIG4 was purchased from Sigma-Aldrich.

Statistical Methods

Pearson correlation coefficients measure the strength of the linear relationship between two variables. The correlation coefficients are -1 to $+1$, respectively, indicating negative correlation and positive correlation, respectively, while 0 indicates no correlation (Wang Y. et al., 2020). The key factors in the model score, tumor purity, tumor mutation burden, M2 macrophages, and CD8$^+$ T lymphocytes were assessed using this test.

RESULTS

M2 Macrophages, Tumor Purity, and Tumor Mutation Burden

The results of our methodology are explained in **Figure 1**.

We summed up the following clinical data composed by M2 macrophages, tumor mutation burden, and clinical following survival data. M2, and M1, and M2/M1 macrophages were inputted as phenotype files to WGCNA. The detailed information is displayed in **Supplementary Table 1**.

M2 Macrophages Co-expression Network Conduction

We performed WGCNA analysis with TCGA–KIRC. A hierarchical clustering tree was built using the dynamic hybrid cutting method, where each leaf on the tree represents a gene, and each branch represents a co-expression module; 21 co-expression models were generated (**Figure 2A**). The correlation coefficients

TABLE 1 | The Module and gene significance for M2 macrophage-related genes in the purple module.

ID	moduleColor	GS.MacrophagesM2	p.GS.M2	GS.CD8.T	p.GS.CD8.
CD27	purple	−0.497	1.32E-18	0.774	2.91E-56
PSMB9	purple	−0.493	2.94E-18	0.715	1.65E-44
CTSW	purple	−0.488	7.07E-18	0.787	2.57E-59
CD3E	purple	−0.483	1.57E-17	0.734	6.70E-48
CST7	purple	−0.482	1.77E-17	0.799	2.39E-62
CD3D	purple	−0.480	2.90E-17	0.755	5.32E-52
SIT1	purple	−0.479	3.05E-17	0.753	1.02E-51
HLA-F	purple	−0.476	5.88E-17	0.689	3.94E-40
IL2RG	purple	−0.475	6.28E-17	0.649	2.22E-34
GZMA	purple	−0.468	2.22E-16	0.7789	2.89E-57
NKG7	purple	−0.468	2.22E-16	0.762	1.28E-53
CD8B	purple	−0.467	2.28E-16	0.832	6.65E-72
PRF1	purple	−0.466	2.80E-16	0.744	9.06E-50
CD8A	purple	−0.466	2.93E-16	0.830	3.40E-71
LCK	purple	−0.465	4.62E-16	0.694	2.54E-42
APOBEC3G	purple	−0.461	6.88E-16	0.715	2.18E-44
HLA-B	purple	−0.459	9.61E-16	0.682	4.66E-39
CXCR3	purple	−0.458	1.03E-15	0.684	2.29E-39
IRF1	purple	−0.457	1.30E-15	0.671	2.24E-37
CD2	purple	−0.449	4.11E-15	0.729	6.85E-47
DUSP2	purple	−0.449	4.43E-15	0.763	1.13E-53
CCL5	purple	−0.447	5.76E-15	0.588	4.94E-27
HLA-E	purple	−0.423	1.08E-14	0.688	1.58E-35
PSME2	purple	−0.409	1.98E-14	0.525	1.11E-43

GS, Gene significance.

between each phenotype and co-expression module of TCGA are shown in **Figure 2B**. The results showed that the purple module had the strongest negatively correlation with M2 macrophage cell proportion in the TCGA–KIRC cohort (Cor = −0.45; $P = 4e^{-15}$) and had the strongest correlation with CD8^{+} T cell proportion in the TCGA–KIRC cohort (Cor = 0.73; $P = 6e^{-47}$) (**Figure 2B**). Based on these findings, we have supplemented the scatter plots of the correlation between the factors in the purple module (**Figures 2C–E**). The horizontal axis is the correlation between the gene and the module, which is used to measure the relationship between the gene and the co-expression module, and the vertical axis is the correlation between the gene and the

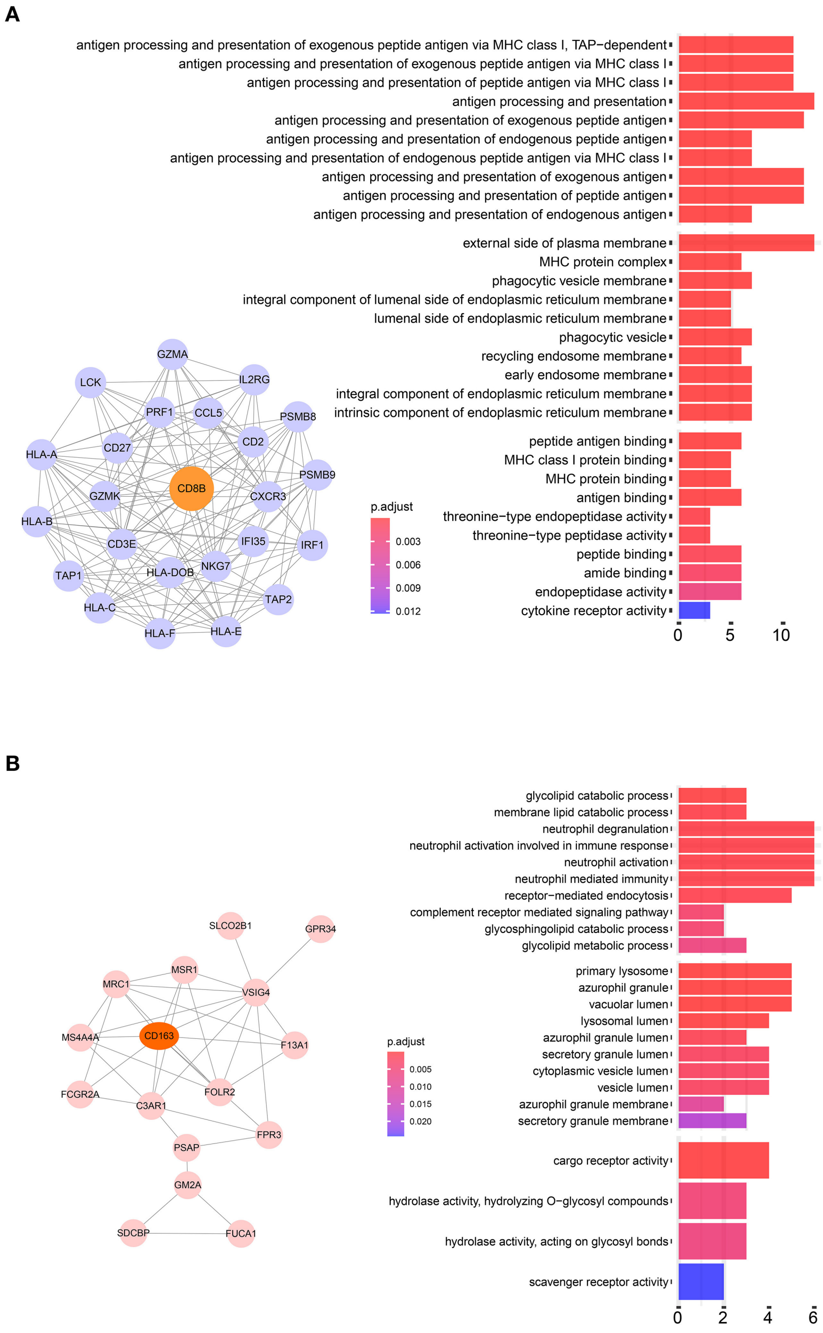


FIGURE 3 | (A) Pathway analysis of 24 negatively correlated co-expressed genes in M2 macrophages in the purple module. These genes were most significantly enriched in the antigen processing and presentation of exogenous peptide antigen via MHC class I, which suggested a declining effect on tumor antigen peptide process. **(B)** Pathway analysis of 16 negatively correlated co-expressed genes in M2 macrophages in the brown module. These genes were most significantly enriched in neutrophil activation involved in immune responses.

macrophage. By drawing the scatter diagram above, we screened out genes that are related to both M2 macrophages and the co-expression module.

M2 Related Genes Function Analysis

Twenty-four M2 macrophage negatively co-expressing genes were identified with coefficient <-0.4 in the TCGA–KIRC purple module. The gene significance for M2 macrophage-related genes in the purple module is shown in **Table 1**. Top 20 M2 macrophage cell proportion positively co-expressing genes were identified in the TCGA–KIRC pink module. The 24 M2 macrophage negatively co-expressing genes were most significantly enriched in the antigen processing and presentation of exogenous peptide antigen via MHC class I, which suggested a declining effect on the tumor antigen peptide process (**Figure 3A**). The 20 M2 macrophage negatively co-expressing genes were most significantly enriched in neutrophil activation involved in immune responses (**Figure 3B**).

M2 Related Genes Prognosis Analysis

To analyze their influence on overall survival, we performed survival analysis. F13A1, FCGR2A, HLA.DOB, ILR2GHLA, DUSP2, PSME2, CD27, IFI35, LIMD2, NFKB2, IL2RB, CCL5, VSIG4, APOBEC3G, GZMA, and PSMB10 were prognosis risk factors for clear renal cell carcinoma. HLA-E, MRC1, GPR34, KCTD12, LIPA, PSAP, MFSD1, EHD1, FUCA1, and CPVL

TABLE 2 | The Module and gene significance for M2 macrophage-related genes in the pink module.

ID	moduleColor	GS.MacrophagesM2	p.GS.M2
GPR34	pink	0.467	2.31E-16
MS4A4A	pink	0.452	2.93E-15
MFSD1	pink	0.446	6.88E-15
FUCA1	pink	0.435	3.55E-14
CD163	pink	0.428	1.07E-13
FOLR2	pink	0.427	1.13E-13
LIPA	pink	0.424	1.78E-13
SLCO2B1	pink	0.418	4.43E-13
PSAP	pink	0.415	6.44E-13
SDCBP	pink	0.404	3.17E-12
C3AR1	pink	0.395	9.95E-12
F13A1	pink	0.391	1.60E-11
KCTD12	pink	0.386	3.14E-11
MSR1	pink	0.384	4.12E-11
CPVL	pink	0.365	4.15E-10
FCGR2A	pink	0.362	5.48E-10
FPR3	pink	0.358	9.40E-10
GM2A	pink	0.353	1.68E-09
VSIG4	pink	0.347	3.12E-09
MRC1	pink	0.337	9.28E-09

GS, Gene significance.

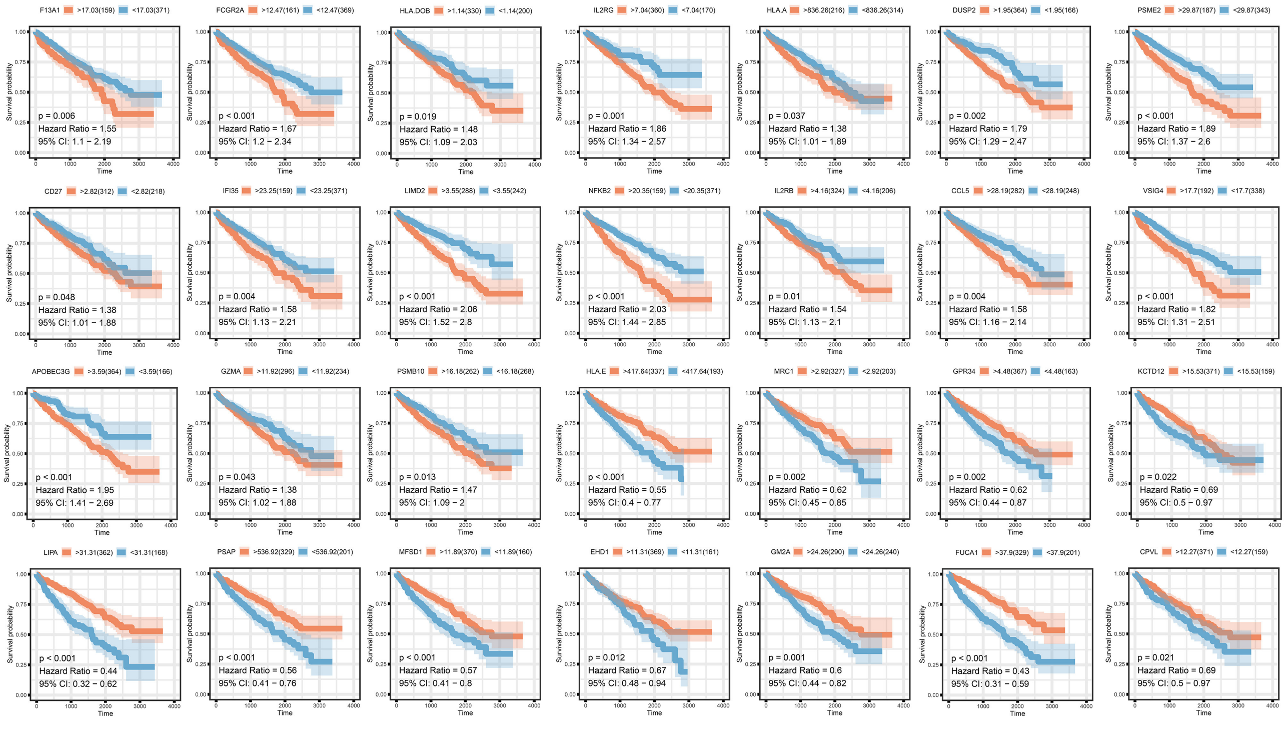


FIGURE 4 | Survival analysis of selected co-expressed genes in purple and pink modules.

were prognosis-protective factors for clear renal cell carcinoma (**Figure 4**).

M2 Macrophage-Related Prognosis Signature

We then generated a multi-Cox regression risk score model based on M2 macrophage-related genes (**Tables 1**, **2**). Risk score = 0.025 * F13A1 – 0.008 * FUCA1 + 0.034 * FCGR2A – 0.016 * KCTD12 – 0.08 * MFSD1 – 0.003 * HLA-E + 0.012 * SDCBP – 0.071 * MRC1 – 0.086 * LCK + 0.02 * PSME2 + 0.016 * VSIG4 + 0.215 * TAP2. Detailed information of the prognosis model is displayed in **Supplementary Table 2**. The patients in high-risk groups for renal clear cell cancer (TCGA: $P < 0.001$; HR = 5.31) (**Figure 5**) showed survival risk vs. low

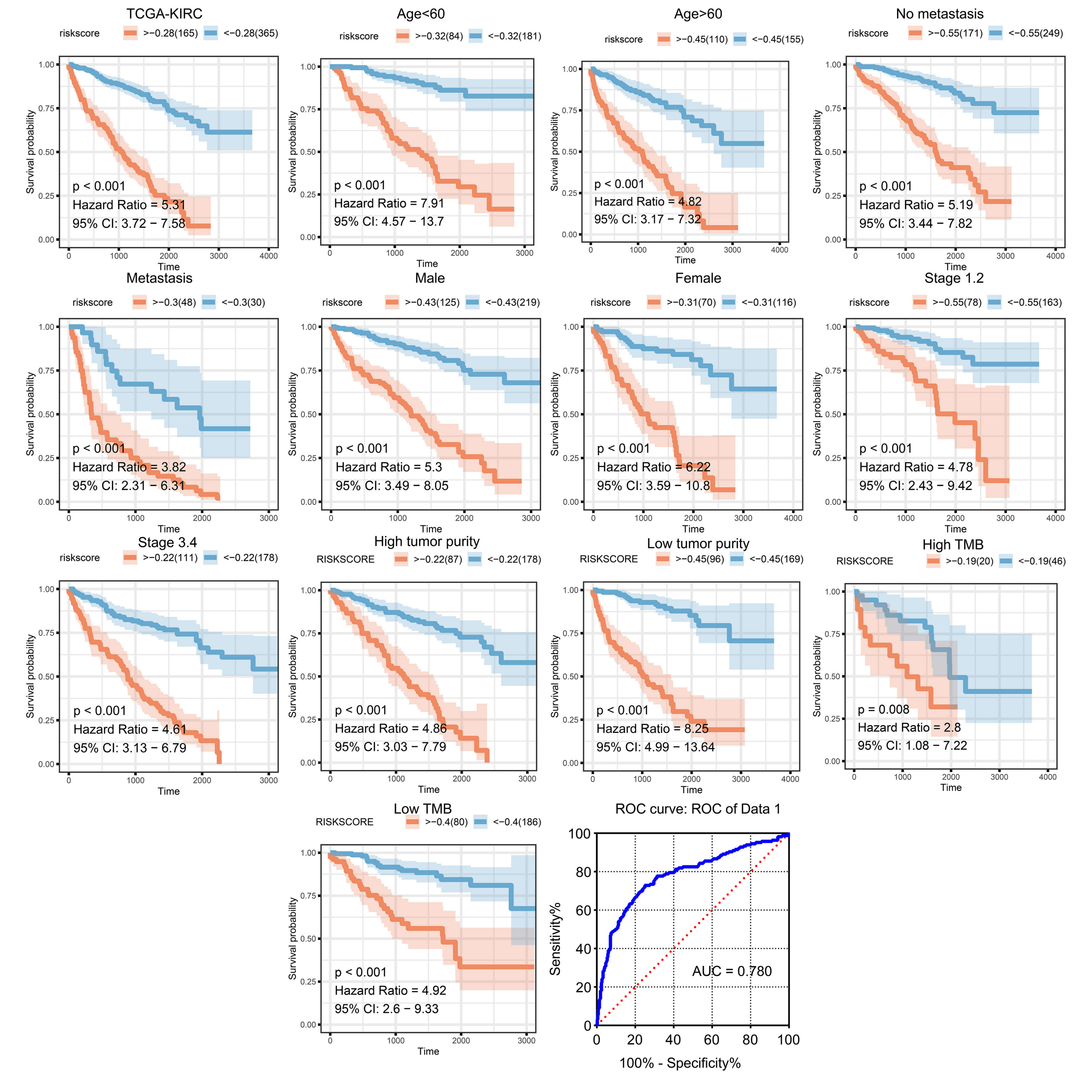


FIGURE 5 | Validation of the prognostic model in clinical subgroups. The patients in high-risk groups for renal clear cell cancer (TCGA: $P < 0.001$; HR = 5.31) showed survival risk against low expression groups, with the area under curve (AUC) = 0.780. The risk score was evaluated in clinical subgroups, including age, gender, stage, metastasis, tumor purity, and tumor mutation burden. P-values of all subgroups validations were <0.05, indicating that this model has good predictive ability.

expression groups, with the area under the curve (AUC) = 0.780 (**Figure 5**). The risk score was evaluated in various subgroups, including age, gender, stage, metastasis, tumor purity, and tumor mutation burden. The results were significant in these subgroups (**Figure 5**).

Immune Environment Correlation

Significant associations between M2 frequency and the genes involved in the risk signature are indicated in **Figure 6**, and the highest correlation of MFSD1 was 0.49 (**Figure 6A**); the correlation of LCK was the lowest at −0.47 (**Figure 6B**). TAP2, PSME2, HLA-E, and LCK were negatively related to M2 macrophage proportions. We then analyzed the correlations with CD8$^+$ T cell and tumor mutation burden of these four genes. TAP2 ($P < 0.001$; Cor = 0.60), PSME2 ($P < 0.001$; Cor = 0.52), HLA - E ($P < 0.001$; Cor = 0.69), and LCK ($P < 0.001$; Cor = 0.69) (**Figure 7A**) positively related to CD8$^+$ T cell and negatively correlated with tumor purity (**Figure 7B**). This result suggested that M2 macrophages were negatively related to antigen processing.

HPA

The prognostic value and immune phenotype correlation were determined for these M2 macrophage-related genes. We compared the various expression levels of these genes between normal and tumor tissues. HPA001804 is an antibody against F13A1, which showed higher intensity in tumor tissue than in normal tissue. HPA056371 is an antibody against FUCA1, which showed higher intensity in the normal tissue than in tumor tissue. CAB012245 is an antibody against SDCBP, which showed a higher intensity in tumor tissue than in normal tissue. HPA003903 is an antibody against VSIG4, which showed higher intensity in tumor tissue than in normal tissue. HPA031454 is an antibody against HLA-E, which showed lower intensity in tumor tissue than in normal tissue. HPA001312 is an antibody against TAP2, which showed lower intensity in tumor tissue than in normal tissue. The protein levels

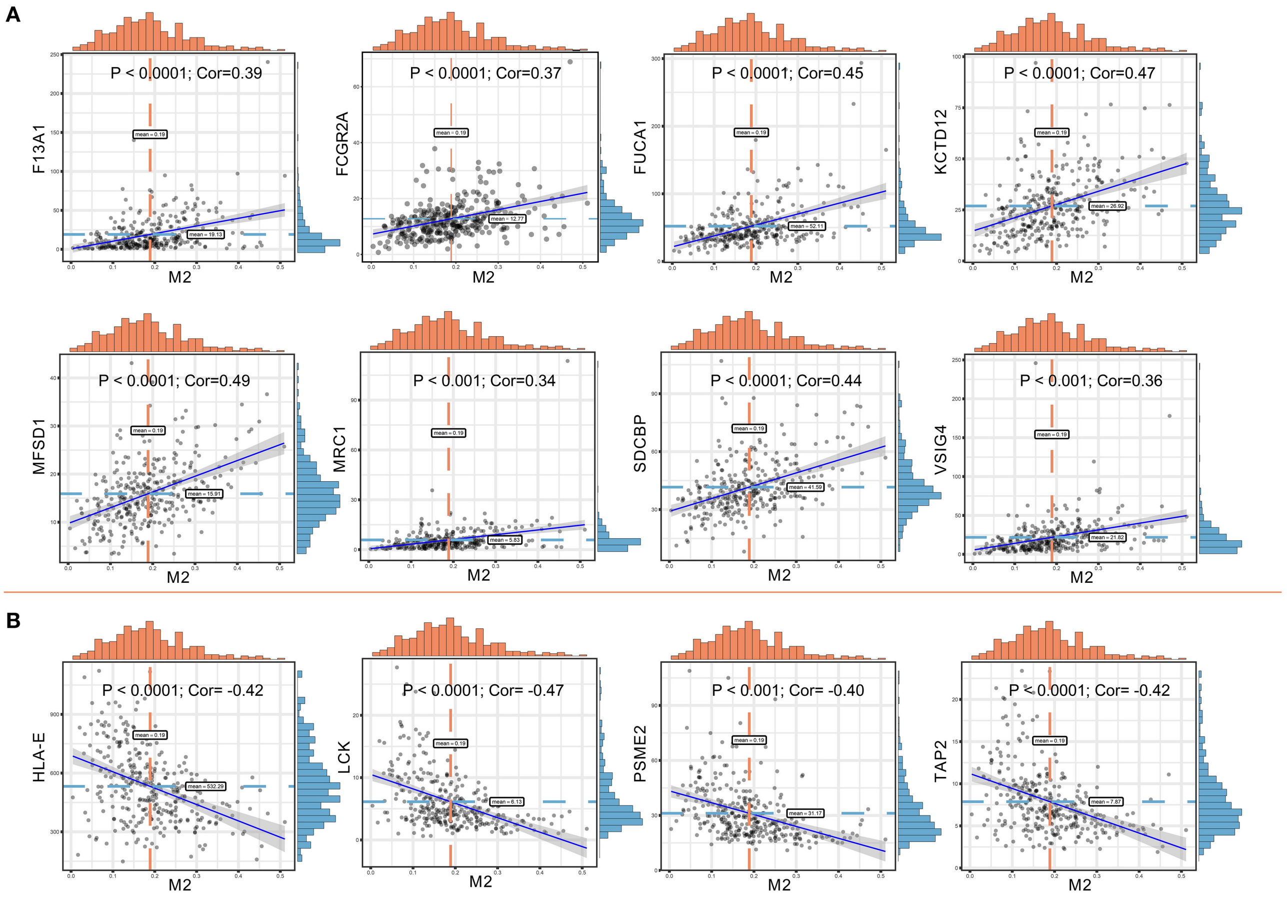


FIGURE 6 | (A) Co-expressed genes with a significant positive correlation with M2 macrophages. The correlation coefficients are as follows: F13A1 – M2: Cor = 0.39; FCGR2A – M2: Cor = 0.37; FUCA1 – M2: Cor = 0.45; KCTD12 – M2: Cor = 0.47; MFSD1 – M2: Cor = 0.49; MRC1 – M2: Cor = 0.34; SDCBP – M2: Cor = 0.44; VSIG4 – M2: Cor = 0.36. **(B)** Co-expressed genes with a significant negative correlation with M2 macrophages. The correlation coefficients are as follows: HLA-E – M2: Cor = −0.42; LCK – M2: Cor = −0.47; PSME2 – M2: Cor = −0.40; TAP2 – M2: Cor = −0.42.

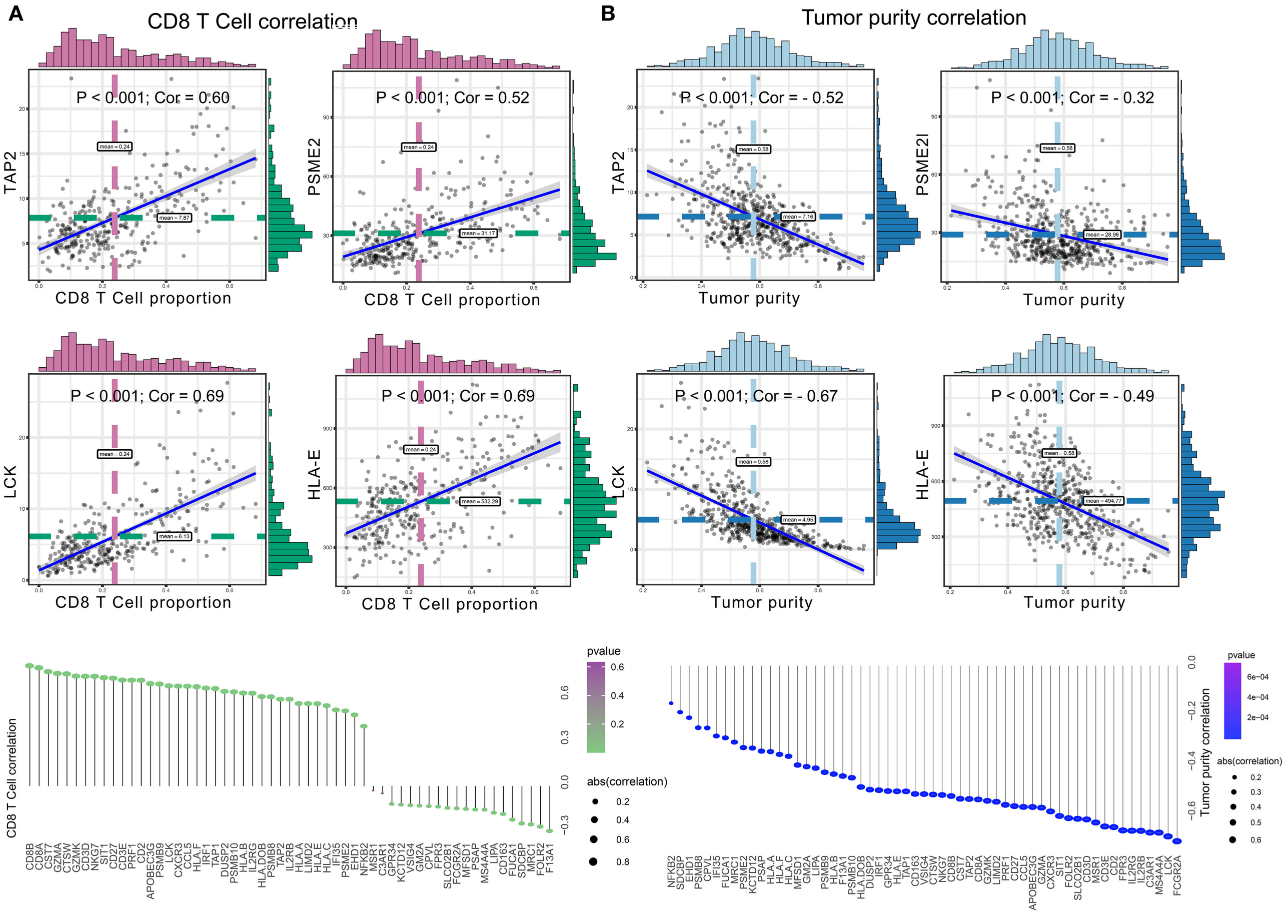


FIGURE 7 | (A) The correlation between co-expressed gene of M2 macrophage and CD8$^+$ T cell, with significantly positive relations as TAP2 ($P < 0.001$; Cor = 0.60), PSME2 ($P < 0.001$; Cor = 0.52), HLA – E ($P < 0.001$; Cor = 0.69), and LCK ($P < 0.001$; Cor = 0.69). **(B)** The correlation between co-expressed genes of M2 macrophage and tumor purity, with significantly negative relations as TAP2 ($P < 0.001$; Cor = −0.52), PSME2 ($P < 0.001$; Cor = −0.32), LCK ($P < 0.001$; Cor = −0.67), and HLA-E ($P < 0.001$; Cor = −0.49).

of these M2 macrophage genes were similar to the results of prognosis analysis at the transcription level (**Figure 8**). Subsequently, the M2 correlations for VSIG4, FUCA1, F13A1, SDCBP, HLA-E, and TAP2 were verified in the four GEO datasets (**Supplementary Figure 1**).

VSIG4

VSIG4 is thought to positively correlate with M2 macrophages; therefore, we conducted a combined analysis of VSIG4 and M2-type macrophages. Combining VSIG4 elevated the predictive accuracy of M2 macrophages even more than either of them alone; the hazard of the "high VSIG4 expression + high M2 macrophage" group showed more survival risk than the other group (Kaplan–Meier analysis, low VSIG4 expression + low M2 macrophage; HR = 1.458; **Figure 9A**). Subsequently, we compared VSIG4 protein expression levels between normal renal tissues and clear renal cell carcinoma and found that VSIG4 protein expression levels in tumors were higher than the normal tissues (**Figure 9B**). Then, various tumor infiltration deconvolution methods were applied; we found that VSIG4 was one of the most commonly associated M2 macrophage biomarkers (**Figure 9C**).

DISCUSSION

In the tumor microenvironment, the chemotactic effects of the functional differences between the types of tumor-associated macrophages are not completely clear. The biological cytological role of M2/M1 macrophages in tumor tissues still needs to be explored. The present study is based on a bioinformatics algorithm to determine some of the M2 macrophage co-expression networks. Through the analysis of various modules, we tried to explain the biological function of co-expressed genes with M2 macrophages and related pathway changes from the perspective of bioinformatics. Our data processing and analysis processes are shown in the flowchart (**Figure 1**).

F13A1, FCUA1, HLA-E, VSIG4, SDCBP, and TAP2 were the most common co-expressed genes in M2 macrophages. In terms

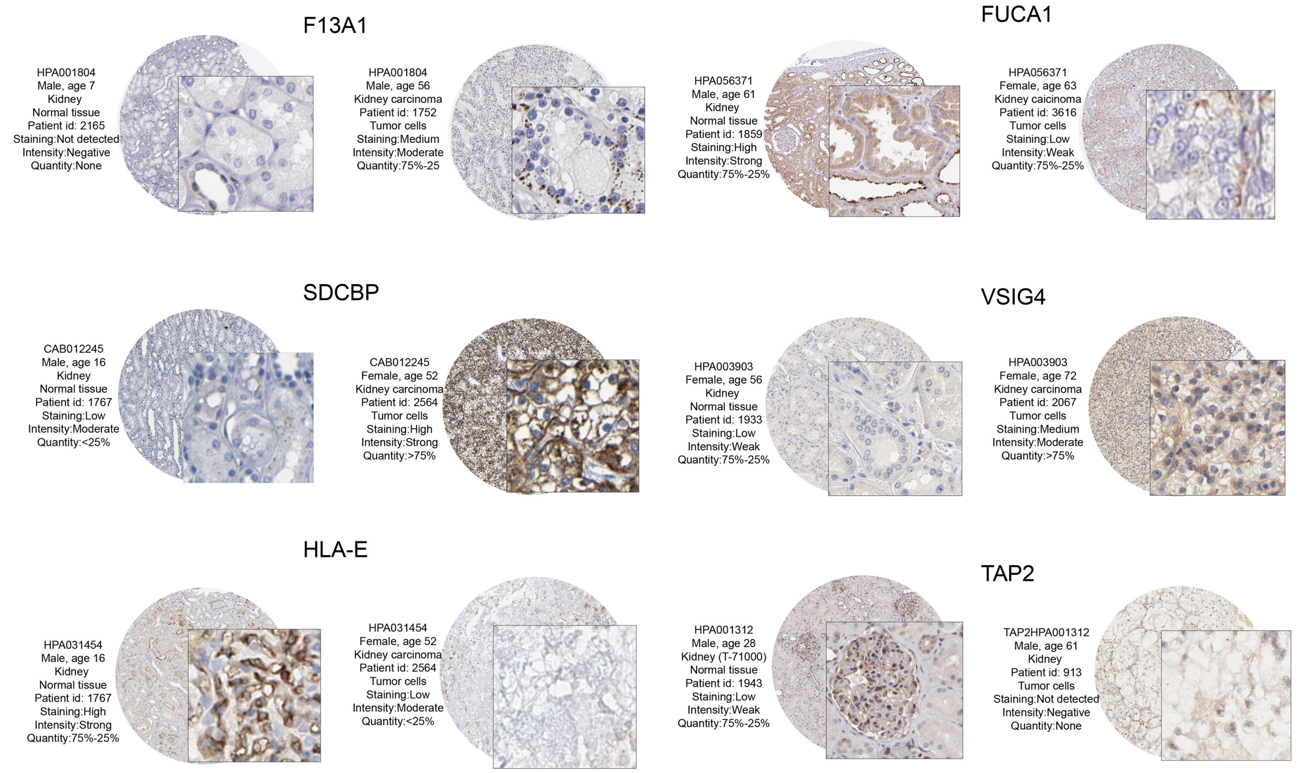


FIGURE 8 | From the HPA database to verify protein expression-level differences of these candidate genes. Of these, F13A1, SDCBP, and VSIG4, and corresponding immunohistochemical samples, the degree of renal clear cell carcinoma tissue staining is higher than in normal kidney tissue. In FUCA1, HLA – E, and TAP2, and corresponding immunohistochemical samples, the degree renal clear cell carcinoma tissue staining is lower than in normal kidney tissue. These M2 macrophage gene protein levels at the transcription level were similar to those of the prognostic analysis.

of function enrichment, the 24 negatively co-expressed genes in M2 macrophages were most significantly enriched in antigen processing and presentation of exogenous peptide antigen via MHC class I. The 16 negatively co-expressing genes in M2 macrophages were most significantly enriched in neutrophil activation involved in immune response. M1 macrophages tend to adopt a Th1 response gene expression pattern and can secrete various cytokines that present MHC II and B7 molecules so as to present antigen efficiently (Herberman et al., 1979). This mechanism resists pathogen invasion, monitors tumor pathological changes, and generates Th1 immune responses in macrophages. By contrast, M2 macrophages have poor tumor antigen processing ability.

F13A1 encodes the coagulation factor XIII A subunit which has a catalytic function. In a human stem cell study, mRNA transcription expressed by F13A1 increased as myeloid progenitors differentiated into macrophages and erythroblasts (De Paoli et al., 2015). The protein encoded by FCUA1 is a lysosomal enzyme involved in the degradation of fucose-containing glycoproteins and glycolipids. Downregulation of FUCA1 enhances autophagy and inhibits macrophage infiltration so as to inhibit tumor growth (Xu L. et al., 2020). VSIG4 is a transmembrane receptor of the immunoglobulin superfamily that is specifically expressed in macrophages and mature dendritic cells. It is a newly discovered B7 family-related macrophage protein that inhibits T cell activation and has a potential role in cancer (Kim et al., 2016). VSIG4 negatively regulates macrophage activation by reprogramming mitochondrial pyruvate metabolism (Li et al., 2017). HLA-E belongs to the HLA class I heavy chain paralogs. This class I molecule is a heterodimer consisting of a heavy chain and a light chain (beta-2 microglobulin). The heavy chain is anchored in the membrane. HLA-E binds a restricted subset of peptides derived from the leader peptides of other class I molecules. HlA-E is a non-classical HLA-I molecule that is best known for its role in protecting natural killer cells. Camilli et al. found that HLA-E was significantly increased during the differentiation of monocytes and macrophages (Camilli et al., 2016). The expression of HLA-E is related to the poor clinical results of anti-PD-1 immunotherapy. From the surface of M2 tumor-associated macrophages (TAMs), HLA-E antigen binds to the receptor CD94/NKG2A, which inhibits the expression of NK cell subpopulations and activated cytotoxic T lymphocytes, protecting cells from being destroyed (Marchesi et al., 2013). Epithelial-derived cancer cells, tumor macrophages, and CD141^{+} traditional dendritic

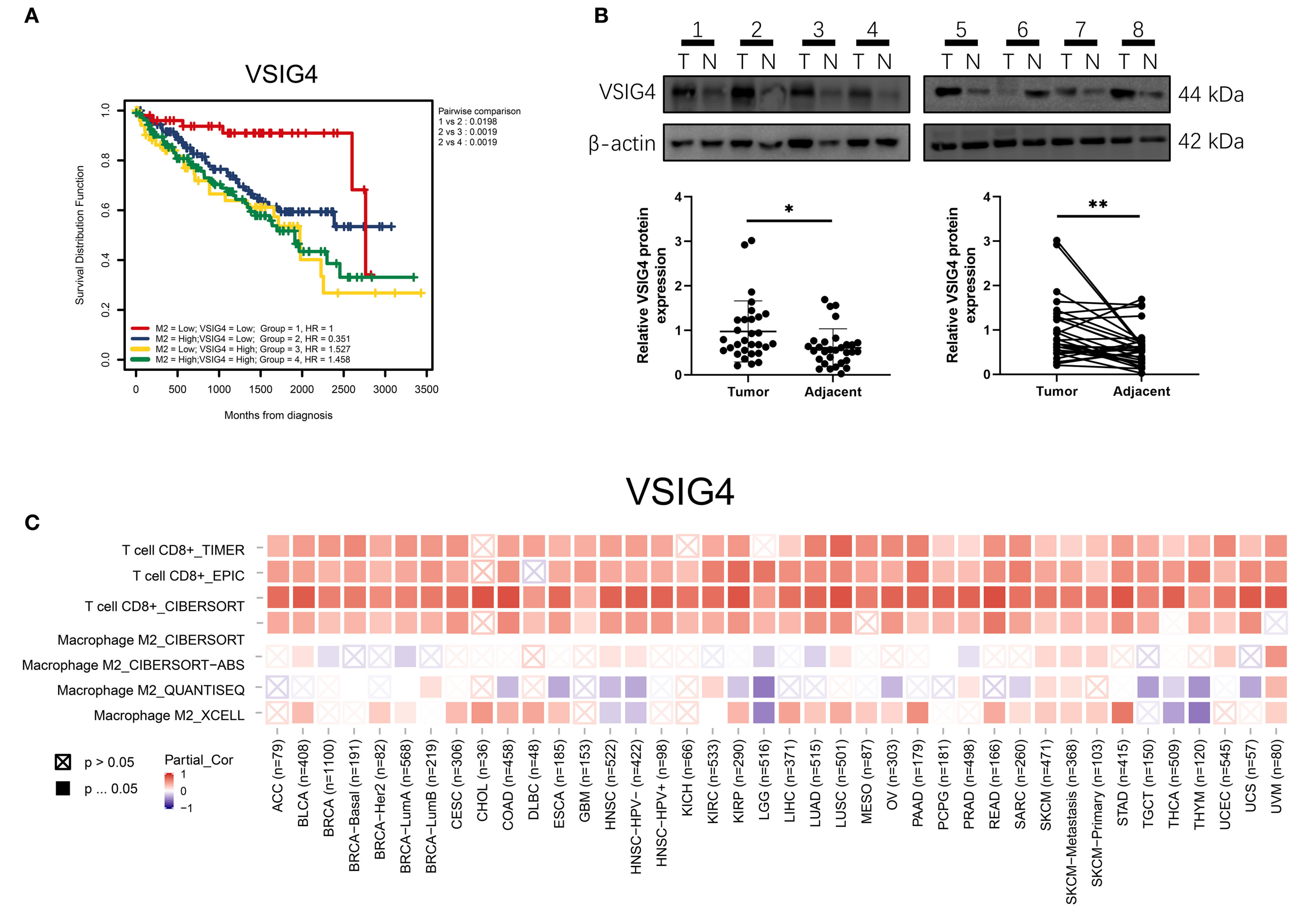


FIGURE 9 | (A) Combining high VSIG4 and high M2 macrophage showed more survival risk than the other group. **(B)** The VSIG4 protein expression levels were higher in clear renal cell carcinoma tissues than in normal tissues according to western blotting. **(C)** Pan-cancer analysis of VSIG4 in TCGA.

cells promote the enrichment of HLA-E in carcinomas. CD8$^+$ tumor-infiltrating T lymphocytes with high PD-1 content are prevented from surviving in the tumor microenvironment by the interaction of enriched HLA-E and CD94/NKG2A inhibition (Abd Hamid et al., 2019).

This study has some limitations, including lack of cross-validation of multicenter data. There is also lack of experimental verification of M2 macrophage biomarkers in renal clear cell cancer. We found that using the co-expression method of network-building, we can explicitly identify biomarkers, demonstrating the correctness of the logic based on bioinformatics.

In conclusion, we found that F13A1, FCUA1, HLA-E, VSIG4, SDCBP, and TAP2 were biomarkers of M2-type macrophages using a co-expression network of infiltrated immune cells, and we proposed six candidate-related factors. The biomarkers and related processes of M2 macrophages in the tumor microenvironment were explained from the perspective of bioinformatics, providing a strategy to explore the polarization of macrophages.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The TCGA-BLCA dataset used in this study were obtained from TCGA database (https://cancergenome.nih.gov/). GEO datasets used in this study were obtained from GEO database (https://www.ncbi.nlm.nih.gov/geo/).

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

YW, KY, and JB designed the study. YW, KY, JLin, JLi, and JB analyzed and wrote the article.

All authors read and agreed to the final version of the manuscript.

FUNDING

This work was supported by the China Shenyang Science and Technology Plan (20-205-4-015).

ACKNOWLEDGMENTS

We appreciate the free use of TCGA and GEO databases.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.615655/full#supplementary-material

Supplementary Figure 1 | (A) The multi-dataset correction of GSE8050, GSE12606, GSE14762, and GSE36895 using the "sva" package. **(B)** The verification analysis of M2 related genes in the combined cohorts.

Supplementary Table 1 | The detail information of tumor mutation, tumor purity, and M2 macrophage proportion.

Supplementary Table 2 | The process data of multi-Cox regression risk score model.

REFERENCES

Abd Hamid, M., Wang, R. Z., Yao, X., Fan, P., Li, X., Chang, X. M., et al. (2019). Enriched HLA-E and CD94/NKG2A interaction limits antitumor CD8+ tumor-infiltrating T lymphocyte responses. *Cancer Immunol. Res*. 7, 1293–1306. doi: 10.1158/2326-6066.CIR-18-0885

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*. 25, 25–29. doi: 10.1038/75556

Camilli, G., Cassotta, A., Battella, S., Palmieri, G., Santoni, A., Paladini, F., et al. (2016). Regulation and trafficking of the HLA-E molecules during monocyte-macrophage differentiation. *J. Leukoc. Biol*. 99, 121–130. doi: 10.1189/jlb.1A0415-172R

Cervantes-Villagrana, R. D., Albores-García, D., Cervantes-Villagrana, A. R., and García-Acevez, S. J. (2020). Tumor-induced neurogenesis and immune evasion as targets of innovative anti-cancer therapies. *Signal Transduct. Target Ther*. 5:99. doi: 10.1038/s41392-020-0205-z

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., and Alizadeh, A. A. (2018). Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol. Biol*. 1711, 243–259. doi: 10.1007/978-1-4939-7493-1_12

Chowdhury, N., and Drake, C. G. (2020). Kidney cancer: an overview of current therapeutic approaches. *Urol. Clin. North Am*. 47, 419–431. doi: 10.1016/j.ucl.2020.07.009

De Paoli, F., Eeckhoute, J., Copin, C., Vanhoutte, J., Duhem, C., Derudas, B., et al. (2015). The neuron-derived orphan receptor 1 (NOR1) is induced upon human alternative macrophage polarization and stimulates the expression of markers of the M2 phenotype. *Atherosclerosis* 241, 18–26. doi: 10.1016/j.atherosclerosis.2015.04.798

DeNardo, D. G., and Ruffell, B. (2019). Macrophages as regulators of tumour immunity and immunotherapy. *Nat. Rev. Immunol*. 19, 369–382. doi: 10.1038/s41577-019-0127-6

Díaz-Montero, C. M., Rini, B. I., and Finke, J. H. (2020). The immunology of renal cell carcinoma. *Nat. Rev. Nephrol*. 16, 721–735. doi: 10.1038/s41581-020-0316-3

Herberman, R. B., Holden, H. T., Djeu, J. Y., Jerrells, T. R., Varesio, L., Tagliabue, A., et al. (1979). Macrophages as regulators of immune responses against tumors. *Adv. Exp. Med. Biol*. 121B, 361–379. doi: 10.1007/978-1-4684-8914-9_35

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

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., et al. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*. 8:R183. doi: 10.1186/gb-2007-8-9-r183

Jiang, J., Sun, X., Wu, W., Li, L., Wu, H., Zhang, L., et al. (2017). Corrigendum: Construction and application of a co-expression network in Mycobacterium tuberculosis. *Sci. Rep*. 7:40563. doi: 10.1038/srep40563

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 45, D353–D361. doi: 10.1093/nar/gkw1092

Karnevi, E., Andersson, R., and Rosendahl, A. H. (2014). Tumour-educated macrophages display a mixed polarisation and enhance pancreatic cancer cell invasion. *Immunol. Cell Biol*. 92, 543–552. doi: 10.1038/icb.2014.22

Kim, K. H., Choi, B. K., Kim, Y. H., Han, C., Oh, H. S., Lee, D. G., et al. (2016). Extracellular stimulation of VSIG4/complement receptor Ig suppresses intracellular bacterial infection by inducing autophagy. *Autophagy* 12, 1647–1659. doi: 10.1080/15548627.2016.1196314

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

Lawrence, T., and Natoli, G. (2011). Transcriptional regulation of macrophage polarization: enabling diversity with identity. *Nat. Rev. Immunol*. 11, 750–761. doi: 10.1038/nri3088

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

Li, J., Diao, B., Guo, S., Huang, X., Yang, C., Feng, Z., et al. (2017). VSIG4 inhibits proinflammatory macrophage activation by reprogramming mitochondrial pyruvate metabolism. *Nat. Commun*. 8:1322. doi: 10.1038/s41467-017-01327-4

Li, Y., Chen, Z., Wu, L., and Tao, W. (2020). Novel tumor mutation score versus tumor mutation burden in predicting survival after immunotherapy in pan-cancer patients from the MSK-IMPACT cohort. *Ann. Transl. Med*. 8:446. doi: 10.21037/atm.2020.03.163

Linehan, W. M., and Ricketts, C. J. (2019). The Cancer Genome Atlas of renal cell carcinoma: findings and clinical implications. *Nat. Rev. Urol*. 16, 539–552. doi: 10.1038/s41585-019-0211-5

Marchesi, M., Andersson, E., Villabona, L., Seliger, B., Lundqvist, A., Kiessling, R., et al. (2013). HLA-dependent tumour development: a role for tumour associate macrophages. *J. Transl. Med*. 11:247. doi: 10.1186/1479-5876-11-247

Motzer, R. J., Hutson, T. E., Cella, D., Reeves, J., Hawkins, R., Guo, J., et al. (2013). Pazopanib versus sunitinib in metastatic renal-cell carcinoma. *N. Engl. J. Med*. 369, 722–731. doi: 10.1056/NEJMoa1303989

Peña-Llopis, S., Vega-Rubín-de-Celis, S., Liao, A., Leng, N. A., Pavía-Jiménez, Wang, S., et al. (2012). BAP1 loss defines a new class of renal cell carcinoma. *Nat. Genet*. 44, 751–759. doi: 10.1038/ng.2323

Pollard, J. W. (2004). Tumour-educated macrophages promote tumour progression and metastasis. *Nat. Rev. Cancer* 4, 71–78. doi: 10.1038/nrc1256

Pripp, A. H. (2018). [Pearson's or Spearman's correlation coefficients]. *Tidsskr. Nor. Laegeforen*. 138:42. doi: 10.4045/tidsskr.18.0042

Stickel, J. S., Weinzierl, A. O., Hillen, N., Drews, O., Schuler, M. M., Hennenlotter, J., et al. (2009). HLA ligand profiles of primary renal cell carcinoma maintained in metastases. *Cancer Immunol. Immunother*. 58, 1407–1417. doi: 10.1007/s00262-008-0655-6

Uhlén, M., Fagerberg, L., Hallström, 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, C., Wang, Y., Hong, T., Ye, J., Chu, C., Zuo, L., et al. (2020). Targeting a positive regulatory loop in the tumor-macrophage interaction impairs the progression of clear cell renal cell carcinoma. *Cell Death Differ*. doi: 10.1038/s41418-020-00626-6

Wang, Y., Roche, O., Yan, M. S., Finak, G., Evans, A. J., Metcalf, J. L., et al. (2009). Regulation of endocytosis via the oxygen-sensing pathway. *Nat. Med.* 15, 319–324. doi: 10.1038/nm.1922

Wang, Y., Yan, K., Lin, J., Wang, J., Zheng, Z., Li, X., et al. (2020). Three-gene risk model in papillary renal cell carcinoma: a robust likelihood-based survival analysis. *Aging* 12, 21854–21873. doi: 10.18632/aging.104001

Weinzierl, A. O., Maurer, D., Altenberend, F., Schneiderhan-Marra, N., Klingel, K., Schoor, O., et al. (2008). A cryptic vascular endothelial growth factor T-cell epitope: identification and characterization by mass spectrometry and T-cell assays. *Cancer Res.* 68, 2447–2454. doi: 10.1158/0008-5472.CAN-07-2540

Xu, L., Li, Z., Song, S., Chen, Q., Mo, L., Wang, C., et al. (2020). Downregulation of α-l-fucosidase 1 suppresses glioma progression by enhancing autophagy and inhibiting macrophage infiltration. *Cancer Sci.* 111, 2284–2296. doi: 10.1111/cas.14427

Xu, W., Atkins, M. B., and McDermott, D. F. (2020). Checkpoint inhibitor immunotherapy in kidney cancer. *Nat. Rev. Urol.* 17, 137–150. doi: 10.1038/s41585-020-0282-3

Yang, Z., Wei, S., Deng, Y., Wang, Z., and Liu, L. (2020). Clinical significance of tumour mutation burden in immunotherapy across multiple cancer types: an individual meta-analysis. *Jpn. J. Clin. Oncol.* 50, 1023–1031. doi: 10.1093/jjco/hyaa076

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612. doi: 10.1038/ncomms3612

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.

ORIGINAL RESEARCH
published: 23 February 2021
doi: 10.3389/fgene.2021.638458

Expression Profile Analysis Identifies a Novel Seven Immune-Related Gene Signature to Improve Prognosis Prediction of Glioblastoma

Li Hu[†], *Zhibin Han*[†], *Xingbo Cheng, Sida Wang, Yumeng Feng and Zhiguo Lin*[*]

Department of Neurosurgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Jingting Yu,
Salk Institute for Biological Studies,
United States
Kuixi Zhu,
University of Arizona, United States
Yu Dong,
Shanghai Jiao Tong University, China

***Correspondence:**
Zhiguo Lin
linzhiguo@hotmail.com;
linzhgu@sina.com

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

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

Received: 06 December 2020
Accepted: 01 February 2021
Published: 23 February 2021

Citation:
Hu L, Han Z, Cheng X, Wang S, Feng Y and Lin Z (2021) Expression Profile Analysis Identifies a Novel Seven Immune-Related Gene Signature to Improve Prognosis Prediction of Glioblastoma. Front. Genet. 12:638458. doi: 10.3389/fgene.2021.638458

Glioblastoma multiform (GBM) is a malignant central nervous system cancer with dismal prognosis despite conventional therapies. Scientists have great interest in using immunotherapy for treating GBM because it has shown remarkable potential in many solid tumors, including melanoma, non-small cell lung cancer, and renal cell carcinoma. The gene expression patterns, clinical data of GBM individuals from the Cancer Genome Atlas database (TCGA), and immune-related genes (IRGs) from ImmPort were used to identify differentially expressed IRGs through the Wilcoxon rank-sum test. The association between each IRG and overall survival (OS) of patients was investigated by the univariate Cox regression analysis. LASSO Cox regression assessment was conducted to explore the prognostic potential of the IRGs of GBM and construct a risk score formula. A Kaplan–Meier curve was created to estimate the prognostic role of IRGs. The efficiency of the model was examined according to the area under the receiver operating characteristic (ROC) curve. The TCGA internal dataset and two GEO external datasets were used for model verification. We evaluated IRG expression in GBM and generated a risk model to estimate the prognosis of GBM individuals with seven optimal prognostic expressed IRGs. A landscape of 22 types of tumor-infiltrating immune cells (TIICs) in glioblastoma was identified, and we investigated the link between the seven IRGs and the immune checkpoints. Furthermore, there was a correlation between the IRGs and the infiltration level in GBM. Our data suggested that the seven IRGs identified in this study are not only significant prognostic predictors in GBM patients but can also be utilized to investigate the developmental mechanisms of GBM and in the design of personalized treatments for them.

Keywords: glioblastoma, expression profile, immune-related genes, prognosis prediction, overall survival

INTRODUCTION

Glioblastoma constitutes the most recurrent and aggressive primary malignant tumor of the central nervous system (Yan et al., 2012). In spite of multimodal conventional treatments consisting of neurosurgical resection as well as radiotherapy with accompanying adjuvant alkylating agent temozolomide chemotherapy, the prognosis for glioblastoma multiform (GBM) individuals remains dismal, with a median survival time ranging from 9.4 to 19.0 months (Yang et al., 2014).

This poor outcome is due to the highly invasive nature, malignant progression, drug resistance, and tumor recurrence, which are regulated by a large number of oncogenes and tumor suppressor genes (Liu et al., 2015; Cao et al., 2019). Next-generation sequencing technologies have made great progress recently, enabling scientists to gain profound insights into the molecular level of GBM pathophysiology (Aldape et al., 2015). As a consequence, many prospective diagnostic and prognostic biosignatures have been discovered, which enable a more distinct classification and a more precise outcome estimation of GBM. Nonetheless, given the dismal prognosis of GBM, a multiple-gene signature derived model is still urgently required to estimate the prognosis and treatment response more accurately for GBM patients.

The immune microenvironment has been chronicled to play a pivotal function in tumor biology (Hanahan and Weinberg, 2011), and cancer immunotherapy has been demonstrated to have a significant preclinical or clinical value to many patients with some sensitive types of cancer (Schumacher and Schreiber, 2015; Steven et al., 2016; Odunsi, 2017; Morrison et al., 2018; Christofi et al., 2019). Increasing research evidence supports the idea that although the brain constitutes an immunologically specific site, the immune microenvironment provides ample opportunities for immunotherapy of brain tumors (Lim et al., 2018). Many kinds of immunotherapy, including GBM vaccines, oncolytic viral therapies, immune-checkpoint suppressors, and chimeric antigen receptor T cell therapy, have been tested in clinical trials, but the results are not satisfactory. Tumors are insensitive to immunotherapy due to the immunosuppressive tumor microenvironment, defects in tumor antigen presentation, and characteristics of the physical microenvironment, including hypoxia and necrosis (Lim et al., 2018; Pombo Antunes et al., 2020). The precise mechanism of immune escape is unclear. Glioblastoma usually has a low mutational load and lower T cell invasion relative to other tumor types (Li et al., 2016). Thus, it is imperative to better comprehend the progress and mechanisms of the GBM immune microenvironment. Multiple recent studies have suggested that immune gene expression profile biosignatures may be used as a prediction for clinical outcomes in many cancers (Bremnes et al., 2016; Campbell et al., 2017; Öjlert et al., 2019). Li et al. (2017) created a personalized immune-related gene prognostic biosignature to improve the prognosis of individuals with NSCLC.

In a previous study, a prognostic immune-related gene signature with nine IRGs based on a total of 161 samples from the Cancer Genome Atlas database (TCGA) was generated (Liang et al., 2020), and the 9-IRG model was identified as an independent predictor in glioblastoma. These researchers established a crosstalk network between prognostic immune-related genes (IRGs) and transcription factors. Correlations between immune infiltration cells and risk score were also identified. However, the potential molecular mechanisms were not clarified in their study. Thus, it is necessary to elucidate the function of these genes in the risk score and poor survival outcomes.

Here, we generated a seven immune-linked gene biosignature to exhibit the connection between gene expression and GBM prognosis, and we verified this biosignature in the TCGA and GEO dataset. These data may provide a novel reference for the prognostic prediction of GBM. We also confirmed the relevance of the seven IRGs to immune checkpoints, immune cell infiltration, oncogenesis pathway, and drug sensitivity. As a result, we not only generated a predictive model for GBM prognosis but also indicated the potential function of these IRGs in the occurrence and development of glioblastoma.

MATERIALS AND METHODS

Data Sources and Preliminary Processing

The RNA-Seq data of 169 GBM samples and five normal brain samples, as well as the clinical data of these GBM patients, such as age, gender, molecular subtype, gene mutation status, survival time, and survival status, were obtained from the TCGA dataset[1]. Additionally, the GBM patients' microarray and clinical data were collected from independent datasets in the GEO database, including GSE74187 ($n = 60$) and GSE4412 ($n = 59$). These gene expression data were generated and annotated on GPL6480 or GPL97 platform. The immune-related gene set, including 2,498 genes, was downloaded from the ImmPort database. The RNA-Seq and microarray data were normalized using scale method, and the data were pre-processed through the following steps: (1) patients with unavailable clinical and/or survival information were removed, (2) only the expression profiles of IRGs were preserved, and (3) genes with exceeding low abundance were filtered out (the expression value was 0 in more than half of the samples, or the average expression value was less than 0.3 in the samples). Finally, 1,100 genes were used for univariate Cox regression analysis and LASSO analysis.

Differential Gene and Functional Enrichment Analysis

The expression analysis of 2,498 immune-linked genes was conducted to identify the differentially expressed IRGs by the limma R package [false discovery rate (FDR) < 0.05 and $\log_2$ $|$ fold change$|$ > 1] (Ritchie et al., 2015). We conducted functional enrichment analyses to identify potential molecular biomechanisms of the differentially expressed IRGs *via* GO analysis and KEGG pathways (Yang et al., 2018). GOplot package was used for illustrating the relationship between genes and enriched KEGG pathways. Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005) was employed to examine the signaling cascades in which the IRGs were enriched between the high- and low-risk subgroups.

Establishment of the Immune-Associated Gene Biosignature

The univariate Cox regression analysis was applied to investigate the association between each IRG and OS of patients based on the TCGA dataset. To build the immune-related risk model,

[1] https://tcga-data.nci.nih.gov/tcga/

the genes with p value < 0.01 were considered as candidate survival-associated IRGs. The LASSO regression model was used to determine the most significant survival-correlated IRGs. First, the GBM patients in TCGA dataset were randomly divided into training and internal validation cohorts at a 4:1 ratio, forming a training cohort (n = 134) and an internal validation cohort (n = 33). The LASSO regression was employed based on 10-fold cross-validation to minimize the risk of overfitting. LASSO tends to "shrink" the regression coefficients to zero as λ increases. The optimal λ that yielded minimum cross validation error in 10-fold cross validation was chosen. The risk score was calculated by using the sum of normalized expression weighted by the LASSO regression coefficients (Zhong et al., 2020):

$$\text{Risk score} = \text{EmRNA1}\times\text{CmRNA1}+\text{EmRNA2}\times\text{CmRNA2}+ \\ \text{EmRNAn}\times\text{CmRNAn}$$

where E designates the expression level of each gene; and C designates the lasso regression coefficient of each gene.

The patients were separated into low- and high-risk groups according to the median of the risk score. OS of the patients in the two groups was analyzed by the log-rank test with "survival" package in R. Receiver operating characteristic (ROC) curve and the corresponding area under the ROC curve (AUC) were calculated to evaluate the prognostic value of the risk score by using "ROC" package.

CIBERSORT and Assessment of Tumor-Infiltrating Immune Cells

CIBERSORT is a computational technique that predicts the cell type signature in mix tissues through gene expression levels (Newman et al., 2015). Cell types can be identified using RNA mixtures in nearly any tissue (Yang et al., 2019). For this study, we employed CIBERSORT to examine the 22 types of immune cells in tumor tissues and show the percentages of 22 sets of tumor-infiltrating immune cells (TIICs) with bar plots and a corheatmap.

Analysis of Immune Infiltration

To analyze the correlation between the risk signature and infiltrating levels of six immune cells, including B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells, Spearman's correlation was calculated and the strength of correlation for the absolute value of r was as follows: r between 0 and 0.3 indicates a weak correlation; r between 0.3 and 0.7 indicates a moderate correlation; r between 0.7 and 1.0 indicates a strong correlation (Akoglu, 2018).

Statistical Analysis

Boxplot was generated using the "ggplot2" package in R. Heat map was generated using the "pheatmap" package in R. A correlation analysis of the seven immune genes was performed using the R "corrplot" package in the Pearson's method. Circular plot was generated using the "circlize" package in R. Student's t test was used to compare data from subgroups. Pearson's correlation test was used to analyze the correlation between the IRGs signature and the expression of immune checkpoint genes. K-M survival curves were compared using log-rank test. All statistical analyses were conducted on R software (version 3.6.0). A p value of < 0.05 was considered to indicate significance. Other statistical methods were described throughout the study.

RESULTS

Identification of Differentially Expressed IRGs in GBM

The mRNA levels of 2,498 IRGs in GBM (n = 169) and normal brain tissues (n = 5) from TCGA were compared *via* the Wilcoxon rank-sum test. In total, 595 differentially expressed IRGs comprising 416 upregulated genes and 179 downregulated genes were identified (**Supplementary Table 1**). The volcano plot and heat map of differentially expressed IRGs are shown in **Figures 1A,B**.

Functional Characterization of DEIRGs

The gene functional enrichment assessment showed that immune responses were the most common. The most significant biological terms were "regulation of leukocyte activation," "plasma membrane protein complex" and "receptor ligand activity" among biological processes, cellular components, and molecular functions, respectively (**Figure 2A**). With regard to the KEGG cascades, most of signaling cascades were linked to immune reactions, and cytokine-cytokine receptor crosstalk was the most significantly enriched term (**Figure 2B**). For better visualization, two heatmaps of these values were plotted using the logFC, including one for GO terms (**Figure 2C**) and the other for KEGG pathways (**Figure 2D**). Some GO terms and KEGG cascades were linked to certain immune processes.

Identification of Prognostic Genes

The univariate Cox regression model was applied to select IRGs with the patient OS, and a total of 15 IRGs were discovered to be significantly associated with OS ($p < 0.01$). These genes were subjected to the LASSO regression analysis to calculate the correlation coefficients. The signature performed best when only seven genes were included (**Figures 3A,B**). For this analysis, we used LASSO regression to obtain the following seven optimal IRGs (risk genes) for incorporation into the prognostic risk model in TCGA training cohort (**Supplementary Figure 1**): Bone Morphogenetic Protein Receptor Type 1A (BMPR1A), Cathepsin B (CTSB), NFKB Inhibitor Zeta (NFKBIZ), TNF Superfamily Member 14 (TNFSF14), C-X-C Motif Chemokine Ligand 2 (CXCL2), Semaphorin-4F (SEMA4F), and Oncostatin M Receptor (OSMR). Among these genes, CTSB, NFKBIZ, TNFSF14, CXCL2, SEMA4F, and OSMR were characterized as high-risk genes (estimating a poor prognosis), whereas BMPR1A was identified as low-risk genes (functioning as a protective factor) with regard to the OS of patients (see detailed information in **Table 1**).

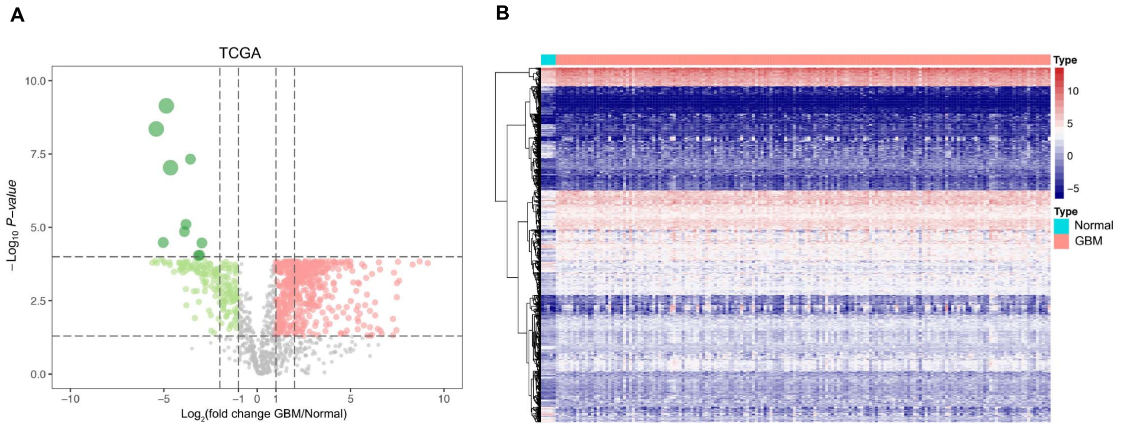


FIGURE 1 | Identification of differentially expressed IRGs between GBM and normal brain tissues. **(A)** Volcano plots showing the $\log_2$ (fold change) of mRNA in GBM compared to normal brain tissues, and the corresponding-$\log_{10}$ (*P* value) in TCGA datasets. Genes with adjusted P value below 0.05 and fold change above one (below –1) were marked with red (green) dots. **(B)** Heatmap of the differentially expressed IRGs in TCGA datasets.

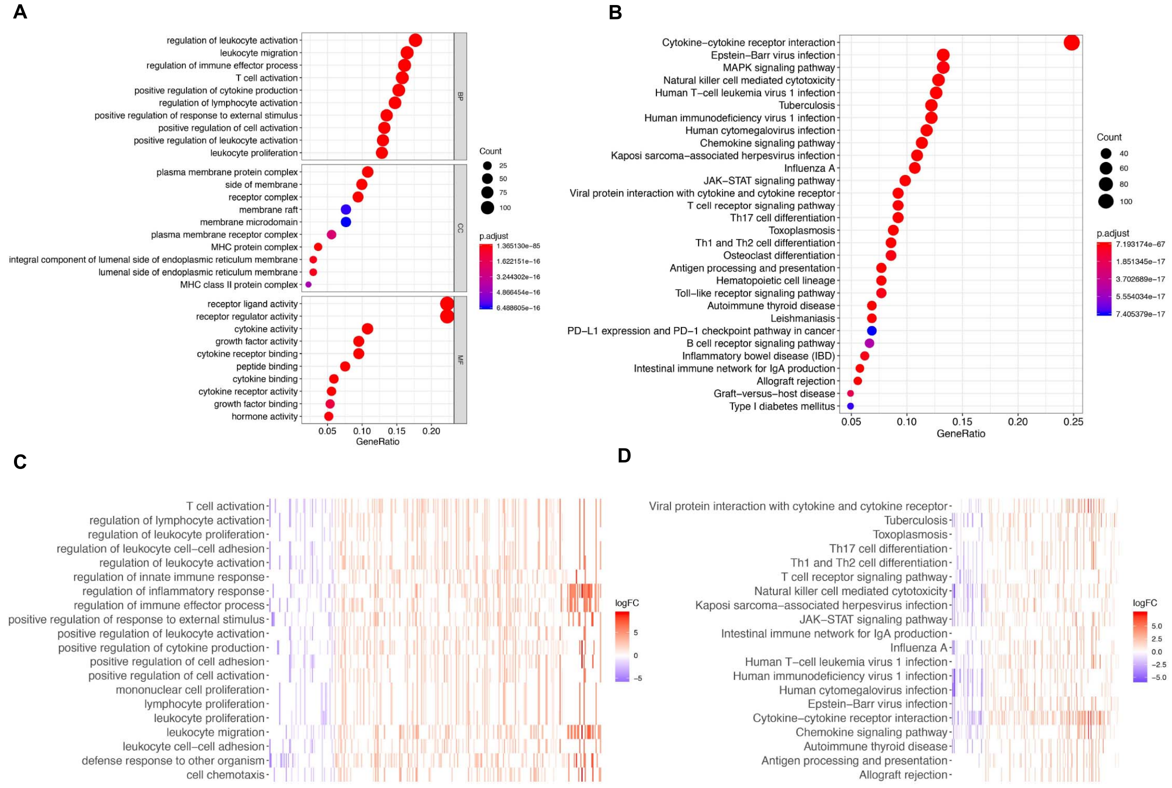


FIGURE 2 | GO terms and Enrichment of KEGG pathways for differentially expressed IRGs. **(A)** GO biological process analysis for the immune-related DEGs. **(B)** KEGG pathway enrichment analysis for the immune-related DEGs. **(C)** Heatmap of the GO terms by logFC. **(D)** Heatmap of the KEGG pathways by logFC.

Construction of a Seven-Gene Prognostic Biosignature

The LASSO regression analysis was used to screen the risk genes for estimating the prognosis of GBM individuals (Friedman et al., 2010; Simon et al., 2011). We utilized mRNA contents and predicted the regression coefficients of the risk genes to compute a risk score for each GBM individual. The prognostic estimation model was created, which incorporated seven immune-linked genes. The following formula was used for the calculation:

$$\text{Risk score} = (-0.194)\ \text{BMPR1A}+0.011\ \text{CTSB}+0.050\ \text{NFKBIZ} \\ +0.081\ \text{TNFSF14}+0.090\ \text{CXCL2} \\ +0.217\ \text{SEMA4F}+0.250\ \text{OSMR}$$

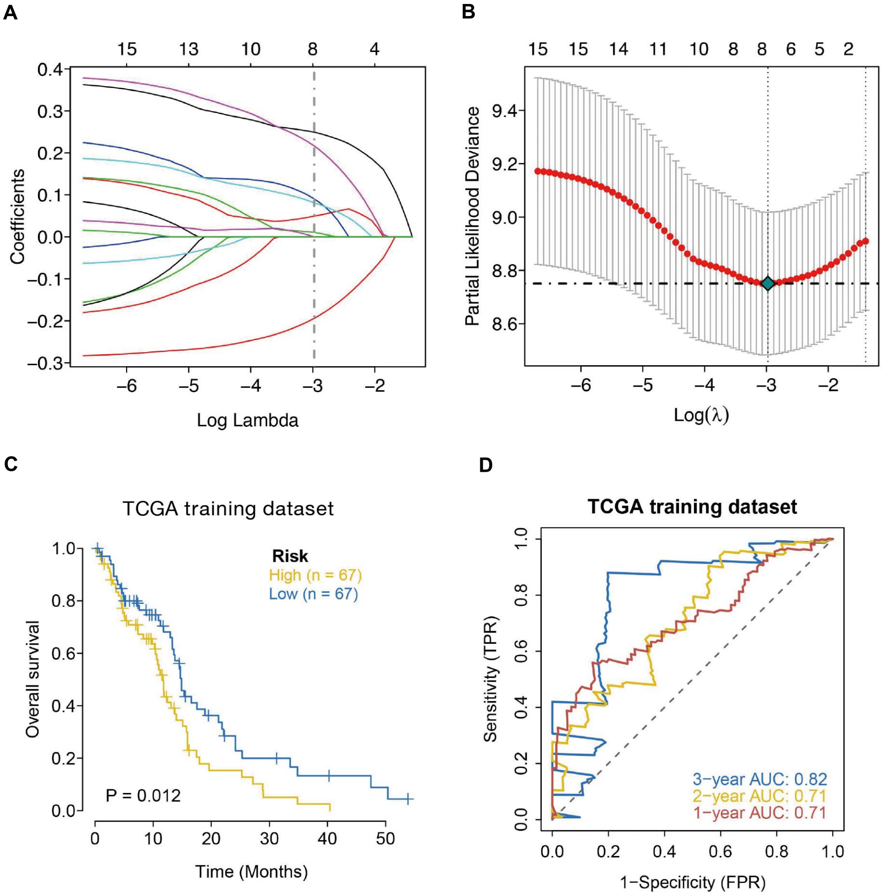


FIGURE 3 | Seven-immune-related gene signature prognostic risk model analysis of GBM patients in TCGA dataset. **(A)** LASSO coefficient profiles of the 15 IRGs in TCGA-GBM. **(B)** A coefficient profile plot was generated against the log (lambda) sequence. Selection of the optimal parameter (lambda) in the LASSO model for TCGA. **(C)** Kaplan–Meier survival curves for high-risk and low-risk groups. **(D)** ROC curves to examine the predictive accuracy of the model for OS at 1-, 2-, and 3-years.

According to the formula, we calculated the risk scores of each GBM individual and clustered them into low-risk and high-risk classes according to the median risk score. According to the log-rank test, the Kaplan–Meier curve revealed that the prognosis in the high-risk class was worse compared to the low-risk class in TCGA training cohort (p = 0.012) (**Figure 3C**). We employed the time-dependent ROC curves to explore the estimation accuracy of the model for OS in TCGA training cohort. The prognostic model area under the ROC values were 0.71 at 1-year, 0.71 at 2-year, and 0.82 at 3-year (**Figure 3D**). Suggesting our 7-gene model had a favorable efficiency in predicting prognosis.

Verification of the Immune-Linked Gene Biosignature

The prognostic value of the seven IRGs signature was further evaluated in three validation sets (TCGA internal validation set, GSE74187, and GSE4412 datasets). The risk score for each patient was calculated following the same formula. Patients in three validation sets were classified into high- and low-risk groups based on the median of the risk score. Survival analysis in the three validation sets confirmed a lower survival rate in the high-risk group (**Figures 4A–C**). The AUC of ROC curves for 1-, 2-, and 3-year survival rate in the validation dataset were 0.79, 0.91, and 0.93 (TCGA internal validation cohort) 0.64, 0.67, and

TABLE 1 | Risk genes in the prognostic risk model.

Gene	Coef	HR	Low. 95%CI	Upp. 95%CI	*p*-value
BMPR1A	−0.194	0.691	0.556	0.859	8.86E-4
CTSB	0.011	1.280	1.104	1.484	1.06E-3
NFKBIZ	0.050	1.442	1.200	1.731	9.10E-5
TNFSF14	0.081	1.320	1.138	1.532	2.58E-4
CXCL2	0.090	1.350	1.152	1.583	2.11E-4
SEMA4F	0.217	1.490	1.199	1.852	3.25E-3
OSMR	0.250	1.475	1.239	1.757	1.30E-5

7 prognostic immune-related genes screened out by the univariate Cox regression and LASSO Cox proportional hazards regression.

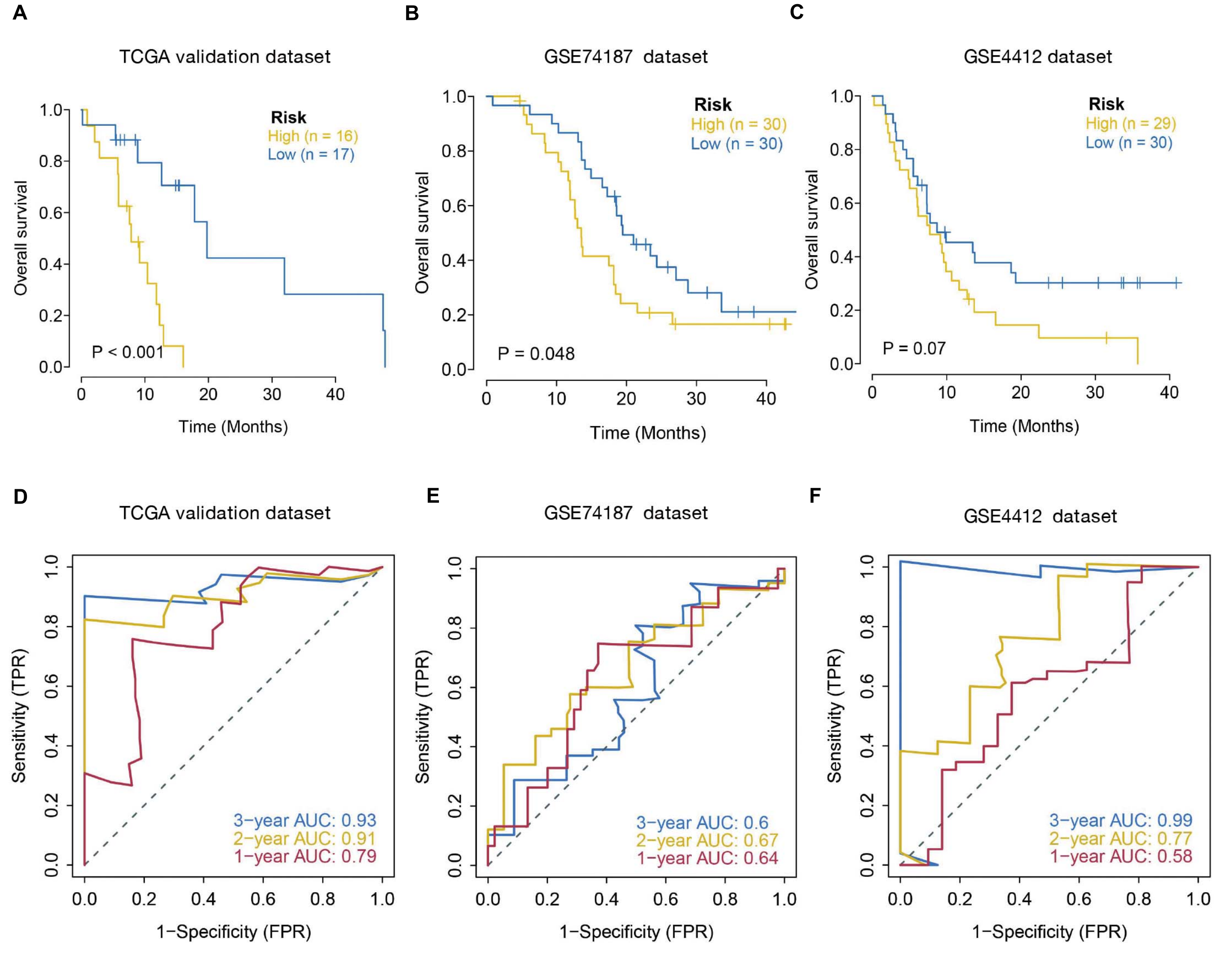


FIGURE 4 | Validation of seven-immune-related gene signature prognostic risk model of GBM patients in validation datasets. (A) Kaplan–Meier survival curves for high-risk and low-risk groups in TCGA internal validation dataset (*p* < 0.001). **(B)** Kaplan–Meier survival curves for high-risk and low-risk groups in GSE74187 dataset (*p* = 0.048). **(C)** Kaplan–Meier survival curves for high-risk and low-risk groups in GSE4412 dataset (*p* = 0.07). **(D–F)** ROC curves to examine the predictive accuracy of the model for OS at 1-, 2-, and 3- years in validation cohorts.

0.6 (GSE74187); 0.58, 0.77, and 0.99 (GSE4412) (**Figures 4D–F**). In summary, the prognosis model created according to the expression patterns of these seven prognosis-distinct immune-linked genes had high estimation accuracy and stability in identifying immune features. These data demonstrated that our prognostic risk model precisely estimates the prognosis of GBM individuals.

Relationship Between the Risk Score and Clinical Factors

The relationship between the seven IRGs signature and clinical factors, including age, gender, IDH1 mutation, 1p/19q mutation, and subtype was further investigated using data from the TCGA dataset. The results showed that a higher risk score was always associated with IDH1 mutation, 19q mutation, and subtype. No differences were observed between the risk score and age, gender, or 1p mutation (**Supplementary Figure 2**).

Functional Annotations and Signaling Pathway Enrichment in High- and Low-Risk Score Groups

Because the monitoring of disease outcome is imperative for clinical management, we aimed to identify molecular biosignatures that could be utilized as viable prognostic indicators. Functional gene annotation and KEGG enrichment analyses focused on the above mentioned seven prognosis-distinct immune-linked genes were conducted (Yu et al., 2012). We demonstrated that these survival-linked IRGs were most abundant in gene ontology (GO) terms linked to "cell adhesion mediated by integrin," "granulocyte migration," "platelet degranulation," "regulation of leukocyte adhesion to vascular endothelial cell," "rna capping" and "transcription preinitiation complex assembly" (**Figure 5A**). Gene set enrichment analysis (GSEA) was performed to identify the prospective cascades that differentiated the high- or low-risk

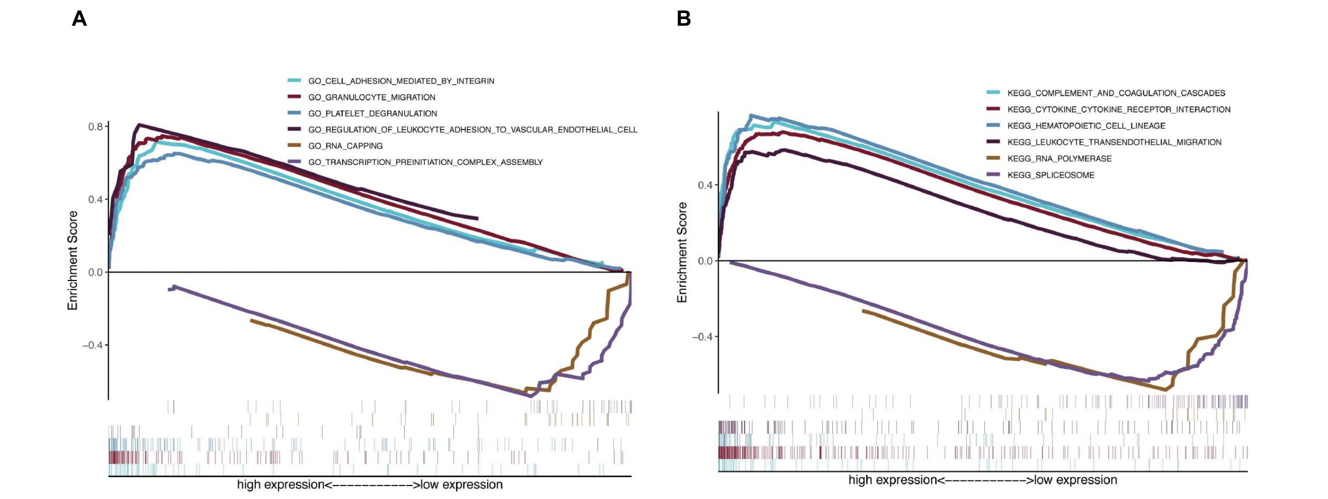


FIGURE 5 | Functional gene annotations and KEGG enrichment analysis between high and low risk groups. **(A)** ClusterProfiler was selected for functional gene annotations. **(B)** GSEA analysis was performed to identify the potential pathways differentiate the high and low risk groups.

groups. The following cascades were significantly enriched: "complement and coagulation cascades," "cytokine cytokine receptor interaction," "hematopoietic cell lineage," "leukocyte transendothelial migration," "rna polymerase," and "spliceosome" (**Figure 5B**). These results suggested that the prognosis-specific immune-related gene risk score using the seven IRGs may affect these cascades and estimate the survival of GBM patients.

Correlation Between the Risk Score and Immune Response

To better comprehend the connection between the risk score and immune response, we calculated the association between the risk score and the expression levels of core immune checkpoints in GBM, such as CD28, TIM-3, B7-H3, PD-1, B7-H4, CD40, LAG3, and PD-L1. Interestingly, the Circos plot (Gu et al., 2014) showed that the risk score was strongly linked to expression levels of B7-H3, CD40, and PD-L1 in TCGA cohorts (**Figure 6A**).

Distribution of Immune Invasion in Glioblastoma

We first assessed immune invasion in glioma tissue in 22 subpopulations of immune cells by employing the CIBERSORT algorithm. In **Figure 6B**, the percentage of immune cells in each GBM sample is shown in different colors, and the lengths of the bars indicate the immune cell population levels. We then speculated that the divergence in TIIC proportions may function as a critical feature of individual differences and possess prognostic significance. Based on the chart, we established that glioma tissues had comparatively high proportions of M1, M0, and M2 macrophages as well as monocytes, which were responsible for approximately 70% of the 22 subpopulations of immune cells. In contrast, B cell and neutrophil proportions were comparatively low, and they were responsible for approximately 10% of the immune cell subpopulations (**Figure 6B**). Proportions of different types of immune cells subsets were weakly and then moderately correlated (**Figure 6C**). Populations with a negative correlation consisted of monocytes/M2 macrophages (Pearson's correlation = −0.41) and resting NK cells/activated NK cells (Pearson's correlation = −0.43). Given the important role of these hub immune genes, the genetic variations of five of them with a mutation rate ≥ 5% were further explored (**Supplementary Figure 3**).

Prognostic Model Associates With Immune Invasion in GBM

Clinical studies on immunotherapy have verified that tumor-invading lymphocytes in the tumor microenvironment possess an estimation significance for prognosis and treatment using immunotherapy in some solid tumors (Bremnes et al., 2016; Lee et al., 2016; Badalamenti et al., 2019). Given that our risk score was centered on seven immune-linked genes, we investigated whether it was linked to the invading levels of six immune cell types in the TCGA GBM cohort acquired from TIMER. We examined the link between the expression levels of seven immune-linked genes and the invading contents of six immune cell types. The findings demonstrated that the expression of these seven genes exhibited remarkably positive correlation with immune cell invasion. The expressions of CTSB, NFKBIZ, CXCL2, and OSMR were all correlated with the invading levels of dendritic cells (**Supplementary Figure 4**). To better understand the impact of the seven IRGs signature on the infiltration of immune cells, the relevance of the risk score and six immune cells was investigated. Results indicated that the risk score was positively related to neutrophil cells (r = 0.188), dendritic cells (r = 0.404), and CD4+ T cells (r = 0.169) (**Supplementary Figure 5**). Collectively, these data indicated that our model system is partially linked to the invading level of immune cells in the tumor microenvironment of GBM. Particularly, BMPR1A was significantly correlated with the infiltrating levels of CD4+ T cells, macrophages, and dendritic cells. TNFSF14 and OSMR

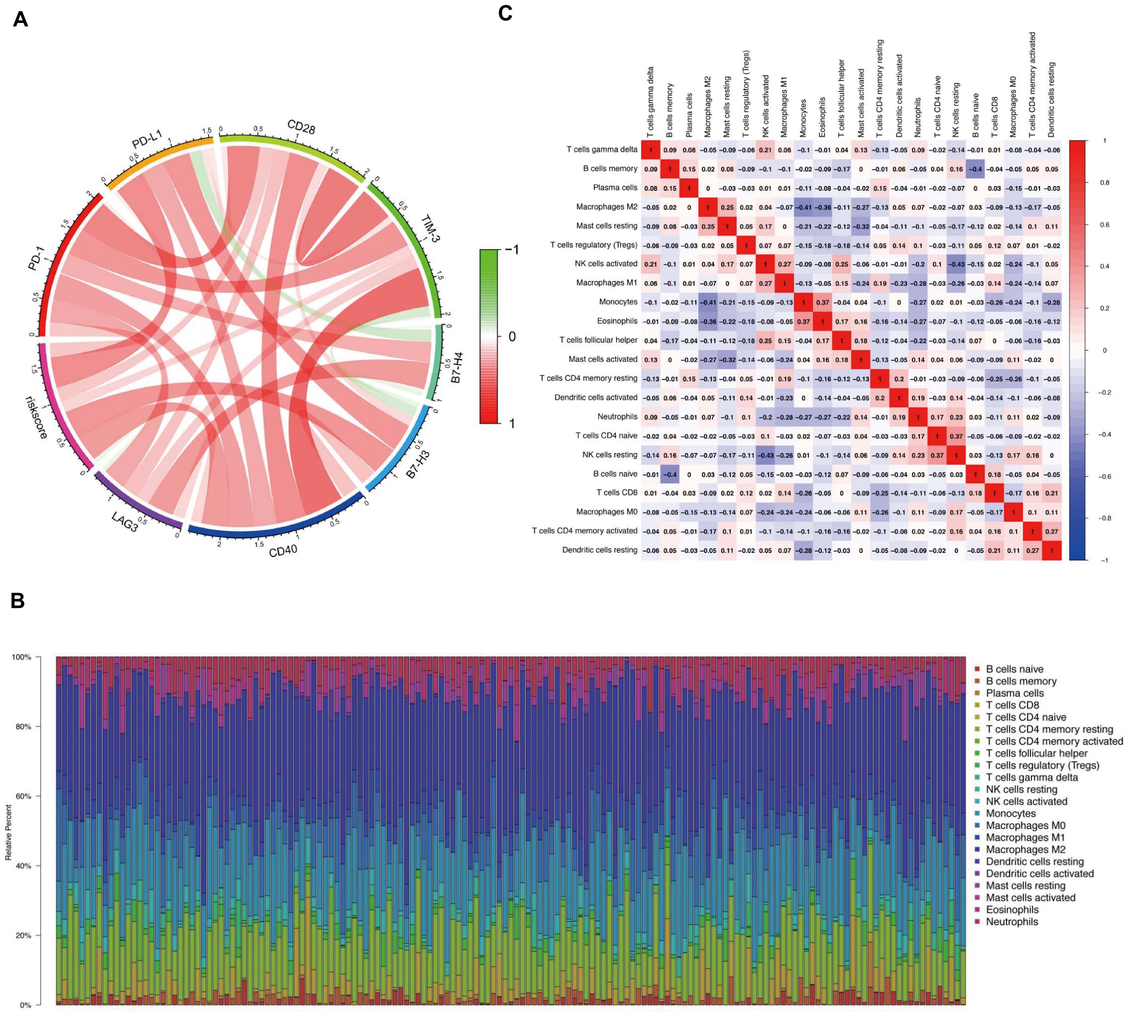


FIGURE 6 | Correlation between the risk score and immune response and the distribution of immune infiltration in GBM. **(A)** Circos plot shows the relationship between the risk score and the expression levels of some important immune checkpoints in GBM. **(B)** The proportions of immune cells in each GBM sample are indicated with different colors, and the lengths of the bars in the bar chart indicate the levels of the immune cell populations. **(C)** Correlation matrix for all 22 immune cell proportions. Some immune cells were negatively related, represented in blue, and others were positively related, represented in red. The darker the color, the higher the correlation.

were significantly correlated with the invading levels of CD4+ T cells and dendritic cells. CXCL2 was significantly associated with the invading levels of dendritic cells (**Figure 7**).

Effects of Prognosis-Specific Immune-Related Genes on Oncogenic Pathways

To further elucidate the molecular mechanisms for prognosis-specific IRGs participating in tumorigenesis, we explored the link between the expression of individual genes and activation or repression of 10 core signaling cascades based on a pathway score computed from the sum of the relative protein contents for all positive modulatory constituents less that of all negative modulatory constituents (Akbani et al., 2014). Our data demonstrated that seven genes were highly correlated to the activation or suppression of numerous oncogenic cascades (**Supplementary Figure 6**). For example, CTSB was highly correlated with the repression of DNA damage response and AR hormone, as well as the activation of apoptosis and EMT signaling pathways. CXCL2 was associated with the inhibition of cell cycle, DNA damage response, and AR hormone, as well as activation of apoptosis, EMT, and RAS/MAPK signaling pathways. These results suggested that prognosis-specific IRGs are linked to alterations of diverse oncogenic cascades.

Hub Gene Drug Sensitivity

GSCALite constitutes a web-based analysis portal for gene set cancer analysis (Liu C. J. et al., 2018), based on which the drug sensitivity of the hub genes was analyzed to provide support on drug-targeted therapy (**Supplementary Figure 7**). Low NFKBIZ

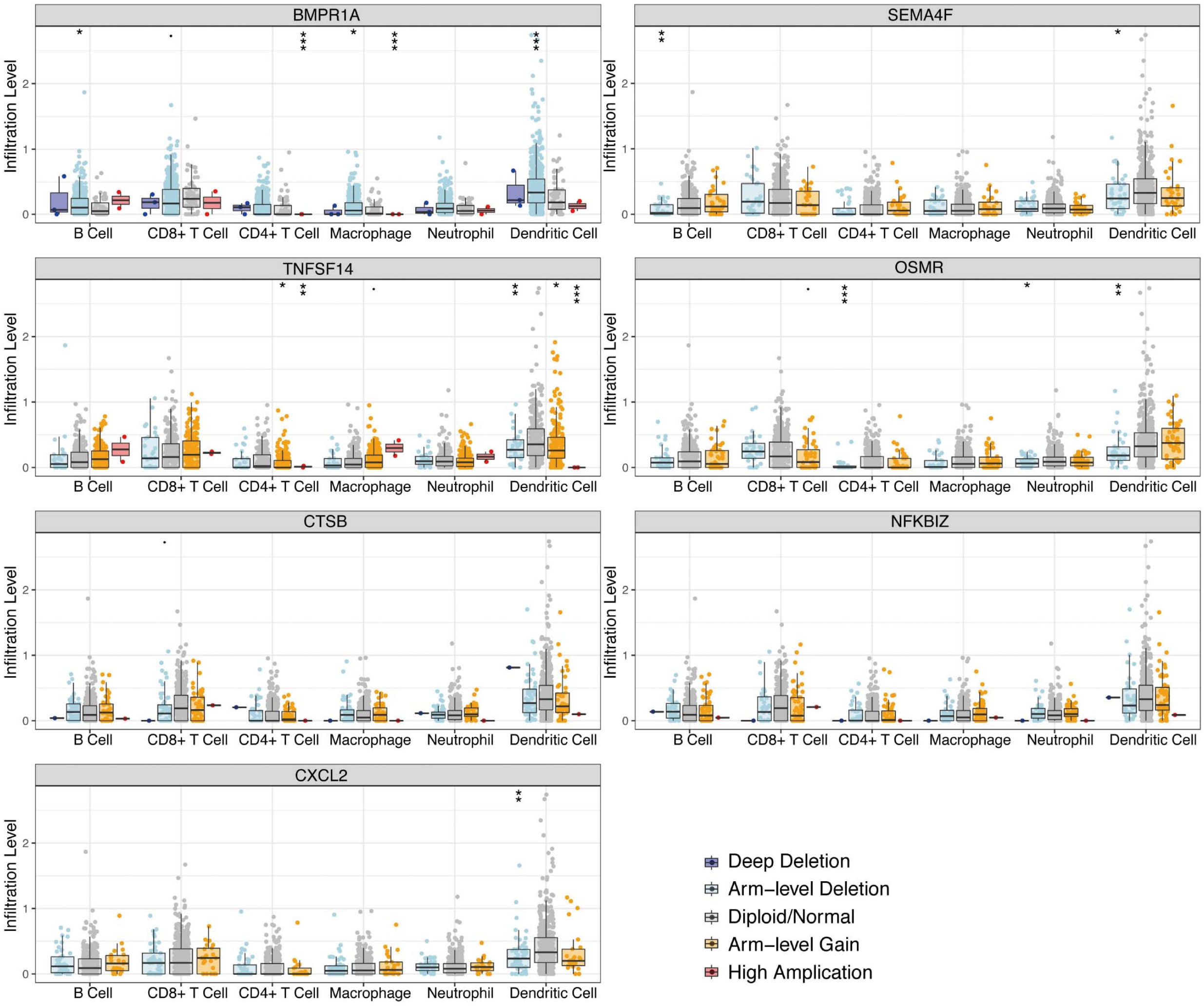


FIGURE 7 | Correlations of seven immune-related gene copy member with immune infiltration level in GBM. These seven-immune-related gene CNV affects the infiltrating levels of different immune cells in GBM. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

level is resistant to 11 drugs or small molecules, low BMPR1A level is resistant to seven drugs or small molecules, low SEMA4F level is resistant to 16 drugs or small molecules, and low levels of OSMR, CXCL2, and CTSB are resistant to more than 32 drugs or small molecules.

DISCUSSION

Glioblastoma is a fatal human cancer. Despite of the years of research focused on GBM biology and the numerous clinical trials to evaluate new treatments, the prognosis of individuals with glioblastoma remains dismal (Thakkar et al., 2014). Patients diagnosed with GBM undergo treatments, including neurosurgery, radiotherapy, and chemotherapy, with unsatisfactory survival.

There has been great advancement in the comprehension of the genetic and molecular underpinnings of glioblastoma with the emergence and progression of microarray technology and sequencing technology. The IDH1 mutant was found in an integrated genomic analysis in 2008 (Parsons et al., 2008). Many studies have been performed in recent years and suggest that mutated IDH1 participates in the pathogenesis of glioma. According to the WHO categorization of central nervous system tumors, glioblastoma is divided into IDH-mutant and IDH-wildtype subtypes (Louis et al., 2016). This categorization is based entirely on histological features. There are many specific genetic changes in glioblastoma cases, and the most frequently mutated or deregulated gene is epidermal growth factor receptor (EGFR), which is amplified in approximately 60% of glioblastomas (Huang et al., 2007). Many deregulations with certain pathways, such as PI3K, P53, and RB, have also been identified. Overall, these studies show the prospect of the gene signature in tumor diagnosis and prognosis, and they provide new evidence for tumor biology. With the progression of bioinformatics and open access of high-throughput data, researchers have studied multiple gene prognostic signatures for GBM, which result in more accuracy than single gene prognostic signatures (Colman et al., 2010; Yin et al., 2019).

The CNS has been considered as an immune-favored system based on the initial experimental data documented more than 50 years ago (Medawar, 1948; Billingham et al., 1954), but many findings have suggested that the immune microenvironment provides sufficient opportunities to treat brain tumors with immunotherapy even though the brain is an immunologically distinct region (Schiffer et al., 2017). Scientists have great interest in utilizing immunotherapy to treat glioblastoma because it has shown considerable improvements in the management of numerous solid tumors, including melanoma, renal cell carcinoma, and NSCLC. There are many ongoing clinical trials for immunotherapy, but the results are not satisfactory. Thus, we need more knowledge about the GBM immune microenvironment.

Herein, we constructed a robust seven immune-linked gene biosignature for risk stratification in glioblastoma patients. In contrast to a previous studies (Liang et al., 2020), we used univariate Cox regression analysis and LASSO regression assessment to classify genes as independent prognostic indicators. Among them, CTSB, NFKBIZ, TNFSF14, CXCL2, SEMA4F, and OSMR were characterized as high-risk genes, whereas BMPR1A was identified as low-risk gene.

The protease cathepsin B (CTSB) has been identified to highly express in cancer (Mijanovic et al., 2019), and associate with poor prognosis of a variety of cancers, including breast cancer, pancreatic cancer, and lung squamous cell carcinoma, which could be used as an independent predictor of these tumors (Gong et al., 2013; Zhang et al., 2014). It was previously found that the absence of CTSB delays the growth and invasion of pancreatic neuroendocrine tumors (Gocheva et al., 2006). Here, we identified CTSB as a risk pattern based on our risk model, which is in consistent with previous studies. NFKBIZ mutation is associated with ulcerative colitis, and the repeated inflammation and repair are closely related to the occurrence of colorectal cancer (Kakiuchi et al., 2020). Thus, chronic inflammation might be related to GBM. TNFSF14 is also known as LIGHT, which has been studied at preclinical level for more than 10 years and has shown the prospect of strengthening cancer immunotherapy (Skeate et al., 2020). CXCL2 can promote the recruitment of MDSC and is associated with the prognosis of bladder cancer (Zhang et al., 2017). SEMA4F is expressed in adults and related to the neural guidance of embryos. It can induce neurogenesis in prostate cancer, thus promoting cancer growth and migration (Ding et al., 2013). The cytokine receptor for oncostatin M (OSMR) regulates self-renewing brain tumor stem cells and promotes the resistance of GBM to ionizing radiation (Sharanek et al., 2020). In breast cancer, BMPR1-knockdown can inhibit RANKL production through p38 pathway, thereby inhibiting breast cancer-induced osteolysis (Liu Y. et al., 2018). Above all, the above mentioned seven genes play important roles in the occurrence and development of tumors.

We next created a landscape of 22 subtypes of immune cells and acquired the status of immune infiltration in the GBM microenvironment. Our results were similar to those of previous studies (Lu et al., 2019; Liang et al., 2020). Furthermore, we analyzed the relationship between the expression levels of seven immune-linked genes and the invading levels of six immune cells. The data demonstrated that the expression of these seven genes exhibited positive correlation with immune cell invasion (**Supplementary Figure 4**). All these findings indicated that our prognostic model may aid in understanding the immune status of glioblastoma patients. We also generated a circo plot to show the relationship between the risk score and expression levels of core immune checkpoints in GBM. This study may provide new targets or effective biomarkers for glioblastoma immunotherapy.

In summary, the immunotherapy of GBM patients should be individualized to obtain a better curative effect. Our study provides a prognosis prediction based on IRGs, which may reflect the immune status of GBM patients. However, our study had limitations as our study was based on databases and bioinformatics analyses. Immunohistochemistry, flow cytometry, and RT-PCR should be used to verify our research results.

CONCLUSION

In our study, IRGs were identified to generate a prediction model of glioblastoma patient prognosis. We also explored the connection between these genes and the immune cells and immune checkpoints. Further research on these genes may provide new insights in GBM biology and promote immunotherapy.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LH and ZH performed all experiments, prepared figures, and drafted the manuscript. LH, ZH, XC, SW, and YF participated in data analysis and interpretation of results. LH and ZL designed the study and participated in data analysis. All authors have read and approved the manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (Grant Nos. 81772678, 81802755, and U20A20383), the China Postdoctoral Science Foundation (Grant No. 2018M630372), and Heilongjiang Postdoctoral Fund (Grant No. LBH-Z17166).

ACKNOWLEDGMENTS

We thank TopEdit (https://topeditsci.com/) for its linguistic assistance during the preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.638458/full#supplementary-material

REFERENCES

Akbani, R., Ng, P. K., Werner, H. M., Shahmoradgoli, M., Zhang, F., Ju, Z., et al. (2014). A pan-cancer proteomic perspective on the cancer genome Atlas. *Nat. Commun.* 5:3887.

Akoglu, H. (2018). User's guide to correlation coefficients. *Turk. J. Emerg. Med.* 18, 91–93. doi: 10.1016/j.tjem.2018.08.001

Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G., and von Deimling, A. (2015). Glioblastoma: pathology, molecular mechanisms and markers. *Acta Neuropathol.* 129, 829–848.

Badalamenti, G., Fanale, D., Incorvaia, L., Barraco, N., Listi, A., Maragliano, R., et al. (2019). Role of tumor-infiltrating lymphocytes in patients with solid tumors: can a drop dig a stone? *Cell. Immunol.* 343:103753.

Billingham, R. E., Brent, L., Medawar, P. B., and Sparrow, E. M. (1954). Quantitative studies on tissue transplantation immunity. I. The survival times of skin homografts exchanged between members of different inbred strains of mice. *Proc. R. Soc. Lond. B Biol. Sci.* 143, 43–58. doi: 10.1098/rspb.1954.0053

Bremnes, R. M., Busund, L. T., Kilvaer, T. L., Andersen, S., Richardsen, E., Paulsen, E. E., et al. (2016). The role of tumor-infiltrating lymphocytes in development, progression, and prognosis of non-small cell lung cancer. *J. Thorac Oncol.* 11, 789–800. doi: 10.1016/j.jtho.2016.01.015

Campbell, M. J., Baehner, F., O'Meara, T., Ojukwu, E., Han, B., Mukhtar, R., et al. (2017). Characterizing the immune microenvironment in high-risk ductal carcinoma in situ of the breast. *Breast Cancer Res. Treat.* 161, 17–28. doi: 10.1007/s10549-016-4036-0

Cao, M., Cai, J., Yuan, Y., Shi, Y., Wu, H., Liu, Q., et al. (2019). A four-gene signature-derived risk score for glioblastoma: prospects for prognostic and response predictive analyses. *Cancer Biol. Med.* 16, 595–605. doi: 10.20892/j.issn.2095-3941.2018.0277

Christofi, T., Baritaki, S., Falzone, L., Libra, M., and Zaravinos, A. (2019). Current perspectives in cancer immunotherapy. *Cancers* 11:1472. doi: 10.3390/cancers11101472

Colman, H., Zhang, L., Sulman, E. P., McDonald, J. M., Shooshtari, N. L., Rivera, A., et al. (2010). A multigene predictor of outcome in glioblastoma. *Neuro Oncol.* 12, 49–57. doi: 10.1093/neuonc/nop007

Ding, Y., He, D., Florentin, D., Frolov, A., Hilsenbeck, S., Ittmann, M., et al. (2013). Semaphorin 4F as a critical regulator of neuroepithelial interactions and a biomarker of aggressive prostate cancer. *Clin. Cancer Res.* 19, 6101–6111.

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

Gocheva, V., Zeng, W., Ke, D., Klimstra, D., Reinheckel, T., Peters, C., et al. (2006). Distinct roles for cysteine cathepsin genes in multistage tumorigenesis. *Genes Dev.* 20, 543–556.

Gong, F., Peng, X., Luo, C., Shen, G., Zhao, C., Zou, L., et al. (2013). Cathepsin B as a potential prognostic and therapeutic marker for human lung squamous cell carcinoma. *Mol. Cancer* 12:125.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812. doi: 10.1093/bioinformatics/btu393

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

Huang, P. H., Mukasa, A., Bonavia, R., Flynn, R. A., Brewer, Z. E., Cavenee, W. K., et al. (2007). Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proc. Natl. Acad. Sci. U.S.A.* 104, 12867–12872. doi: 10.1073/pnas.0705158104

Kakiuchi, N., Yoshida, K., Uchino, M., Kihara, T., Akaki, K., Inoue, Y., et al. (2020). Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis. *Nature* 577, 260–265.

Lee, N., Zakka, L. R., Mihm, M. C. Jr., and Schatton, T. (2016). Tumour-infiltrating lymphocytes in melanoma prognosis and cancer immunotherapy. *Pathology* 48, 177–187. doi: 10.1016/j.pathol.2015.12.006

Li, B., Cui, Y., Diehn, M., and Li, R. (2017). Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non-small cell lung cancer. *JAMA Oncol.* 3, 1529–1537. doi: 10.1001/jamaoncol.2017.1609

Li, B., Severson, E., Pignon, J. C., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 17, 174. doi: 10.1186/s13059-016-1028-7

Liang, P., Chai, Y., Zhao, H., and Wang, G. (2020). Predictive analyses of prognostic-related immune genes and immune infiltrates for glioblastoma. *Diagnostics (Basel)* 10:177. doi: 10.3390/diagnostics10030177

Lim, M., Xia, Y., Bettegowda, C., and Weller, M. (2018). Current state of immunotherapy for glioblastoma. *Nat. Rev. Clin. Oncol.* 15, 422–442. doi: 10.1038/s41571-018-0003-5

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.-F., Hu, Y.-Y., Jin, T., Xu, K., Wang, S.-H., Du, G.-Z., et al. (2015). Matrix Metalloproteinase-9/Neutrophil gelatinase-associated lipocalin complex activity in human glioma samples predicts tumor presence and clinical prognosis. *Dis. Markers* 2015:138974. doi: 10.1155/2015/138974

Liu, Y., Zhang, R. X., Yuan, W., Chen, H. Q., Tian, D. D., Li, H., et al. (2018). Knockdown of bone morphogenetic proteins Type 1a receptor (BMPR1a) in breast cancer cells protects bone from breast cancer-induced osteolysis by suppressing RANKL expression. *Cell. Physiol. Biochem.* 45, 1759–1771.

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Lu, J., Li, H., Chen, Z., Fan, L., Feng, S., Cai, X., et al. (2019). Identification of 3 subpopulations of tumor-infiltrating immune cells for malignant transformation of low-grade glioma. *Cancer Cell Int.* 19:265. doi: 10.1186/s12935-019-0972-1

Medawar, P. B. (1948). Immunity to homologous grafted skin; the fate of skin homografts transplanted to the brain, to subcutaneous tissue, and to the anterior chamber of the eye. *Br. J. Exp. Pathol.* 29, 58–69.

Mijanovic, O., Brankovic, A., Panin, A. N., Savchuk, S., Timashev, P., Ulasov, I., et al. (2019). Cathepsin B: a sellsword of cancer progression. *Cancer Lett.* 449, 207–214. doi: 10.1016/j.canlet.2019.02.035

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

Morrison, A. H., Byrne, K. T., and Vonderheide, R. H. (2018). Immunotherapy and prevention of pancreatic cancer. *Trends Cancer* 4, 418–428. doi: 10.1016/j.trecan.2018.04.001

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.

Odunsi, K. (2017). Immunotherapy in ovarian cancer. *Ann. Oncol.* 28(Suppl. 8), viii1–viii7.

Öjlert, Å. K., Halvorsen, A. R., Nebdal, D., Lund-Iversen, M., Solberg, S., Brustugun, O. T., et al. (2019). The immune microenvironment in non-small cell lung cancer is predictive of prognosis after surgery. *Mol. Oncol.* 13, 1166–1179.

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., Angenendt, P., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science* 321, 1807–1812.

Pombo Antunes, A. R., Scheyltjens, I., Duerinck, J., Neyns, B., Movahedi, K., and Van Ginderachter, J. A. (2020). Understanding the glioblastoma immune microenvironment as basis for the development of new immunotherapeutic strategies. *Elife* 9:e52176.

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.

Schiffer, D., Mellai, M., Bovio, E., and Annovazzi, L. (2017). The neuropathological basis to the functional role of microglia/macrophages in gliomas. *Neurol. Sci.* 38, 1571–1577.

Schumacher, T. N., and Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science* 348, 69–74.

Sharanek, A., Burban, A., Laaper, M., Heckel, E., Joyal, J. S., Soleimani, V. D., et al. (2020). OSMR controls glioma stem cell respiration and confers resistance of glioblastoma to ionizing radiation. *Nat. Commun.* 11:4116.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1–13.

Skeate, J. G., Otsmaa, M. E., Prins, R., Fernandez, D. J., Da Silva, D. M., and Kast, W. M. (2020). TNFSF14: LIGHTing the way for effective cancer immunotherapy. *Front. Immun.* 11:922. doi: 10.3389/fimmu.2020.00922

Steven, A., Fisher, S. A., and Robinson, B. W. (2016). Immunotherapy for lung cancer. *Respirology* 21, 821–833.

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

Thakkar, J. P., Dolecek, T. A., Horbinski, C., Ostrom, Q. T., Lightner, D. D., Barnholtz-Sloan, J. S., et al. (2014). Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiol. Biomarkers Prev.* 23, 1985–1996.

Yan, W., Zhang, W., You, G., Zhang, J., Han, L., Bao, Z., et al. (2012). Molecular classification of gliomas based on whole genome gene expression: a systematic report of 225 samples from the Chinese Glioma cooperative group. *Neuro Oncol.* 14, 1432–1440.

Yang, L.-J., Zhou, C.-F., and Lin, Z.-X. (2014). Temozolomide and radiotherapy for newly diagnosed glioblastoma multiforme: a systematic review. *Cancer Invest.* 32, 31–36.

Yang, X., Deng, Y., He, R. Q., Li, X. J., Ma, J., Chen, G., et al. (2018). Upregulation of HOXA11 during the progression of lung adenocarcinoma detected via multiple approaches. *Int. J. Mol. Med.* 42, 2650–2664.

Yang, X., Shi, Y., Li, M., Lu, T., Xi, J., Lin, Z., et al. (2019). Identification and validation of an immune cell infiltrating score predicting survival in patients with lung adenocarcinoma. *J. Transl. Med.* 17:217.

Yin, W., Tang, G., Zhou, Q., Cao, Y., Li, H., Fu, X., et al. (2019). Expression profile analysis identifies a novel five-gene signature to improve prognosis prediction of glioblastoma. *Front. Genet.* 10:419. doi: 10.3389/fgene.2019.00419

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

Zhang, H., Ye, Y. L., Li, M. X., Ye, S. B., Huang, W. R., Cai, T. T., et al. (2017). CXCL2/MIF-CXCR2 signaling promotes the recruitment of myeloid-derived suppressor cells and is correlated with prognosis in bladder cancer. *Oncogene* 36, 2095–2104.

Zhang, W., Wang, S., Wang, Q., Yang, Z., Pan, Z., and Li, L. (2014). Overexpression of cysteine cathepsin L is a marker of invasion and metastasis in ovarian cancer. *Oncol. Rep.* 31, 1334–1342.

Zhong, S., Chen, H., Yang, S., Feng, J., and Zhou, S. (2020). Identification and validation of prognostic signature for breast cancer based on genes potentially involved in autophagy. *PeerJ* 8:e9621. doi: 10.7717/peerj.9621

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.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 23 February 2021
doi: 10.3389/fgene.2021.639254

Identification and Validation of a Novel Immune-Related Four-lncRNA Signature for Lung Adenocarcinoma

Jixin Wang[1]†, Xiangjun Yin[2], Yin-Qiang Zhang[3] and Xuming Ji[2]**

[1] Zhejiang University-University of Edinburgh Institute, Zhejiang University, Zhejiang, China, [2] School of Basic Medical Science, Zhejiang Chinese Medical University, Zhejiang, China, [3] Department of Hepatic Diseases, Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing, China

Lung adenocarcinoma (LUAD) is a major subtype of lung cancer, the prognosis of patients with which is associated with both lncRNAs and cancer immunity. In this study, we collected gene expression data of 585 LUAD patients from The Cancer Genome Atlas (TCGA) database and 605 subjects from the Gene Expression Omnibus (GEO) database. LUAD patients were divided into high and low immune-cell-infiltrated groups according to the single sample gene set enrichment analysis (ssGSEA) algorithm to identify differentially expressed genes (DEGs). Based on the 49 immune-related DE lncRNAs, a four-lncRNA prognostic signature was constructed by applying least absolute shrinkage and selection operator (LASSO) regression, univariate Cox regression, and stepwise multivariate Cox regression in sequence. Kaplan–Meier curve, ROC analysis, and the testing GEO datasets verified the effectiveness of the signature in predicting overall survival (OS). Univariate Cox regression and multivariate Cox regression suggested that the signature was an independent prognostic factor. The correlation analysis revealed that the infiltration immune cell subtypes were related to these lncRNAs.

OPEN ACCESS

Edited by:
Shengli Li,
Shanghai Jiao Tong University, China

Reviewed by:
Rong-Rong He,
Jinan University, China
Zhaoyuan Fang,
Chinese Academy of Sciences, China

****Correspondence:***
Xuming Ji
jixuming724@163.com
Yin-Qiang Zhang
zhyq992@163.com

†ORCID:

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

***Received:** 08 December 2020*
***Accepted:** 18 January 2021*
***Published:** 23 February 2021*

Citation:
Wang J, Yin X, Zhang Y-Q and Ji X (2021) Identification and Validation of a Novel Immune-Related Four-lncRNA Signature for Lung Adenocarcinoma. Front. Genet. 12:639254. doi: 10.3389/fgene.2021.639254

Keywords: lung adenocarcinoma, lncRNA, survival analysis, immune infiltrate, GSEA

INTRODUCTION

Lung cancer is one of the most common types of malignancy that is a leading cause of death worldwide. The frequency of lung adenocarcinoma (LUAD) has exceeded lung squamous cell carcinoma (LUSC), which makes LUAD the most common histological subtype of primary lung cancer (Lortet-Tieulent et al., 2014). The high mortality is mainly because lung cancer is typically diagnosed at an advanced stage. Patients who have mutations in epidermal growth factor receptor (EGFR) are recommended to receive molecule-targeted therapy by administrating anti-EGFR inhibitors (Duffy and O'Byrne, 2018). For those who do not have specific mutations, immunotherapies targeting inhibitory receptors have recently emerged as an effective therapy for advanced cancer. The most studied way is using antibodies to block the programmed cell death-1 (PD-1)/programmed cell death ligand-1 (PD-L1) pathway, an immune checkpoint that is exploited by tumor cells (Sacher and Gandhi, 2016). These anti-PD-1/PD-L1 treatments do not need specific mutations such as EGFR, KRAS, or ALK (Brody et al., 2017) and are available for more patients.

A large proportion of tumor cells are immune infiltrating cells (Yu et al., 2018). Tumor immune cell infiltration is vital for the effect of immunotherapy and therefore the prognosis of LUAD patients because the tumor-specific antigens need to be recognized by the antigen–antibody complementary determining region in immune cells (Sela-Culang et al., 2014; Jiang et al., 2017; Liu et al., 2018). Higher CD8-T cell infiltration seems to better respond to anti-PD-1/PD-L1 administration (Pagès et al., 2016).

Long non-coding RNA (lncRNA), a type of non-coding RNA with a length longer than 200 nucleotides, accounts for a large proportion of the human genome. Studies have suggested that lncRNAs regulate gene expression and are associated with many biological processes such as development (Quinn and Chang, 2016). For example, PTTG3P up-regulation has been discovered to promote cell viability and contribute to the poor survival of LUAD patients (Shih et al., 2020). Moreover, lncRNA is related to many aspects of cancer immunity including the recognition and killing of cancer cells, cell migration, and T cell infiltration (Yu et al., 2018).

Therefore, it is reasonable to predict the survival of LUAD patients and guide clinical treatment using immune-related lncRNAs. To further explore the possible roles of immune-related lncRNAs that play in prognosis and immunotherapy, we analyzed the transcriptome data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) database to build an immune-related lncRNA signature.

MATERIALS AND METHODS

LUAD Data Collection and Grouping

TCGA-LUAD datasets including RNA expression profile (n = 585) processed by HTcount, patients survival (n = 738), and phenotype (n = 125) information were downloaded from TCGA[1]. RNA expression profiles of GSE19188 (normal: 65; tumor: 45), GSE27262 (normal: 25; tumor: 25), GSE30219 (normal: 14; tumor: 84), and GSE31210 (normal: 20; tumor: 226) were downloaded from https://www.ncbi.nlm.nih.gov/ and each was normalized by RMA algorithm using R package *affy*. Tumor and normal tissue samples were selected from the above GEO datasets and divided into tumor (n = 375) and normal (n = 124) groups. The survival data of LUAD patients in GSE30219 (n = 84), GSE31210 (n = 226), and GSE50081 (n = 106) were collected and integrated with RNA expression data (n = 416) for prognostic model validation.

Metagene of 28 immune cell subtypes was obtained from https://www.cell.com/cms/10.1016/j.celrep.2016.12.019/attachment/f353dac9-4bf5-4a52-bb9a-775e74d5e968/mmc3.xlsx (Charoentong et al., 2017) to evaluate the infiltration level of immune cells by the single sample gene set enrichment (ssGSEA) method.

Validation of the Data Grouping

ssGSEA and hierarchical cluster were used to divide the subjects into a high immune-cell-infiltrated group and a low immune-cell-infiltrated group. ESTIMATE algorithm was used to validate the grouping by comparing the stromal score, immune score, ESTIMATE score, and tumor purity of the two groups.

Identification of Immune-Related Differentially Expressed lncRNAs

R package *edgeR* was used to find the differentially expressed genes (DEGs) between two pairs: tumor and non-tumor cells (GSE31210, GSE30219, GSE19188, and GSE27262), and high immune-infiltrated and low immune-infiltrated cells (TCGA dataset). $|$ Log fold change$|$ >1 and $p < 0.05$ were used to choose the DEGs because the data have been log-transformed. Then, the lncRNAs that appeared in both groups will be regarded as immune-related lncRNAs.

Prognostic Signature Construction and Validation

The least absolute shrinkage and selection operator (LASSO) regression was used to find out the prognosis-related lncRNAs in the immune-related lncRNAs because it is a robust feature selection algorithm. The survival data of 585 TCGA patients were used. Then univariate Cox regression filtered out those prognostic lncRNAs with $p < 0.005$. Finally, stepwise multivariate Cox regression based on AIC (Akaike information criterion) value was used on the identified lncRNAs to select the ones that minimize AIC to attain the best model fit. The eventual risk score was calculated based on the coefficients of every lncRNAs as below:

$$\text{risk score} = \sum_{\text{i}=1}^{\text{n}} \text{coefi} \times \text{id} \tag{1}$$

All the subjects were divided into high-risk and low-risk groups with respect to the median risk score. Then, the Kaplan–Meier curve was constructed to compare the overall survival (OS) between these two groups. Although the sequencing and processing methods were different for training and testing datasets, the relative gene expression level should be similar. Therefore, it is reasonable to use GEO datasets to test the prognostic lncRNA signature based on the defined coefficients. The area under the ROC curve (AUC), an evaluation of the performance of the model based on true-positive rate and false-positive rate, was plotted to assess the model. Univariate and multivariate Cox regressions were then used to explore whether the risk signature was an independent prognostic factor.

Gene Set Enrichment Analysis (GSEA)

Hallmark gene sets were fetched from the MsigDB database using *msigdbr* (v7.0.1) package in R. The gene list was ranked by the Wald test statistics. R package *fgsea* (v1.14.0) was used to perform GSEA and visualize the top enriched gene sets.

Pearson Correlation Analysis

Infiltration values of immune cell subtypes for LUAD were downloaded from the TIMER database[2] (Li et al., 2016). The

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

[2] http://timer.cistrome.org/

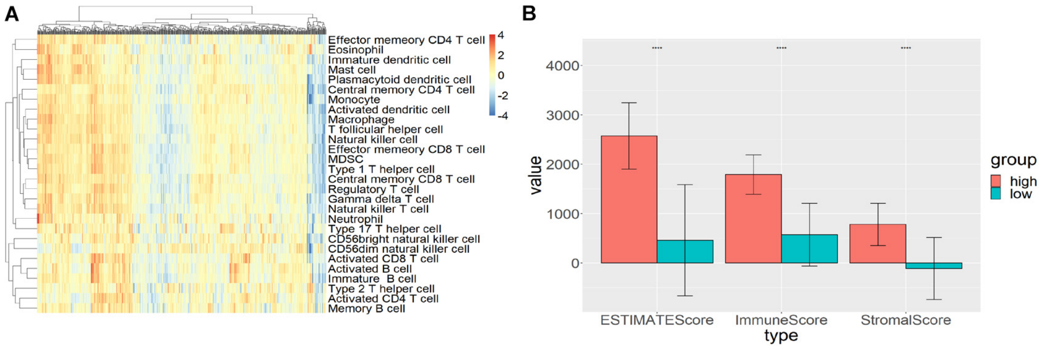


FIGURE 1 | High and low immune-cell-infiltrated groups. **(A)** The GSEA scores for 28 types of immune cells from *GSVA* package using ssGSEA method. Red represented high GSEA score for high immune cell infiltration, blue represented low GSEA score for low immune cell infiltration. **(B)** The stromal score, immune score, and ESTIMATE score for high and low immune-cell-infiltrated groups.

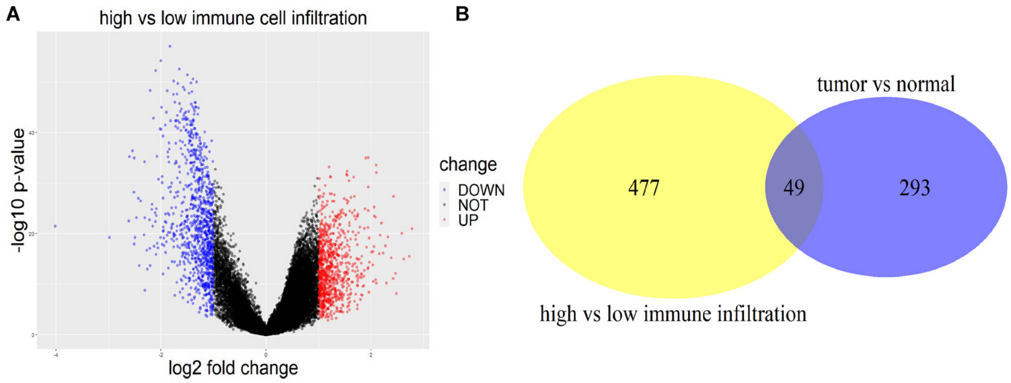


FIGURE 2 | DEGs between high and low immune infiltration groups, and between tumor and normal tissues. **(A)** The volcano plot of DEGs between the high immune-cell-infiltrated and low immune-cell-infiltrated group. Red indicated DEGs up-regulated in the high infiltration group, while blue indicated the down-regulated ones. **(B)** The yellow circle represented the DE lncRNAs between high and low immune infiltration groups. The purple circle represented the DE lncRNAs between tumor and normal tissues. Forty-nine lncRNAs appeared in both groups.

Pearson correlation was calculated between risk scores and infiltration value.

Statistical Analysis

All statistical methods were accomplished by R (4.0.1) using packages *gsva, estimate, glmnet, survival*, and *fgsea*. Two-tailed $p < 0.05$ indicated significant difference if not specified.

RESULTS

Gene Expression Data Grouping and Validation

Single sample gene set enrichment analysis (ssGSEA) and hierarchical clustering algorithm were used to divide the subjects into high immune cell infiltration ($n = 193$) and low immune cell infiltration ($n = 392$) groups. R package *GSVA* was used to calculate the GSEA score for each sample (**Figure 1A**). Then, the *hclust* package was used to hierarchically cluster the samples based on the Euclidean distance of these scores. The two groups derived from clustering were validated by the ESTIMATE algorithm. Compared with the high immune-cell-infiltrated group, the tumor purity of the low immune-cell-infiltrated group was significantly higher while the stromal score, immune score, and ESTIMATE score were significantly lower ($p < 0.0001$; **Figure 1B**).

Identification of Immune-Related DEGs

R package *edgeR* was used to figure out the DEGs between tumor and normal tissues with a threshold of $|\log2$ fold change$| > 1$ and $p < 0.05$ using four datasets (GSE31210, GSE19188, GSE30219, and GSE27262). The DEGs were first identified within each dataset, and then the genes verified in more than one dataset were extracted. In total, 2931 DEGs including 342 lncRNAs of LUAD patients were identified for tumor and normal tissues. With the same criterion, 1,886 (874 up-regulated and 1,012 down-regulated) immune-related DEGs including 526 lncRNAs were found using TCGA data between high and low immune-cell-infiltrated groups (**Figure 2A**). Two-way Venn analysis was carried out to filter the immune-related DEGs for LUAD patients (**Figure 2B**).

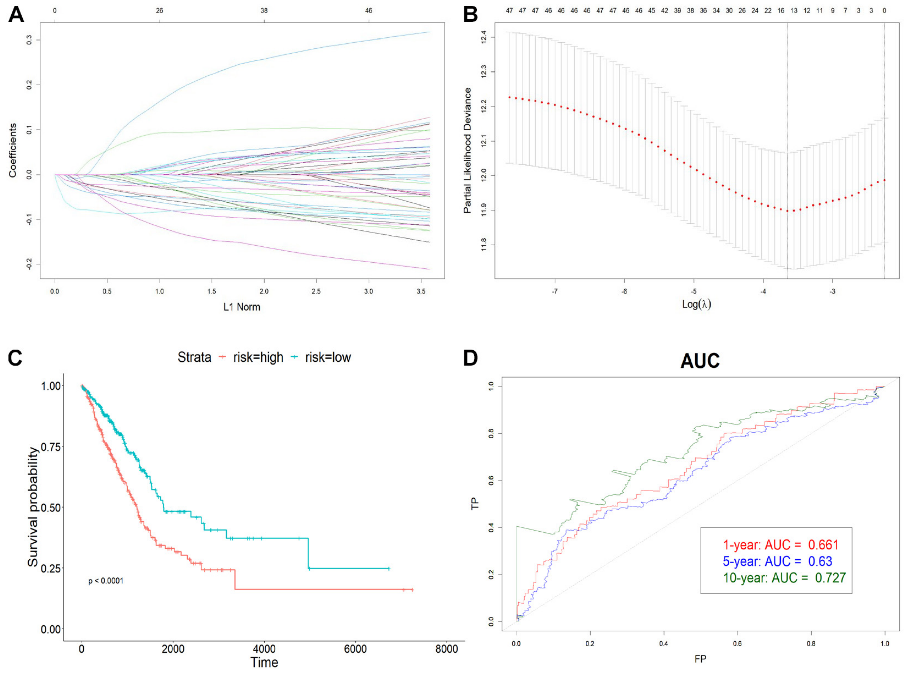


FIGURE 3 | Construction of the immune-related lncRNA signature. **(A)** The LASSO coefficient profiles of 19 prognosis-related lncRNAs. Each colored line showed the change of the coefficient of one lncRNA with the normalization factor. **(B)** Partial likelihood deviance was plotted against the logarithm of lambda in the 10-fold cross-validation. The red dots indicated the deviance and the gray vertical lines indicated standard error of the deviance. The gray vertical dotted line corresponded to the optimal lambda with the lowest partial likelihood deviance. **(C)** Kaplan–Meier curve of high-risk and low-risk groups. **(D)** The AUC of 1-, 5-, and 10-year OS. The *x*-axis represented the false-positive (FP) rate, and the *y*-axis represented the true-positive (TP) rate. The signature predicted the 10-year survival best.

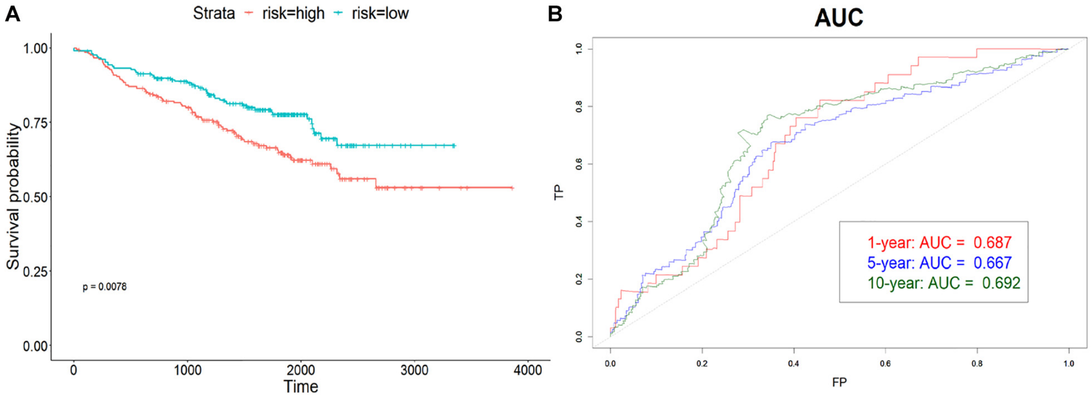


FIGURE 4 | Validation of the signature. **(A)** Kaplan–Meier curve of high-risk and low-risk groups in combined validation dataset. **(B)** The AUC of 1-, 5-, and 10-year OS. The validation data also predicted 10-year survival best.

Immune-Related lncRNA Prognostic Signature Construction Using Regressions

To avoid overfitting, 19 prognostic lncRNAs were selected from the 49 DE lncRNAs using LASSO regression with 10-fold cross-validation (**Figures 3A,B**). Univariate Cox regression was then carried out to increase the robustness with a threshold of $p < 0.005$ and filtered nine lncRNAs for the subsequent step. The four-lncRNA signature was finally constructed by a stepwise multivariate Cox regression with coefficients. The risk score was calculated as below:

Risk score = −0.088*HSPC078 − 0.083*DRAIC − 0.045*AP004608.1 − 0.125*MIR223HG, which is the sum of the multiplication of lncRNA expression and each coefficient.

To determine how well the risk score could predict OS, LUAD patients were divided into high-risk and low-risk groups with respect to the median risk score. The Kaplan–Meier curve showed that the OS of the high-risk

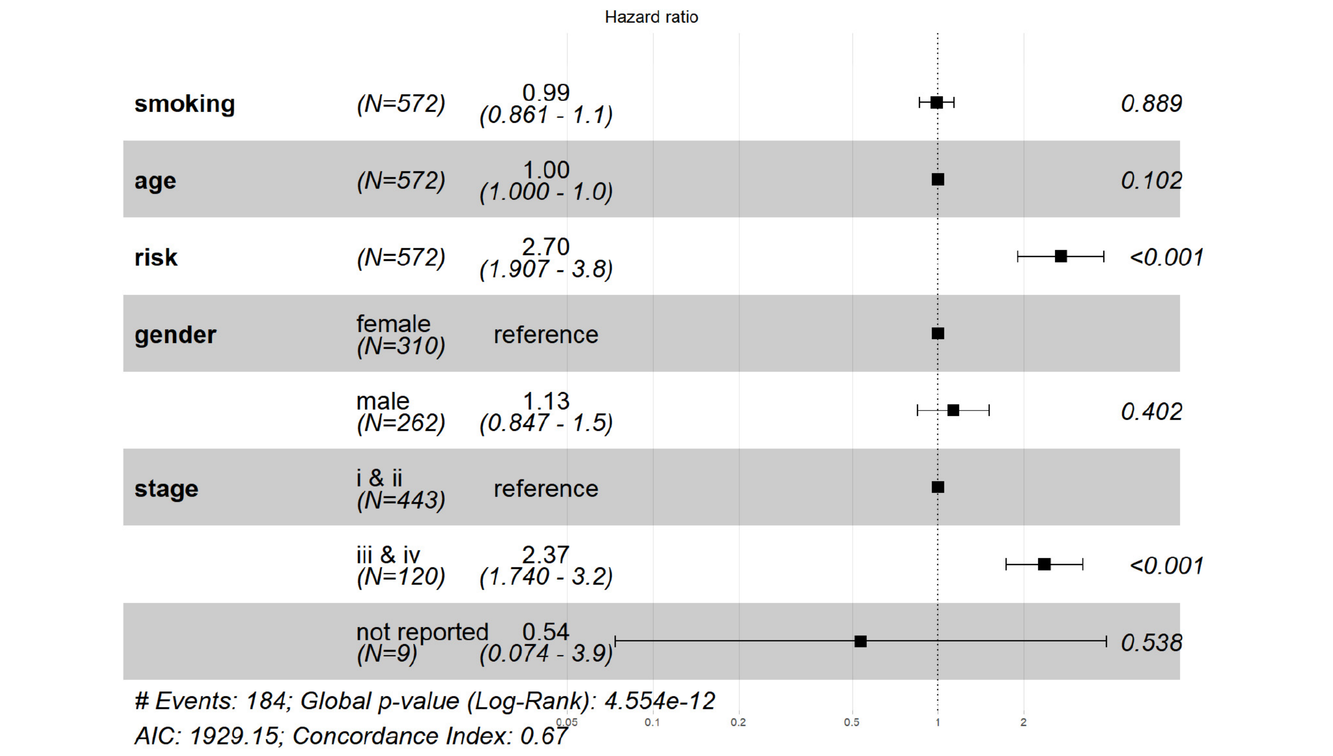


FIGURE 5 | Verification that the signature is an independent prognostic factor. From left to right, the column represented: factor name, number of subjects, HR (lower and upper 95% value), the HR plot, and the *p*-value. The result of multivariate Cox regression showed that risk score and tumor stage are significant prognostic factors for LUAD patients.

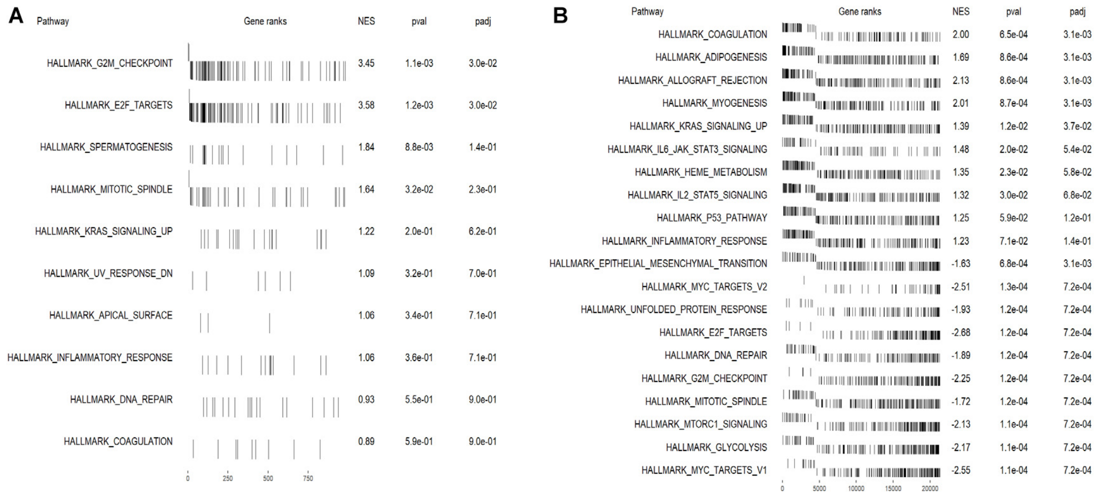


FIGURE 6 | The top enriched gene sets in up- and down-regulated DEGs in the high-risk group. **(A)** Gene sets enriched in DEGs up-regulated in the high-risk group. G2M checkpoints and E2F target-related genes were significantly enriched. **(B)** Gene sets enriched in DEGs down-regulated in the high-risk group. Genes related to interleukin, STAT, KRAS, and p53 were enriched, while cell mitosis-related genes were significantly deprived.

group was significantly worse than that of the low-risk one ($p < 0.0001$) (**Figure 3C**). Also, the AUC plot suggested that the signature could predict the survival well in a long time course (0.661, 0.63, and 0.727 for 1-, 5-, and 10-year survival) (**Figure 3D**).

Validation of the Effectiveness of lncRNA Signature

The model was validated using GSE31210, GSE30219, and GSE50081 datasets using the coefficients trained previously. Samples in each dataset were assigned to the high-risk

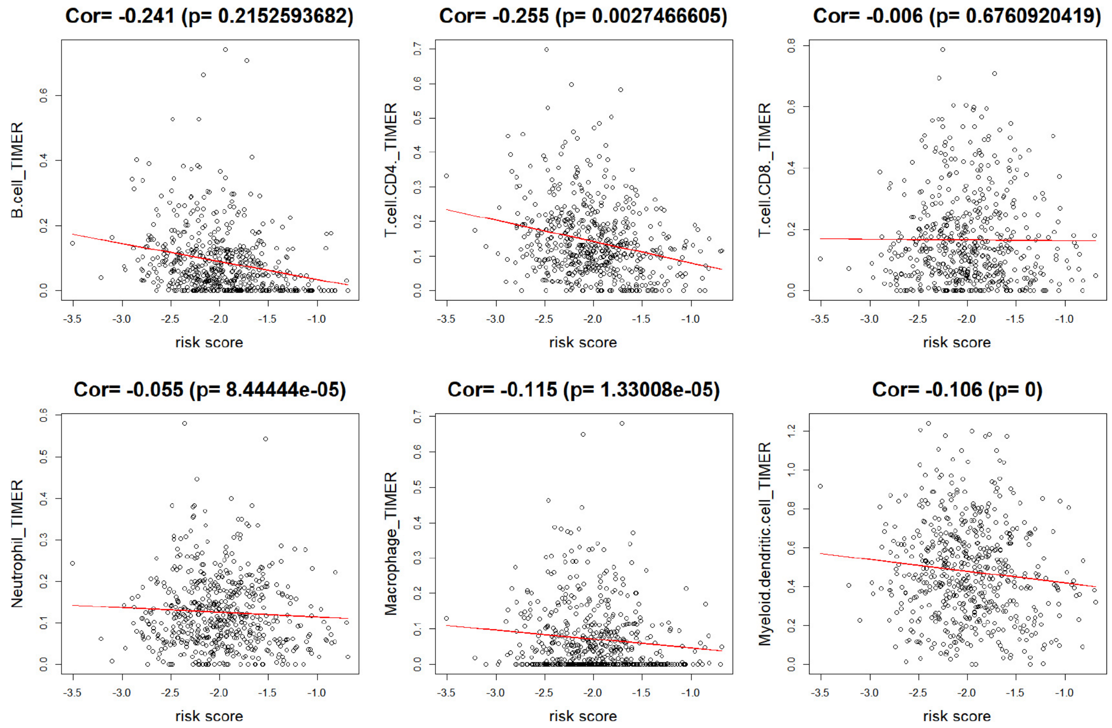


FIGURE 7 | Correlation between risk score and immune cell subtype infiltration. The correlation values of B cells, CD4$^+$ T cells, CD8$^+$ T cells, neutrophils, macrophages, and myeloid dendritic cells were −0.241, −0.255, −0.006, −0.055, −0.115, and −0.106, respectively. Only B cells, CD4$^+$ T cells, macrophages, and myeloid dendritic cells were significantly correlated with risk score ($p < 0.05$).

or low-risk group based on the median risk score. Then, the assignment results of three datasets were combined to plot the Kaplan–Meier curve and AUC. From **Figure 4A**, the survival time of the high-risk group was significantly shorter ($p < 0.01$) than the low-risk group in the combined validating set, which suggested that the risk score can predict the OS well. The area under ROC was 0.687, 0.677, and 0.697 for the 1-, 5-, and 10-year OS (**Figure 4B**). Same as the training set, the lncRNA signature predicted the 10-year survival best.

The Immune-Related Signature Could Serve as an Independent Prognostic Factor

The risk score was then analyzed by Cox regression along with age, gender, tumor stage, and smoking history as an independent factor. The p value of risk score < 0.001 in both univariate and multivariate (**Figure 5**). Cox regression indicated that risk score could serve as an independent prognostic factor. The risk score and advanced tumor stage were risk factors for LUAD patients with a hazard ratio (HR) larger than 1 as shown in **Figure 5**.

Functional Analysis Revealed Related Signaling Pathways and Micro-RNAs

To identify the enriched gene sets for DEGs ranked by the Wald test statistics, the *fgsea* package was used to do GSEA analysis for up- and down-regulated genes in the high-risk group separately. Several mitosis-related gene sets including E2F target (NES = 3.58) and G2M checkpoints (NES = 3.45) were enriched in the up-regulated DEGs in the high immune-cell-infiltrated group (**Figure 6A**). In the down-regulated DEGs, signaling pathways including IL6-JAK-STAT3, KRAS (down-regulated), IL2-STAT5, and p53 pathways were enriched (**Figure 6B**). These results showed that the immune-related lncRNAs may promote cancer progression by advancing cell mitosis.

Also, some micro-RNAs (miRNAs) were related to these immune-related prognostic lncRNAs. From the LncBase database (Paraskevopoulou et al., 2016), we found that 21 miRNAs have been verified to interact with these lncRNAs by experiments. The genes regulated by these miRNAs were enriched in ECM-receptor, viral carcinogenesis, p53 signaling, and hippo signaling pathways (Ioannis et al., 2015; Dimitra et al., 2018). mir-30, mir-10, and mir-181 played important roles in these pathways.

The lncRNA Signature Was Associated With B Cell, CD4$^+$ T Cell, Macrophage, and Myeloid Dendritic Cell Infiltration

To explore the relationship between lncRNAs and the infiltration of some representative immune cells, the Pearson correlation value was calculated between risk scores and TIMER estimated infiltration value. As shown in **Figure 7**, the infiltration values of B cells, CD4^{+} T cells, macrophages, and myeloid dendritic cells were significantly negatively correlated with risk scores. The negative coefficients illustrated that the immune-related lncRNA signature was associated with high infiltration of immune cell subtypes.

DISCUSSION

We obtained data from TCGA and GEO database to identify immune-related differentially expressed lncRNAs of LUAD patients. Patients were grouped into high and low immune-cell-infiltrated groups by GSVA, which was further validated by ESTIMATE. LASSO regression, univariate Cox regression, and stepwise multivariate Cox regression were used to build a four-lncRNA prognostic signature. The risk score was calculated using the coefficients of the four lncRNAs, based on which patients were classified into low-risk and high-risk groups. The OS of the high-risk group was significantly shorter than the low-risk group in both the training and the testing datasets. The AUCs showed that the risk signature has a good prediction of 10-year survival. The lncRNA signature was confirmed to be an independent prognostic factor when analyzed by multivariate Cox regression along with age, gender, tumor stage, and smoking history. Finally, the functional GSEA analysis was performed to investigate how the lncRNAs may affect the OS.

Our model showed consistent results in predicting OS using both RNA-seq and microarray datasets although the coefficients were trained only by the RNA-seq data. This could be explained by the robust prognostic value of the four lncRNAs. All lncRNAs in the signature, including SIGLEC17P, DRAIC, MIR223HG, and AP004608.1, are protective for LUAD patients as shown by the negative coefficients. SIGLEC17P was suppressed in the advanced stage of cancer (iii and iv), which illustrated that the dysfunction of it may be associated with cancer progression (Zhou et al., 2019). Previous studies have implied that DRAIC may inhibit cell migration and invasion and predict longer survival time in LUAD patients (Sakurai et al., 2015). AP004608.1 was a protective lncRNA in pancreatic adenocarcinoma (Wang et al., 2019). MIR223HG has also been identified as a prognostic lncRNA related to tumor microenvironment in another study (Jin et al., 2020) with HR < 1, which was consistent with our results.

The GSEA results indicated that genes highly expressed in the high-risk group could promote cell mitosis, while genes expressed lowly seems to promote p53 IL6-JAK-STAT3 and IL2-STAT5 pathways and decrease KRAS signaling. The tumor suppressor protein p53 was suggested to regulate cell growth by promoting apoptosis and DNA repair under stressful conditions (Kanapathipillai, 2018). KRAS signaling is oncogenic and was reported to regulate tumor-associated immune responses such as inducing cancer cell evasion from immunosurveillance (Dias Carvalho et al., 2017). Therefore, the down-regulation of KRAS could delay cancer progression and benefit immunotherapy. The JAK-STAT gene set was enriched in up-regulated DEGs as a downstream pathway of interferon-gamma signaling, which is an essential responsive cytokine in cytotoxic T cells mediated killing of tumor cells (Barnholt et al., 2009; Ni and Lu, 2018). These pathways enriched in down-regulated DEGs in the high-risk group have contributed to tumor suppression in various ways that are associated with tumor immunity. Also, these lncRNAs were correlated with miRNAs including mir-30, mir-10, and mir-181. Mir-30 was shown to be a tumor suppressor gene by many studies (Braun et al., 2010; Cheng et al., 2012). As mir-10 were de-regulated in many cancers (Lund, 2010), their up-regulation may decrease the progression of cancer. The down-regulation of mir-181 was suggested to regulate PTEN expression and thus inhibit tumor development (Chang et al., 2017).

As shown by the correlation analysis, the negative correlation between risk scores and infiltration values illustrated that higher expression of lncRNA signature was correlated with higher immune cell infiltration and thus longer OS. This might be explained by the fact that the signaling pathways correlated with lncRNA expressions that could also affect[tumor immunity.

In conclusion, we identified a novel four-lncRNA prognostic signature that was associated with the infiltration of immune cell subtypes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/projects/TCGA-LUAD.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the TCGA Ethics, Law and Policy Group. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JW analyzed data and wrote the manuscript. XY searched manuscripts and cleaned data. XJ designed the research and modified the manuscript. Y-QZ designed the research and provided clinical insights. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Beijing Traditional Chinese Medicine Science and Technology Development Project (JJ2014-52). The research was supported by the National Natural Science Foundation of China (Nos. 81774198, 81573871, and 81273634) and the Open Project of Shandong Co-Innovation Center.

REFERENCES

Barnholt, K., Kota, R., Aung, H., and Rutledge, J. (2009). Adenosine blocks IFN-γ-induced phosphorylation of STAT1 on serine 727 to reduce macrophage activation. *J. Immunol*. 183, 6767–6777. doi: 10.1002/ana.21634

Braun, J., Hoang-Vu, C., Dralle, H., and Hüttelmaier, S. (2010). Downregulation of micrornas directs the emt and invasive potential of anaplastic thyroid carcinomas. *Oncogene* 29, 4237–4244. doi: 10.1038/onc.2010.169

Brody, R., Zhang, Y., Ballas, M., Siddiqui, M., Gupta, P., Barker, C., et al. (2017). PD-L1 expression in advanced NSCLC: insights into risk stratification and treatment selection from a systematic literature review. *Lung Cancer* 112, 200–215. doi: 10.1016/j.lungcan.2017.08.005

Chang, S., Chen, B., Wang, X., Wu, K., and Sun, Y. (2017). Long non-coding rna xist regulates pten expression by sponging mir-181a and promotes hepatocellular carcinoma progression. *BMC Cancer* 17:248. doi: 10.1186/s12885-017-3216-6

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep*. 18, 248–262. doi: 10.1016/j.celrep.2016.12.019

Cheng, C. W., Wang, H. W., Chang, C. W., Chu, H. W., Chen, C. Y., Yu, J. C., et al. (2012). Microrna-30a inhibits cell migration and invasion by downregulating vimentin expression and is a potential prognostic marker in breast cancer. *Breast Cancer Res. Treat*. 134, 1081–1093. doi: 10.1007/s10549-012-2034-4

Dias Carvalho, P., Guimarães, C., Cardoso, A., Mendonça, S., Costa, Â, Oliveira, M., et al. (2017). KRAS oncogenic signaling extends beyond cancer cells to orchestrate the microenvironment. *Cancer Res*. 78, 7–14. doi: 10.1158/0008-5472.CAN-17-2084

Dimitra, K., Paraskevopoulou, M. D., Serafeim, C., Vlachos, I. S., Spyros, T., Ilias, K., et al. (2018). Diana-tarbase v8: a decade-long collection of experimentally supported mirna–gene interactions. *Nucleic Acids Res*. 46, D239–D245. doi: 10.1093/nar/gkx1141

Duffy, M., and O'Byrne, K. (2018). Tissue and blood biomarkers in lung cancer: a review. *Adv. Clin. Chem*. 86, 1–21. doi: 10.1016/bs.acc.2018.05.001

Ioannis, S. V., Konstantinos, Z., Maria, D. P., Georgios, G., Dimitra, K., Thanasis, V., et al. (2015). Diana-mirpath v3.0: deciphering microrna function with experimental support. *Nucleic Acids Res*. 43, W460–W466. doi: 10.1093/nar/gkv403

Jiang, R., Tang, J., Chen, Y., Deng, L., Ji, J., Xie, Y., et al. (2017). The long noncoding RNA lnc-EGFR stimulates T-regulatory cells differentiation thus promoting hepatocellular carcinoma immune evasion. *Nat. Commun*. 8:15129.

Jin, D., Song, Y., Chen, Y., and Zhang, P. (2020). Identification of a seven-lncRNA immune risk signature and construction of a predictive nomogram for lung adenocarcinoma. *BioMed. Res. Int*. 2020:7929132. doi: 10.1155/2020/7929132

Kanapathipillai, M. (2018). Treating p53 mutant aggregation-associated cancer. *Cancers* 10:154. doi: 10.3390/cancers10060154

Li, B., Severson, E., Pignon, J., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol*. 17:174.

Liu, J., Zhong, Y., Peng, S., Zhou, X., and Gan, X. (2018). Efficacy and safety of PD1/PDL1 blockades versus docetaxel in patients with pretreated advanced non-small-cell lung cancer: a meta-analysis. *Onco Targets Ther*. 11, 8623–8632. doi: 10.2147/OTT.S181413

Lortet-Tieulent, J., Soerjomataram, I., Ferlay, J., Rutherford, M., Weiderpass, E., and Bray, F. (2014). International trends in lung cancer incidence by histological subtype: adenocarcinoma stabilizing in men but still increasing in women. *Lung Cancer* 84, 13–22. doi: 10.1016/j.lungcan.2014.01.009

Lund, A. (2010). miR-10 in development and cancer. *Cell Death Differ*. 17, 209–214. doi: 10.1038/cdd.2009.58

Ni, L., and Lu, J. (2018). Interferon gamma in cancer immunotherapy. *Cancer Med*. 7, 4509–4516. doi: 10.1002/cam4.1700

Pagès, F., Granier, C., Kirilovsky, A., Elsissy, C., and Tartour, E. (2016). Biomarqueurs prédictifs de réponse aux traitements bloquant les voies de costimulation inhibitrices. *Bull. Cancer* 103, S151–S159. doi: 10.1016/S0007-4551(16)30373-3

Paraskevopoulou, M. D., Vlachos, I. S., Dimitra, K., Georgios, G., Ilias, K., Thanasis, V., et al. (2016). Diana-lncbase v2: indexing microrna targets on non-coding transcripts. *Nucleic Acids Res*. 44, D231–D238. doi: 10.1093/nar/gkv1270

Quinn, J., and Chang, H. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet*. 17, 47–62. doi: 10.1038/nrg.2015.10

Sacher, A., and Gandhi, L. (2016). Biomarkers for the clinical use of PD-1/PD-L1 inhibitors in non-small-cell lung cancer. *JAMA Oncol*. 2, 1217–1222. doi: 10.1001/jamaoncol.2016.0639

Sakurai, K., Reon, B., Anaya, J., and Dutta, A. (2015). The lncRNA DRAIC/PCAT29 locus constitutes a tumor-suppressive nexus. *Mol. Cancer Res*. 13, 828–838. doi: 10.1158/1541-7786.MCR-15-0016-T

Sela-Culang, I., Benhnia, M., Matho, M., Kaever, T., Maybeno, M., Schlossman, A., et al. (2014). Using a combined computational-experimental approach to predict antibody-specific B cell epitopes. *Structure* 22, 646–657. doi: 10.1016/j.str.2014.02.003

Shih, J., Chen, H., Lin, S., Yeh, Y., Shen, R., Lang, Y., et al. (2020). Integrative analyses of noncoding RNAs reveal the potential mechanisms augmenting tumor malignancy in lung adenocarcinoma. *Nucleic Acids Res*. 48, 1175–1191. doi: 10.1093/nar/gkz1149

Wang, Y., Huang, T., Sun, X., and Wang, Y. (2019). Identification of a potential prognostic lncRNA-miRNA-mRNA signature in endometrial cancer based on the competing endogenous RNA network. *J. Cell. Biochem*. 120, 18845–18853. doi: 10.1002/jcb.2920

Yu, W., Wang, H., He, Q., Xu, Y., and Wang, X. (2018). Long noncoding RNAs in cancer-immunity cycle. *J. Cell. Physiol*. 233, 6518–6523. doi: 10.1002/jcp.26568

Zhou, W., Liu, T., Saren, G., Liao, L., Fang, W., and Zhao, H. (2019). Comprehensive analysis of differentially expressed long non-coding RNAs in non-small cell lung cancer. *Oncol. Lett*. 18, 1145–1156. doi: 10.3892/ol.2019.10414

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.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 12 March 2021
doi: 10.3389/fgene.2021.648329

Deep Learning Reveals Key Immunosuppression Genes and Distinct Immunotypes in Periodontitis

Wanchen Ning[1], Aneesha Acharya[2,3], Zhengyang Sun[4], Anthony Chukwunonso Ogbuehi[5], Cong Li[6], Shiting Hua[6], Qianhua Ou[6], Muhui Zeng[6], Xiangqiong Liu[7], Yupei Deng[7], Rainer Haak[8], Dirk Ziebolz[8†], Gerhard Schmalz[8†], George Pelekos[3], Yang Wang[9]* and Xianda Hu[7]**

[1] Department of Conservative Dentistry and Periodontology, Ludwig-Maximilians-University of Munich, Munich, Germany, [2] Dr. D. Y. Patil Dental College and Hospital, Dr. D. Y. Patil Vidyapeeth, Pune, India, [3] Faculty of Dentistry, The University of Hong Kong, Hong Kong, China, [4] Faculty of Mechanical Engineering, Chemnitz University of Technology, Chemnitz, Germany, [5] Faculty of Physics, University of Münster, Münster, Germany, [6] Zhujiang Hospital, Southern Medical University, Guangzhou, China, [7] Laboratory of Cell and Molecular Biology, Beijing Tibetan Hospital, China Tibetology Research Center, Beijing, China, [8] Department of Cariology, Endodontology and Periodontology, University of Leipzig, Leipzig, Germany, [9] State Key Laboratory of Biocatalysis and Enzyme Engineering, Hubei Collaborative Innovation Center for Green Transformation of Bio-Resources, School of Life Sciences, Hubei University, Wuhan, China

Background: Periodontitis is a chronic immuno-inflammatory disease characterized by inflammatory destruction of tooth-supporting tissues. Its pathogenesis involves a dysregulated local host immune response that is ineffective in combating microbial challenges. An integrated investigation of genes involved in mediating immune response suppression in periodontitis, based on multiple studies, can reveal genes pivotal to periodontitis pathogenesis. Here, we aimed to apply a deep learning (DL)-based autoencoder (AE) for predicting immunosuppression genes involved in periodontitis by integrating multiples omics datasets.

Methods: Two periodontitis-related GEO transcriptomic datasets (GSE16134 and GSE10334) and immunosuppression genes identified from DisGeNET and HisgAtlas were included. Immunosuppression genes related to periodontitis in GSE16134 were used as input to build an AE, to identify the top disease-representative immunosuppression gene features. Using K-means clustering and ANOVA, immune subtype labels were assigned to disease samples and a support vector machine (SVM) classifier was constructed. This classifier was applied to a validation set (Immunosuppression genes related to periodontitis in GSE10334) for predicting sample labels, evaluating the accuracy of the AE. In addition, differentially expressed genes (DEGs), signaling pathways, and transcription factors (TFs) involved in immunosuppression and periodontitis were determined with an array of bioinformatics analysis. Shared DEGs common to DEGs differentiating periodontitis from controls and those differentiating the immune subtypes were considered as the key immunosuppression genes in periodontitis.

Results: We produced representative molecular features and identified two immune subtypes in periodontitis using an AE. Two subtypes were also predicted in the validation

OPEN ACCESS

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Mengqi Zhang,
University of Pennsylvania,
United States
Fulong Yu,
Division of Hematology/Oncology,
Boston Children's Hospital,
United States
Tingting Shao,
Harbin Medical University, China

****Correspondence:***
George Pelekos
george74@hku.hk
Yang Wang
yangwang@hubu.edu.cn
Xianda Hu
hellocean@hotmail.com

† These authors share senior authorship

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

Received: *31 December 2020*
Accepted: *18 February 2021*
Published: *12 March 2021*

Citation:
Ning W, Acharya A, Sun Z, Ogbuehi AC, Li C, Hua S, Ou Q, Zeng M, Liu X, Deng Y, Haak R, Ziebolz D, Schmalz G, Pelekos G, Wang Y and Hu X (2021) Deep Learning Reveals Key Immunosuppression Genes and Distinct Immunotypes in Periodontitis. Front. Genet. 12:648329. doi: 10.3389/fgene.2021.648329

set with the SVM classifier. Three "master" immunosuppression genes, PECAM1, FCGR3A, and FOS were identified as candidates pivotal to immunosuppressive mechanisms in periodontitis. Six transcription factors, NFKB1, FOS, JUN, HIF1A, STAT5B, and STAT4, were identified as central to the TFs-DEGs interaction network. The two immune subtypes were distinct in terms of their regulating pathways.

Conclusion: This study applied a DL-based AE for the first time to identify immune subtypes of periodontitis and pivotal immunosuppression genes that discriminated periodontitis from the healthy. Key signaling pathways and TF-target DEGs that putatively mediate immune suppression in periodontitis were identified. PECAM1, FCGR3A, and FOS emerged as high-value biomarkers and candidate therapeutic targets for periodontitis.

Keywords: deep learning, autoencoder (AE), periodontitis, immunosuppression genes, therapeutic targets, bioinformatics

INTRODUCTION

Periodontitis involves the inflammatory destruction of the supporting tissues of teeth. It involves a perturbed local host immune response that is ineffective in countering plaque biofilm microbiota (Meyle and Chapple, 2015). Innate and adaptive immunity work in tandem to counter the infectious challenge posed by oral microbiota, limit the spread of infection, and reestablish periodontal tissue homeostasis (Cekici et al., 2014). This delicately orchestrated process involves the actions of several immune regulatory cell types, including oral epithelial cells (Dutzan et al., 2016), neutrophils (Scott and Krauss, 2011), macrophages, dendritic cells (Zhou et al., 2019), B cells, and T cells (Gemmell et al., 2002). Regulatory T cells (Tregs) have particularly attracted much recent attention as they engender multiple suppressive mechanisms to inhibit various cells involved in innate and adaptive immunity. The role of Tregs in controlling periodontitis due to their immune-suppressive capabilities has been noted (Alvarez et al., 2018). Immune suppression demands the tandem action of multiple immunosuppression genes, several of which have been demonstrated in the context of periodontal pathology. These include programmed cell death 1 (PD1), PD-Ligand 1 (PD-L1) (Bailly, 2020), and Cytotoxic T-Lymphocyte Antigen4 (CTLA4) (Aoyagi et al., 2000), that function as immune checkpoint inhibitors to modulate B-cells, CD8+ T-cells, and CD4+ T-cells, which can amplify infection and promote tissue damage. Therefore, an immune checkpoint blockade has been proposed as a modality to manage periodontitis. However, existing reports have documented very few immunosuppression genes in the context of periodontitis. It is also recognized that immunosuppressive agents impose a risk for periodontal diseases, inducing gingival overgrowth or other alterations in periodontal tissues (Cota et al., 2010). Immunosuppressive medications for immune-related disorders such as rheumatoid arthritis or solid organ transplantation are associated with periodontal disease. However, the underlying molecular mechanisms remain unclear, and few genes have been implicated. For instance, specific Human Leukocyte Antigen (HLA)-DR1 genotype is documented to protect from gingival overgrowth induced by cyclosporine A (Cebeci et al., 1996). A more expansive understanding of immune suppression genes that are relevant to periodontal disease pathology can lead the identification of candidate genes and molecular pathways of significant potential translational value. Such data may enable the development of gene and targeted drug therapy for multiple periodontal diseases.

Experimental studies are limited by scale, incomplete or inaccurate existing databases, and the cost-intensive nature of molecular experiments, so approaches that can predict previously unidentified gene functions, enable gene function discovery, and automate the identification of inaccuracies can be very valuable (Chicco et al., 2014). Deep learning (DL) computational frameworks are capable of these. In this regard, an autoencoder (AE), is essentially a dimensionality reduction tool, as the "building block" of DL, comprises of a three-layered unsupervised artificial neural network that performs extraction of representative features (Lee et al., 2009; Wang et al., 2016). The AE has been implemented as a DL framework to predict survival in liver cancer (Chaudhary et al., 2018), breast cancer (Tan et al., 2014), head and neck squamous cell carcinoma (HNSCC) (Zhao et al., 2019), and when applied to RNA-seq data (Xiao et al., 2018) has shown value in generating key features from gene expression data that are linked to clinical outcomes.

To our knowledge, the present study is the first to integrate multi-omics data pertaining to immunosuppression genes in periodontitis using a DL-based AE combined with a support vector machine (SVM) classifier (Ju et al., 2015) confirmed in a validation set, along with an array of bioinformatic analysis, with an aim to identify the most significant immunosuppression genes relevant to the pathogenesis of periodontitis.

MATERIALS AND METHODS

Study Design

An overview of the workflow of this study is depicted in **Figure 1**. In brief, two cohorts of periodontitis datasets (GSE16134 and GSE10334) and immunosuppression genes were included.

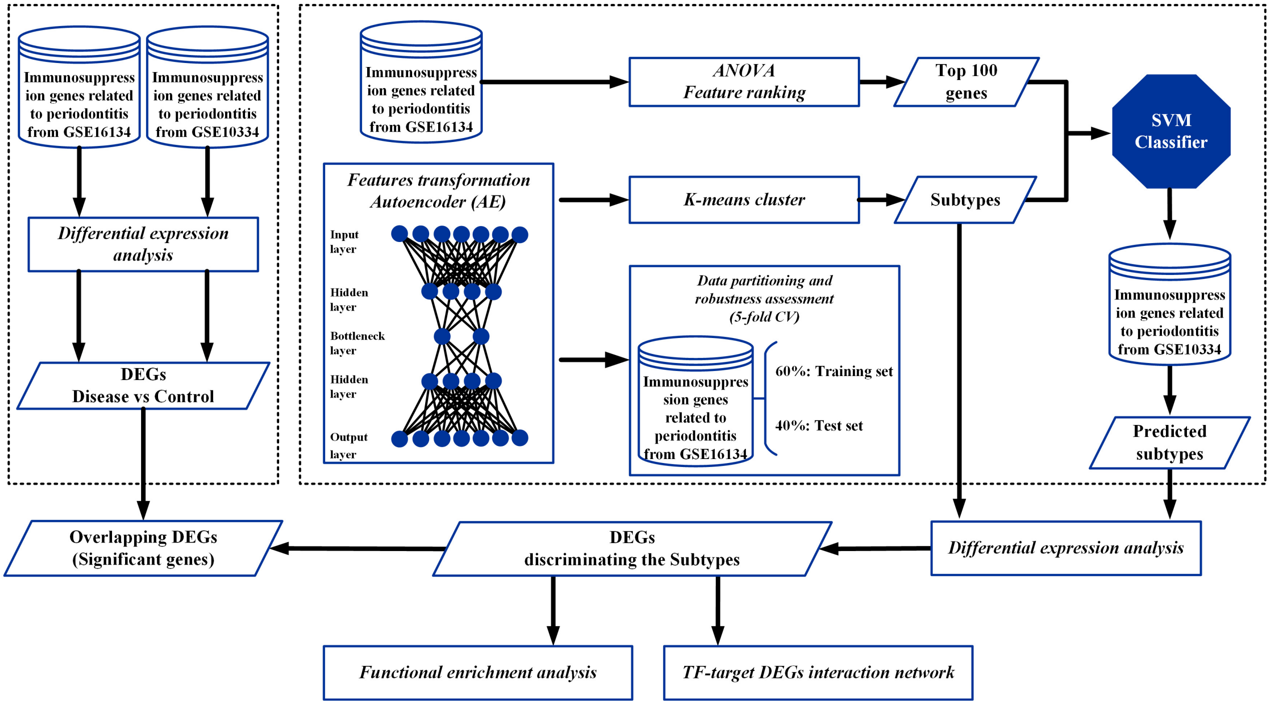

Immunosuppression genes related to periodontitis from GSE16134
Immunosuppression genes related to periodontitis from GSE10334
Differential expression analysis
DEGs Disease vs Control
Immunosuppression genes related to periodontitis from GSE16134
ANOVA Feature ranking
Top 100 genes
SVM Classifier
Features transformation Autoencoder (AE)
Input layer
Hidden layer
Bottleneck layer
Hidden layer
Output layer
K-means cluster
Subtypes
Data partitioning and robustness assessment (5-fold CV)
Immunosuppression genes related to periodontitis from GSE16134
60%: Training set
40%: Test set
Immunosuppression genes related to periodontitis from GSE10334
Predicted subtypes
Overlapping DEGs (Significant genes)
DEGs discriminating the Subtypes
Differential expression analysis
Functional enrichment analysis
TF-target DEGs interaction network

FIGURE 1 | Overall workflow. The flowchart depicts the autoencoder (AE) architecture and workflow combining deep learning (DL) techniques to identify key immunosuppression genes in periodontitis. Immunosuppression genes related to periodontitis from GSE16134 were applied as input features for an AE. The new transformed features in the bottleneck layer of the AE were clustered into different subtypes using K-mean clustering. Then, based on the clustering labels, we selected the top 100 most related genes from GSE16134 based on ANOVA *F* values. The input dataset was split at a 60%/40% ratio (training set/test set) to assess the robustness of the AE, using a 5-fold CV. Subsequently, based on the above labels of GSE16134, an SVM classifier was built and further applied for prediction in a validation set (GSE10334). To explore the biological roles of the different identified subtypes, differentially expressed genes (DEGs) and transcription factors (TFs), differential expression analysis, functional enrichment analysis, and construction of TF-target DEGs interaction network were, respectively, applied. Eventually, to identify the immunosuppression genes that might be most pertinent to periodontitis, the overlapping DEGs among the DEGs discriminating disease (periodontitis) and controls and DEGs discriminating the subtypes classified with the AE and SVM models were determined.

First, immunosuppression genes related to periodontitis from GSE16134 were identified and applied as input features to build an AE model. Second, each of the new transformed features in the bottleneck layer of the AE was clustered into different subgroups using K-mean clustering. In addition, based on the clustering labels, we selected the top 100 most related genes from GSE16134 based on ANOVA *F* values. Data partitioning of the inferring samples of GSE16134 was applied to assess the robustness of the AE, using a 5-fold CV. The samples were randomly split into 5 folds, 3 of which were used as the training set (60%) and the remaining 2 (40%) as the test set. Thereafter, based on the clustering results and the top 100 genes of GSE16134, a SVM classifier was built with a 5-fold CV to identify the optimal hyperparameters, and a validation set (immunosuppression genes related to periodontitis in GSE10334) was applied for SVM to predict the subtypes. To explore the biological roles of the different identified subtypes, differentially expressed genes (DEGs) and transcription factors (TFs), differential expression analysis, functional enrichment analysis, and construction of TF-target DEGs interaction network were, respectively, applied. Finally, to identify the immunosuppression genes that might be most pertinent to periodontitis, the overlapping DEGs among the DEGs discriminating periodontitis and controls and DEGs discriminating the subtypes classified with the DL-based model were determined.

Pre-processing of the Dataset

Transcriptomic data from gingival tissue samples affected with periodontitis and the corresponding controls (GSE16134 and GSE10334) were obtained from the Gene Expression Omnibus (GEO) database of NCBI[1]. Detailed information of the two datasets is listed in **Table 1**. Immunosuppression genes were obtained from databases DisGeNET[2] and HisgAtlas[3]. From these obtained genes, 1,207 immunosuppression genes related to periodontitis were extracted. Next, the two datasets were stacked, and 1,181 immunosuppression genes' expression profiles were found matching in the two datasets. Subsequently, the two datasets were standardized using the "scale" function in R, setting the parameters as (scale = TRUE and center = FALSE).

[1] http://www.ncbi.nlm.nih.gov/geo/
[2] http://www.disgenet.org
[3] http://biokb.ncpsb.org/HisgAtlas/

TABLE 1 | Summary of periodontitis related GEO datasets used in this study.

Data	GPL (General public license)	Gene	Sample control	Sample case
GSE16134	GPL570	24441	69	241
GSE10334	GPL570	24441	64	183

Features Transformation

Immunosuppression gene expression profiles of 241 disease samples in the GSE16134 dataset were selected as the input for the AE. The re-coding of the DL algorithms was performed using the Python library "Keras[4]". An AE is a three-layered neural network consisting of input, hidden, and output layers (Wang et al., 2016), and here an AE with three hidden layers was implemented with 200, 100, and 200 nodes per layer each. One hundred nodes produced by the bottleneck layer were regarded as the new compressed representative features of the data. In accordance with previous research, the AE was set up using the following equations (Chaudhary et al., 2018).

$$y = f_i(x) = tanh\ (w_i.x + b_i)$$

$$x' = F_{1\to k}(x) = f_1^o \ldots {}^o f_{k-1}^o {}^o f_k(x)$$

$$\text{logloss}(x, x') = \sum_{k=1}^{d} (x_k log(x'_k) + (1 - x_k) log\ (1 - x'_k))$$

$$L(x, x') = \text{logloss}(x, x') + \sum_{i=1}^{k} (\partial_w ||W_i||_1 + \partial_a ||F_{1\to i}(x)||_2^2)$$

To control overfitting, the penalty values αα and αw (the activity regularizer of layer output) were set to 0.00002 and 0.00001. In addition, the AE was trained using the gradient descent algorithm with 20 epochs and 50% dropout. Here, an epoch is an iteration that indicates the number of passes of the entire training dataset, while the size 20 is one of the appropriate training cycles calculated in the evaluation of the model.

K-Means Clustering to Identify Subtypes of Immunosuppression Genes in Periodontitis

The 100 nodes from the bottleneck-hidden layer were considered as new features for the analysis and were clustered with the K-means algorithm. The optimal number of clusters was determined based on two metrics: Silhouette index (Rousseeuw, 1987) and Calinski–Harabasz index (Calinski and Harabasz, 1974), using scikit-learn package (Pedregosa et al., 2011).

Comparison of AE With PCA Based Clustering

Principal component analysis (PCA), a conventional dimension reduction approach was applied to compare with the AE performance (Chaudhary et al., 2018). The same number (100) of the principal components were set as the features in the bottleneck layer and clustering performances of AE and PCA were evaluated using the Silhouette index (Rousseeuw, 1987).

Data Partitioning and Robustness Assessment

Data partitioning of the inferring samples of GSE16134 was done to assess the robustness of the model, using a cross-validation (CV)–like procedure, as described in earlier reports (Chaudhary et al., 2018; Zhao et al., 2019). First, the samples were randomly split into 5 folds, 3 of which were used as the training set (60%) and the remaining 2 (40%) as the test set. Using this CV approach,10 new combinations (folds) were obtained. In each, a distinct AE and a classifier were constructed in each training fold and were used for predicting the labels in the test set. Eventually, category labels were inferred using an AE based on all the samples, and these labels were used for predicting labels of the validation dataset.

Supervised Classification

First, the obtained features from GSE16134 were standardized with the "scale" function in R, setting the scale as (center = TRUE and scale = TRUE). Then, the top 100 "most relevant" immunosuppression genes in GSE16134 were selected based on the clustering labels and analysis of variance (ANOVA) *F* values. Since the top 100 genes were also present in GSE10334 dataset, a complementation test for missing genes was not conducted. Subsequently, based on the labels assigned using GSE16134, a SVM classifier was built and further applied for prediction in a validation set (GSE10334). The "scikit-learn" package (Pedregosa et al., 2011) was used to perform a grid search for the identification of the optimal hyperparameters for the SVM model using a 5-fold CV.

Evaluation of the SVM Classifier

Accuracy and area under the curve (AUC) were selected as two metrics to evaluate the performance of the SVM classifier. The percentage of accuracy was calculated as: Accuracy (%) = Predict number / Test number. A receiver operating characteristic (ROC) curve was plotted for the model using the "pROC" (Robin et al., 2011) and the "ggplot2" packages in R[5]. The AUC is the area under the ROC curve, where an AUC value above 70% is considered acceptable (Mandrekar, 2010).

Differential Expression Analysis

Differential expression analysis was performed for each of the datasets (GSE16134 and GSE10334), to identify genes discriminating between the disease and control samples, using the "Linear Models for Microarray data" ("limma") package in R (Ritchie et al., 2015). Genes with *P* value < 0.05, and $|\log \text{FC}| \geq 1$ was selected as differentially expressed genes (DEGs). The DEGs with Log FC ≥ 1 was defined as up-regulated DEGs, while the DEGs with log FC ≤ -1 were defined as down-regulated DEGs.

Differential expression analysis was also similarly conducted for the classified subtypes. Here, genes with *P* value < 0.05, and

[4] https://github.com/fchollet/keras

[5] https://ggplot2.tidyverse.org

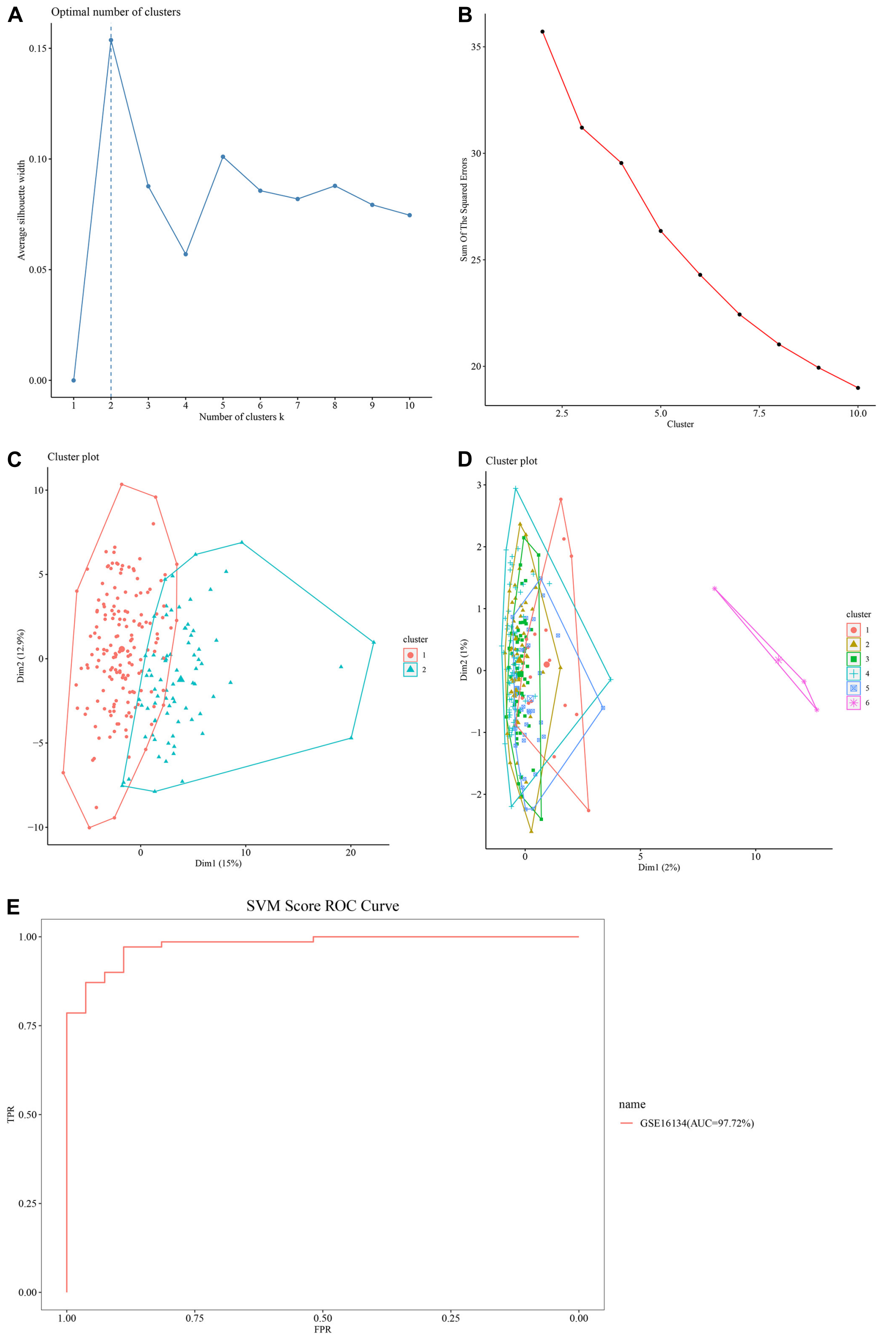


FIGURE 2 | Performance of the autoencoder (AE) and support vector machine (SVM) model. **(A)** Clustering results using the Silhouette index. Horizontal axis: Average silhouette width; Vertical axis: Number of clusters k. The optimal number of clusters is 2. **(B)** Clustering outcomes using Calinski–Harabasz criterion. Horizontal axis: Sum of the squared errors; Vertical axis: Number of clusters k. The optimal number of clusters is 2. **(C,D)** Comparison of AE with principal component analysis (PCA) based clustering. **(C)** The performance of AE based on Silhouette index. The optimum cluster number using AE is 2. Dim = dimensions. **(D)** The performance of PCA based on Silhouette index. The optimum cluster number using PCA is 6. Dim = dimensions. **(E)** Receiver operating characteristic (ROC) curve of the SVM model. Horizontal axis: false discovery rate (FDR); Vertical axis: true positive rate (TPR). The area under the curve (AUC) value of the GSE16134 test set is 97.72%.

$|\log \mathrm{FC}| \geq 0.05$ were selected as DEGs; The DEGs with Log FC ≤ 0.05 were defined as up-regulated DEGs, while the DEGs with log FC ≤ -0.05 were defined as down-regulated DEGs.

To identify the most critical immunosuppression genes in periodontitis, the DEGs discriminating disease and control samples that overlapped with DEGs discriminating the different subtypes were identified and visualized using a Venn diagram. To evaluate the performance of each such identified gene, a ROC curve was plotted as described earlier.

Functional Enrichment Analysis

The DEGs overlapping in the two datasets (GSE16134 and GSE10334) were identified using the "ClusterProfiler" package in R (Yu et al., 2012). The functions of these DEGs were explored by investigating their enriched Gene Ontology (GO) terms, particularly biological processes (BPs) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The GO/BP terms and KEGG pathways with *P* value < 0.05 were regarded as significant functions. The top 30 of the enriched GO/BPs and pathways were chosen to be visualized in a bar plot.

In addition, KEGG pathway analysis was applied to determine the characteristics of different subtypes in GSE16134 and GSE10334 each. KEGG pathways with *P* value < 0.05 were regarded as significant functions. The top 20 of the enriched pathways were listed and visualized using the heatmap function in R (Galili et al., 2017).

Construction of TF-Target DEGs Interaction Network

TF-target gene interaction pairs were downloaded from multiple databases, including TRRUST[6], cGRNB[7], HTRIdb[8], ORTI[9], and TRANSFAC[10]. The TFs targeting DEGs overlapping in the two datasets (GSE16134 and GSE10334) were extracted and used for constructing the TFs-target DEGs interaction network. The network was visualized using Cytoscape (Version 3.7.2) (Shannon et al., 2003), and the topological characteristics of the nodes in the TF-target gene network were determined.

RESULTS

Identification of Two Subtypes of Immunosuppression Genes in GSE16134 by AE

The optimal number of clusters was determined based on two metrics: Silhouette index (**Figure 2A**) and Calinski–Harabasz index (**Figure 2B**). The value of the silhouette coefficient is between $[-1, 1]$ and the score near 1 indicates a highly dense clustering. When $k = 2$, the average silhouette width was nearest to 1 (**Figure 2A**). Using Calinski–Harabasz index, better performance of clustering depends on a higher score and at $k = 2$, the score (sum of the squared errors) was the highest (**Figure 2B**). Therefore, the genes were clustered into two subtypes, defined as S1 and S2.

The AE Performed Better Compared to PCA

The performance of the AE was compared to that of PCA based clustering using Silhouette index. While two optimal clusters were extracted by AE (**Figure 2C**), six optimal clusters were extracted using PCA (**Figure 2D**), indicating that the difference between PCA transformed features was minimal, and it was difficult to cluster them effectively. Furthermore, the PCA landing points were concentrated in one zone, and the division was not clear. Therefore, the AE emerged as more effective and accurate in clustering features.

SVM Model and Its Validation

Using a 5-fold CV, the input dataset (immunosuppression genes related to periodontitis from GSE16134) were split at a 60%/40% ratio for the training set and testing set. The SVM model presented an accuracy of 92.78% (**Table 2**), and the AUC score at 97.72%, above 90% (**Figure 2E**), supporting the model was efficient in distinguishing between classes and thus reliable in predicting significant immunosuppression genes in the GSE10334 dataset (Mandrekar, 2010).

DEGs Involved in Immunosuppression and Periodontitis

Differential expression analysis was applied to the disease and control samples, as well as the two classified subtypes. A total of 236 DEGs consisting of 48 down-regulated DEGs and 188 up-regulated DEGs were identified from the GSE16134 dataset, while a total of 194 DEGs consisting of 42 down-regulated DEGs and 152 up-regulated DEGs were identified

TABLE 2 | Classifier performance outcomes of SVM.

	GSE16134		
	Test	**Predict**	**Accuracy (%)**
Cluster 1	27	21	
Cluster 2	70	69	
Total	97	90	92.78%

TABLE 3 | Outcome of differential gene expression analysis for datasets GSE16134 and GSE10334.

Data (Disease vs. Normal)	DEG (Up)	DEG (Down)	Total	Log FC Abs	*P* value
GSE16134	188	48	236	>1	<0.05
GSE10334	152	42	194	>1	<0.05

TABLE 4 | Differential expression analysis applied to disease samples based on identified subtypes.

Data (Subtype1 vs. Subtype 2)	DEG (Up)	DEG (Down)	Total	Log FC Abs	*P* value
GSE16134	134	85	219	>0.05	<0.05
GSE10334	145	95	240	>0.05	<0.05

[6] https://www.grnpedia.org/trrust/
[7] https://www.scbit.org/cgrnb
[8] http://www.lbbc.ibb.unesp.br/htri/
[9] http://orti.sydney.edu.au/about.html
[10] http://gene-regulation.com/pub/databases.html

from the GSE10334 dataset (**Table 3**). For discriminating the designated subtype labels, a total of 219 DEGs consisting of 85 down-regulated DEGs and 134 up-regulated DEGs were identified in the GSE16134, while a total of 240 DEGs consisting of 95 down-regulated DEGs and 145 up-regulated DEGs were identified in the GSE10334 dataset (**Table 4**). As shown in the Venn diagram (**Figure 3A**), three significant DEGs, Platelet Endothelial Cell Adhesion Molecule

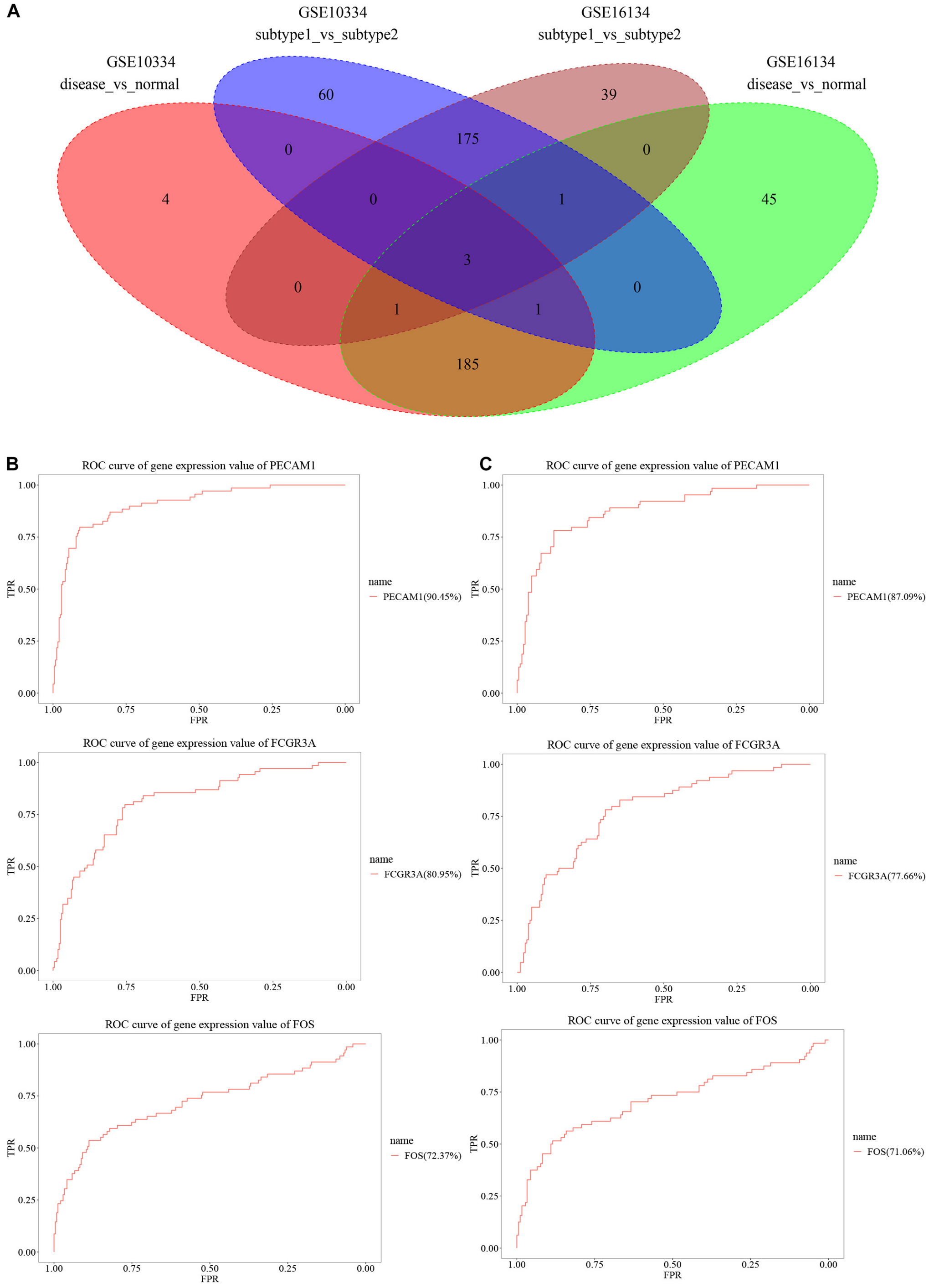


FIGURE 3 | Identification of the significant DEGs. **(A)** Intersection of DEGs discriminating sample type (disease vs. normal) (236 DEGs from GSE16134 and 194 DEGs from GSE10334) and DEGs of the disease samples classified into subtypes (subtype 1 vs. subtype 2) (219 DEGs from GSE16134 and 240 DEGs from GSE10334). **(B,C)** ROC curve of three significant genes (PECAM1, FCGR3A, and FOS) in GSE16134 **(B)** and GSE10334 **(C)**. Horizontal axis: false discovery rate (FDR); Vertical axis: true positive rate (TPR).

TABLE 5 | AUC values of the three most significant genes.

Gene	GSE10334_ROC_AUC (%)	GSE16134_ROC_AUC (%)	Mean (%)
PECAM1	87.09	90.45	88.77
FCGR3A	77.66	80.95	79.31
FOS	71.06	72.37	71.72

(PECAM) 1, Fc Gamma Receptor (FCGR) 3A, and FOS were found intersecting and considered as potentially most robust immunosuppression genes related to periodontitis. Each of the three DEGs has an acceptable performance, with an AUC value above 70%, listed in **Table 5**. The ROC curves of the three genes from GSE16134 and GSE10334 are shown in **Figures 3B,C**, respectively.

Functional Terms Enriched Among the DEGs

Significantly enriched biological processes and signaling pathways related to the immunosuppressive DEGs were identified from those overlapping between GSE16134 and GSE10334. The immunosuppressive DEGs involved in periodontitis were implicated in biological processes, including T cell activation, regulation of lymphocyte activation, regulation of T cell activation, regulation of cell-cell adhesion, and leukocyte cell-cell adhesion (**Figure 4A**). The immune activities were mainly regulated by Th17 cell differentiation, cytokine-cytokine receptor interaction, T cell receptor signaling pathway, Th1 and Th2 cell differentiation, Mitogen-activated Protein Kinase (MAPK) signaling pathway, osteoclast differentiation, and Phosphatidylinositol 3-Kinase (PI3K)-Protein Kinase B (Akt) signaling pathway (**Figure 4B**).

Most pathways of the two subtypes were evident as distinct in GSE16134 (**Figure 5**), indicating significant differences between the two subtypes in terms of immunosuppressive activities in periodontitis. This difference was also detected between the two predicted subtypes in GSE10334 (**Figure 6**). Specifically, subtype S1 of immunosuppressive DEGs in periodontitis from both GSE16134 (**Figure 5A**) and GSE10334 (**Figure 6A**) was mainly enriched in cytokine-cytokine receptor interaction, chemokine signaling pathway, Janus kinase (JAK)- Signal Transducer and Activator of Transcription Protein (STAT) signaling pathway, Hypoxia-inducible Factor (HIF)-1 signaling pathway, and T cell receptor signaling pathway. Of note, subtype S1 from GSE16134 was also enriched in PD-L1 expression and PD-1 checkpoint pathway in cancer (**Figure 5A**). Whereas subtype S2 was mainly associated with MAPK signaling pathway, osteoclast differentiation, and infection of virus and *E. coli* bacteria (**Figures 5B**, **6B**).

Identification of Hub Transcription Factors That Targeted DEGs

The TFs-target DEGs interaction network of the immunosuppression genes in periodontitis is shown in **Figure 7**, consisting of 197 nodes and 447 edges. Top 30 TFs (**Table 6**) with the highest degree were considered to represent those most critical to this network. Of these, the top 10 TFs in the network were determined as the hubs, including Androgen Receptor (AR), Hypoxia-inducible Factor (HIF)1A, Signal Transducer and Activator of Transcription Protein (STAT) 5B, and STAT4, which were not only TFs but also up-regulated DEGs, and Nuclear Factor Kappa B Subunit 1 (NFKB1), MYC, JUN, Tumor Protein (TP)53, FOS, and Forkhead Box (FOX) O3, which were not only TFs but also down-regulated DEGs.

DISCUSSION

In this study, we used a DL-based algorithm, the AE, for identifying the pivotal immunosuppression genes relevant to periodontitis. With this approach, we re-constructed multi-omics data and produced representative molecular features grouped into two immune subtypes and then built an SVM model based on these, which was confirmed using a validation set. Besides, significant pathways and TF-target DEGs involved in immunosuppression during periodontitis were identified.

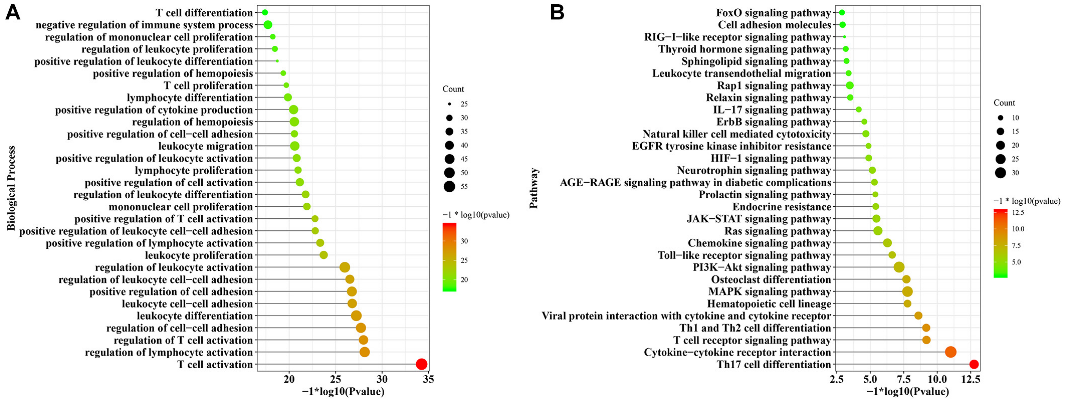


FIGURE 4 | The functional enrichment analysis of the overlapping DEGs common to the two datasets (GSE16134 and GSE10334). **(A)** The significantly enriched biological processes of the overlapped DEGs; **(B)** The significantly enriched signaling pathways of the overlapped DEGs.

A

Pathway	GeneRatio	Pvalue
Cytokine-cytokine receptor interaction	24/98	4.91735E-14
Viral protein interaction with cytokine and cytokine receptor	13/98	1.61152E-10
JAK-STAT signaling pathway	12/98	5.08937E-07
Th1 and Th2 cell differentiation	9/98	1.47271E-06
Chemokine signaling pathway	12/98	3.09937E-06
Th17 cell differentiation	9/98	5.23398E-06
Non-small cell lung cancer	7/98	2.39428E-05
Hematopoietic cell lineage	8/98	2.4112E-05
Natural killer cell mediated cytotoxicity	9/98	2.71695E-05
T cell receptor signaling pathway	8/98	3.45203E-05
Primary immunodeficiency	5/98	8.62334E-05
Neuroactive ligand-receptor interaction	13/98	0.000221902
PD-L1 expression and PD-1 checkpoint pathway in cancer	6/98	0.000701566
African trypanosomiasis	4/98	0.000985574
PI3K-Akt signaling pathway	12/98	0.001115701
Measles	7/98	0.001434426
HIF-1 signaling pathway	6/98	0.002013813
Intestinal immune network for IgA production	4/98	0.002830071
Hepatitis C	7/98	0.002870916
Malaria	4/98	0.003047945

B

Pathway	GeneRatio	Pvalue
MAPK signaling pathway	18/68	1.81E-11
Chronic myeloid leukemia	11/68	2.39E-11
Neurotrophin signaling pathway	12/68	2.14E-10
Human T-cell leukemia virus 1 infection	15/68	2.38E-10
Pancreatic cancer	10/68	5.49E-10
Colorectal cancer	10/68	1.91E-09
Hepatitis B	12/68	7.67E-09
Th17 cell differentiation	10/68	1.65E-08
Fluid shear stress and atherosclerosis	11/68	1.73E-08
Osteoclast differentiation	10/68	9.26E-08
PI3K-Akt signaling pathway	15/68	1.65E-07
Ras signaling pathway	12/68	4.11E-07
Endometrial cancer	7/68	4.64E-07
Thyroid cancer	6/68	5.35E-07
Cellular senescence	10/68	5.95E-07
Pathogenic Escherichia coli infection	11/68	6.18E-07
Shigellosis	12/68	7.71E-07
Epstein-Barr virus infection	11/68	7.94E-07
Salmonella infection	12/68	8.77E-07
Proteoglycans in cancer	11/68	9.19E-07

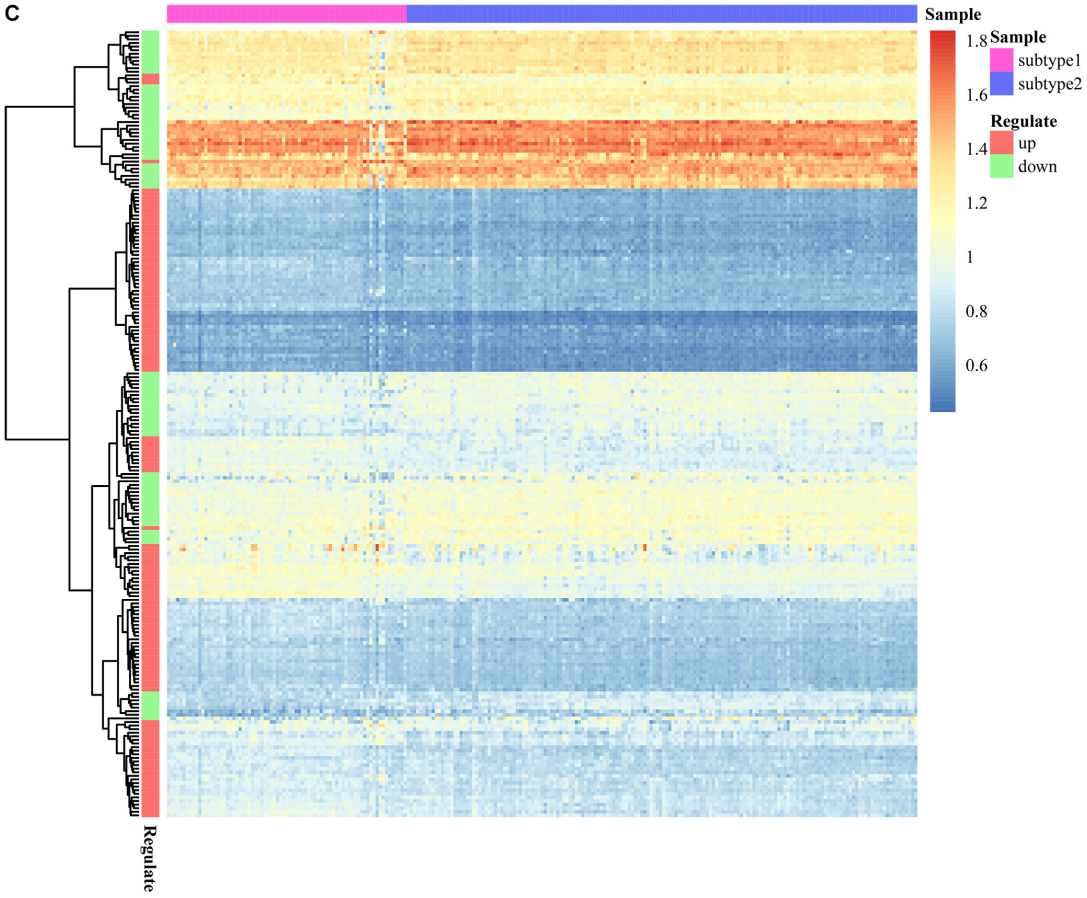
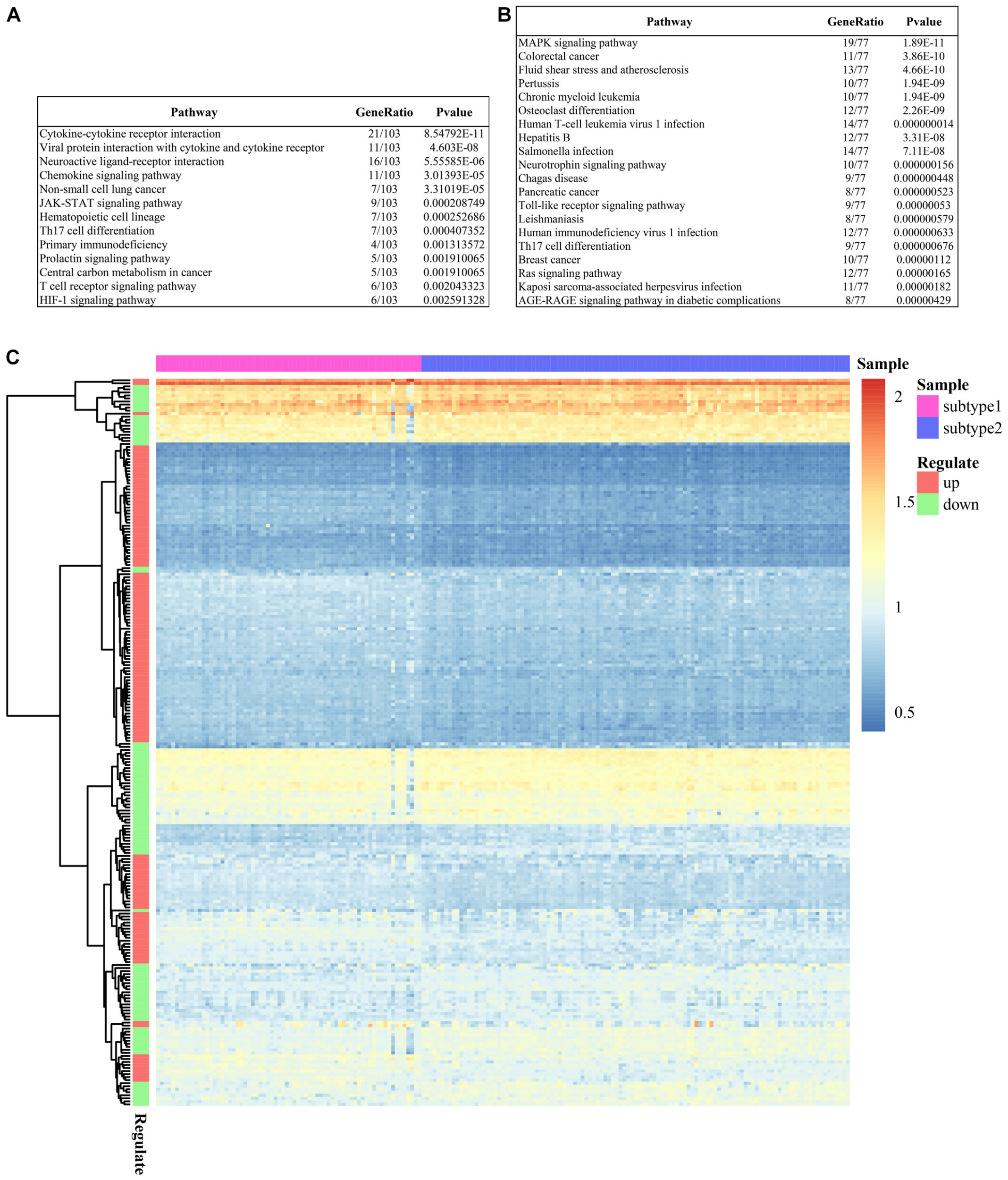


FIGURE 5 | Pathways enriched in the DEGs characterizing the two subtypes in GSE16134. **(A)** Top 20 enriched signaling pathways of DEGs in subtype 1. **(B)** Top 20 enriched signaling pathways of DEGs in subtype 2. **(C)** Heatmap shows the enriched signaling pathways of DEGs in the two subtypes.

Notably, we identified the key characteristics of two immune subtypes of periodontitis. We also identified three "master" immunosuppression genes, PECAM1, FCGR3A, and FOS, as candidate genes central to immune suppressive pathogenic mechanisms in periodontitis.

An AE-based DL approach has demonstrated high efficiency and accuracy in predicting biomarker genes for lung cancer, breast cancer, and HNSCC (Xiao et al., 2018). Akin to these studies, CV results indicated this approach was robust in classifying patients into two subgroups. Furthermore, the AE was more efficient and precise in clustering the distinct features, as compared with the commonly utilized unsupervised ordination method, PCA. In addition, the robustness and reliability of the model were confirmed in a validation set.

The central finding of our study is the identification of three distinct immunosuppression genes, PECAM1, FCGR3A, and FOS, which could be potentially high-value biomarkers or candidate therapeutic targets for periodontitis. PECAM1, also known as CD31, is an immunoglobulin (Ig) gene expressed



Pathway	GeneRatio	Pvalue
Cytokine-cytokine receptor interaction	21/103	8.54792E-11
Viral protein interaction with cytokine and cytokine receptor	11/103	4.603E-08
Neuroactive ligand-receptor interaction	16/103	5.55585E-06
Chemokine signaling pathway	11/103	3.01393E-05
Non-small cell lung cancer	7/103	3.31019E-05
JAK-STAT signaling pathway	9/103	0.000208749
Hematopoietic cell lineage	7/103	0.000252686
Th17 cell differentiation	7/103	0.000407352
Primary immunodeficiency	4/103	0.001313572
Prolactin signaling pathway	5/103	0.001910065
Central carbon metabolism in cancer	5/103	0.001910065
T cell receptor signaling pathway	6/103	0.002043323
HIF-1 signaling pathway	6/103	0.002591328

Pathway	GeneRatio	Pvalue
MAPK signaling pathway	19/77	1.89E-11
Colorectal cancer	11/77	3.86E-10
Fluid shear stress and atherosclerosis	13/77	4.66E-10
Pertussis	10/77	1.94E-09
Chronic myeloid leukemia	10/77	1.94E-09
Osteoclast differentiation	12/77	2.26E-09
Human T-cell leukemia virus 1 infection	14/77	0.000000014
Hepatitis B	12/77	3.31E-08
Salmonella infection	14/77	7.11E-08
Neurotrophin signaling pathway	10/77	0.000000156
Chagas disease	9/77	0.000000448
Pancreatic cancer	8/77	0.000000523
Toll-like receptor signaling pathway	9/77	0.00000053
Leishmaniasis	8/77	0.000000579
Human immunodeficiency virus 1 infection	12/77	0.000000633
Th17 cell differentiation	9/77	0.000000676
Breast cancer	10/77	0.00000112
Ras signaling pathway	12/77	0.00000165
Kaposi sarcoma-associated herpesvirus infection	11/77	0.00000182
AGE-RAGE signaling pathway in diabetic complications	8/77	0.00000429

FIGURE 6 | Pathways enriched in the DEGs characterizing the two subtypes in GSE10334. **(A)** Top 20 enriched signaling pathways of DEGs in subtype1. **(B)** Top 20 enriched signaling pathways of DEGs in subtype 2. **(C)** Heatmap shows the enriched signaling pathways of DEGs in the two subtypes.

in various cells, such as endothelial cells (ECs), platelets, and immune cells. PECAM1 is found to be a co-modulator of T-cell immunity (Huang et al., 2017) and a promoter of endothelial junctional integrity (Marelli-Berg et al., 2013). Periodontal pathogens, particularly *P. gingivalis*, can induce vascular damage through the degradation of PECAM1 (Yun et al., 2005; Farrugia et al., 2020). A protective effect of PECAM1 was also detected in transplant arteriosclerosis (Ensminger et al., 2002). FCGR3A is a member of FCGR families, forming a critical link between humoral and cellular immune responses to periodontal microbiota (Chai et al., 2010; Pavkovic et al., 2018). Previous studies have reported single-nucleotide polymorphisms (SNPs) of FCGR3A (rs396991 and rs4455090) were correlated with periodontitis and might impact susceptibility to periodontitis (Kobayashi et al., 2001; Chai et al., 2010). Besides, FCGR3A polymorphism and the allele rs396991 was identified as an independent susceptibility marker of allograft rejection in patients after organ transplants, highly responsive to natural killer (NK) cells (Paul et al., 2019). FOS was also identified as a significant TF in the study.

Of the top 10 hub TFs, six "leader" immunosuppressive TF-target DEGs with plausible literature evidence were identified

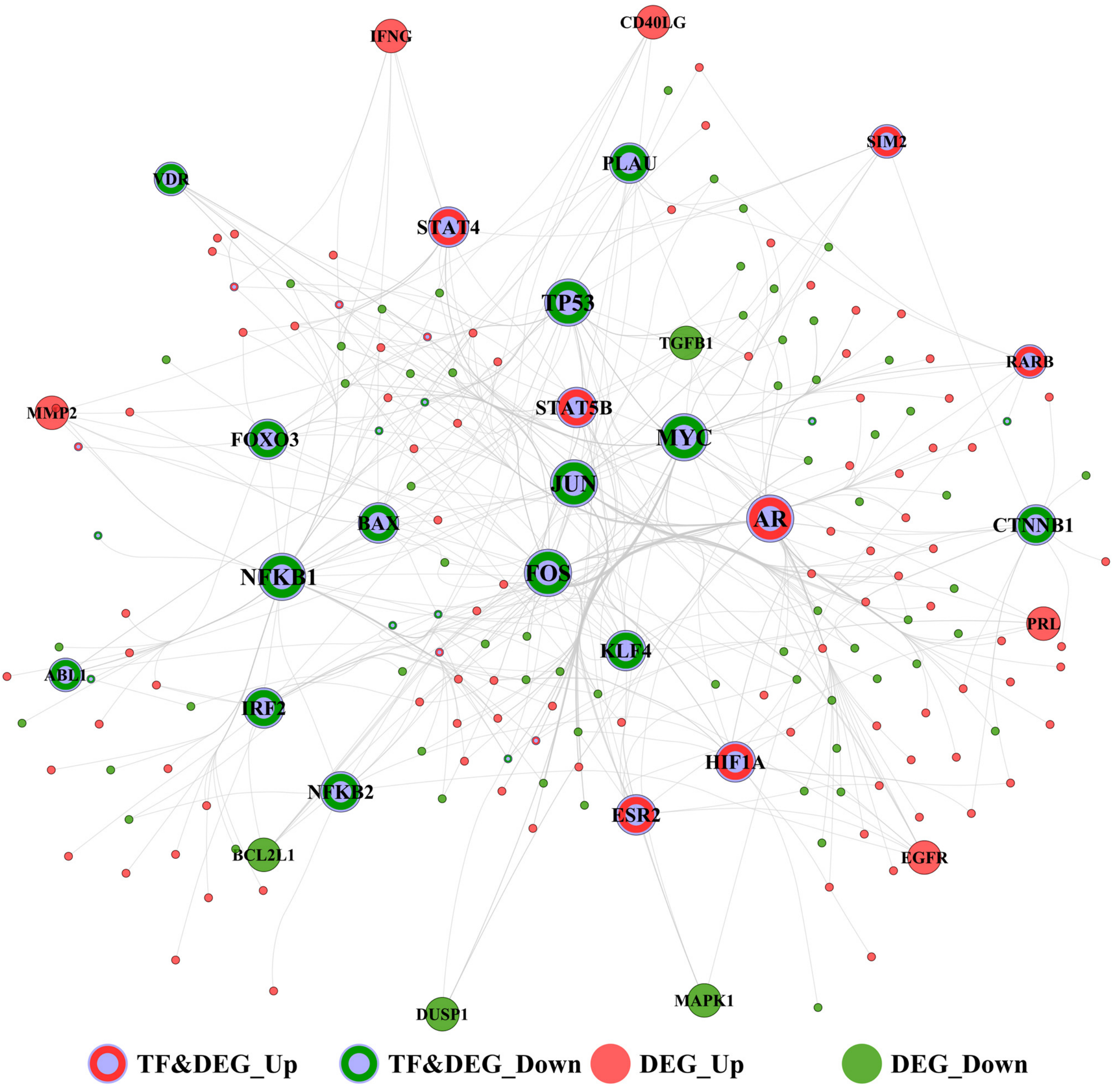


FIGURE 7 | The transcription factor (TF)-target interaction network of GSE16134 and GSE10334 involved in immunosuppression and periodontitis. Top 30 TFs were visualized in the network. Red and gray dots: up-regulated TF and DEG; Green and gray dots: down-regulated TF and DEG; Red dots: up-regulated DEG; Green dots: down-regulated DEG.

as key to periodontitis pathogenesis and included the down-regulated TFs (NFKB1, FOS, and JUN), as well as up-regulated TFs (HIF1A, STAT5B, and STAT4). NFKB1, also termed NF-κB, is a core TF implicated in immune and inflammatory diseases (Tak and Firestein, 2001). Periodontal pathogens can activate NF-κB, and thus inhibition of NF-κB might be a therapeutic target for periodontitis (Ambili et al., 2005). Furthermore, NF-κB is activated in transplanted tissue, and its blockade may be potent in preventing allograft rejection after solid organ transplants, considering the role of NF-κB in T cell activation and differentiation (Molinero and Alegre, 2012). FOS is implicated in periodontitis progression acting via the regulation of T-cell receptor (TCR) signaling (Maekawa et al., 2017). C-Jun (encoded by JUN) signaling is activated by Receptor Activator of Nuclear Factor Kb Ligand (RANKL) and essential for osteoclast differentiation (Ikeda et al., 2004). Activator Protein (AP)-1 is a heterodimer composed of the Fos and Jun subunits, which downregulates osteoprotegerin and

TABLE 6 | The topological characteristics of the top 30 nodes in the TF-target interaction network.

Name	Label	Degree	Average Shortest Path Length	Betweenness Centrality	Closeness Centrality	Clustering Coefficient	Topological Coefficient
AR	TF&DEG_Up	87	1.6327	0.4063	0.6125	0.0270	0.0351
NFKB1	TF&DEG_Down	62	1.7143	0.2926	0.5833	0.0518	0.0457
MYC	TF&DEG_Down	56	1.7398	0.2342	0.5748	0.0610	0.0497
JUN	TF&DEG_Down	49	1.8163	0.1276	0.5506	0.0859	0.0541
TP53	TF&DEG_Down	41	2.0816	0.0906	0.4804	0.0634	0.0603
FOS	TF&DEG_Down	40	1.8776	0.0975	0.5326	0.1077	0.0670
HIF1A	TF&DEG_Up	17	2.0051	0.0433	0.4987	0.1985	0.1076
STAT5B	TF&DEG_Up	16	2.5408	0.0063	0.3936	0.0917	0.1517
FOXO3	TF&DEG_Down	15	2.3418	0.0175	0.4270	0.2000	0.1202
STAT4	TF&DEG_Up	15	2.0459	0.0329	0.4888	0.2190	0.1298
CTNNB1	TF&DEG_Down	14	2.1837	0.0514	0.4579	0.1648	0.1186
BAX	TF&DEG_Down	13	2.2857	0.0264	0.4375	0.1795	0.1348
KLF4	TF&DEG_Down	13	1.9949	0.0268	0.5013	0.2821	0.1560
IRF2	TF&DEG_Down	13	2.5918	0.0283	0.3858	0.0385	0.1110
ESR2	TF&DEG_Up	11	2.2092	0.0060	0.4527	0.3455	0.1706
NFKB2	TF&DEG_Down	10	2.3827	0.0182	0.4197	0.2889	0.1806
PLAU	TF&DEG_Down	10	2.4694	0.0140	0.4050	0.1556	0.1538
EGFR	DEG_Up	9	2.1327	0.0037	0.4689	0.4167	0.2059
VDR	TF&DEG_Down	9	2.4388	0.0030	0.4100	0.2500	0.1993
RARB	TF&DEG_Up	8	2.2245	0.0083	0.4495	0.3571	0.2083
BCL2L1	DEG_Down	7	2.1531	0.0021	0.4645	0.6190	0.2706
SIM2	TF&DEG_Up	7	2.2245	0.0045	0.4495	0.2857	0.2351
ABL1	TF&DEG_Down	6	2.2959	0.0029	0.4356	0.3333	0.2561
TGFB1	DEG_Down	6	2.0816	0.0005	0.4804	0.8667	0.3209
PRL	DEG_Up	6	2.4286	0.0014	0.4118	0.4000	0.2434
CD40LG	DEG_Up	6	2.4694	0.0031	0.4050	0.4000	0.2508
IFNG	DEG_Up	6	2.5918	0.0013	0.3858	0.3333	0.2405
MMP2	DEG_Up	6	2.4031	0.0003	0.4161	0.8000	0.2670
DUSP1	DEG_Down	5	2.2245	0.0022	0.4495	0.4000	0.3270
MAPK1	DEG_Down	5	2.3061	0.0009	0.4336	0.5000	0.3090

is highly expressed in periodontal ligament cells, suggesting their role in bone resorption during periodontitis (Suda et al., 2009). Inhibition of c-Fos/AP-1 by T-5224 (a novel chemical) could attenuate inflammation, T cell proliferation, and allograft rejection in pancreatic islet transplantation (PIT) (Yoshida et al., 2015) and be suggested as a target for immunosuppressive therapy. HIF1A/HIF1, an oxygen-regulated subunit (Corrado and Fontana, 2020), is involved in the immune response of periodontitis, playing a pleitropic role in defending against macrobiotics and facilitating the progression of periodontitis (Wang et al., 2017). HIF1 was also suggested to mediate inflammation and immune responses after organ transplantation, mediating angiogenesis and allograft in the donor organs (Xu et al., 2019). STAT5B and STAT4 are members of the STAT family that play important roles in activating gene transcription through various cytokines. STAT5B and STAT4 can be activated by a variety of cytokines, including Interleukin (IL)12, Type I Interferon (IFNI), IL23, IL2, IL27, and IL35 (Garcia de Aquino et al., 2009; Sanpaolo et al., 2020; Yang et al., 2020), which are prominently involved in mediating immune responses during periodontitis. IFN-γ could stimulate the expression of Indoleamine 2,3-Dioxygenase (IDO)1, a critical immunosuppression protein, in primary human periodontal ligament stem cells (Andrukhov et al., 2017). Thus, evidence suggests STAT5B and STAT4 may mediate immunosuppression during periodontitis.

The immunosuppression DEGs in the two subtypes were functionally related to multiple immune-related biological processes and pathways, and the two subtypes were distinct in their regulating pathways. In subtype S1, PD1/PLL1 checkpoint signaling, T cell receptor signaling, and signaling pathways related to immunosuppressive factors, including cytokines, chemokines, Janus Kinase (JAK) -STAT, and HIF1, are found to activate up-regulated TFs, such as HIF1A, STAT4, and STAT5B (de Souza et al., 2012). Whereas the signaling pathways enriched in subtype S2 primarily regulated the MAPK signaling pathway and osteoclast differentiation, as well as the infection of virus and *E. coli* bacteria, targeting the down-regulated TFs, such as NFKB1, FOS, and JUN (de Souza et al., 2012). Immune response-related pathways were mainly involved in the subtype S1, supporting a hypothesis that periodontitis patients with molecular subtype S1 may be more sensitive to and thus respond comparatively well to the immune-related target therapy.

Considering PDL1/PD1 signaling that characterized the subtype S1, it has been found that peptidoglycans from *P. gingivalis* can lead to the up-regulation of PDL1 expressed by gingival keratinocytes, as well as the overexpression of PD1 expressed on T lymphocytes (Bailly, 2020). The interaction between PDL1 and PD1 can suppress the initial activation and effector function of T cells and thereby promote the progression of periodontal inflammation (Yang et al., 2019). As PDL1-inhibitor has shown significant effects as a cancer therapy in clinical trials (Kim et al., 2020), it may also hold potential as immune therapy for periodontitis patients, especially in the case of immune-compromised patients. The inhibition of the JAK-STAT pathway has been indicated as a potential strategy for immunosuppression therapy, targeting the key cytokines, such as IFNg and IL12 (O'Shea and Plenge, 2012). HIF1A pathway has been found to modulate immunosuppressive molecules, typically VEGF, in periodontitis (Vasconcelos et al., 2016), and tumor microenvironment (El-Sayed Mohammed Youssef et al., 2015). Manipulation of the HIF1A pathway has been proposed as a therapeutic intervention in tumor immunotherapy (Li et al., 2018). The MAPK pathway identified in subtype S2, consists of three family sub-members, extracellular regulated kinases (ERK), c-Jun N-terminal activated kinases (JNK), and p38, and is closely related to osteoblast differentiation (Rodríguez-Carballo et al., 2016). Further, inhibition of p38 may particularly have potential therapeutic value in limiting periodontitis progression at multiple levels of the immune response via its effects on different extracellular stimuli (Kirkwood and Rossa, 2009). Of note, bone resorption, a hallmark of periodontitis, is mainly affected through RANKL, a vital osteoclast differentiation factor (Taubman et al., 2005) and Tumor Necrosis Factor (TNF)-a, majorly activated by MAPK and NF-κB pathways (Ketherin and Sandra, 2018), indicating a key role of these pathways in osteoimmunology.

Altogether, using the DL-based predictive model and bioinformatic analysis, our study provides a predictive and theoretical description of functions and mechanisms relevant to immunosuppression genes active in periodontitis pathogenesis. The validated efficiency and accuracy of the DL-model overcome the bottlenecks of current evidence and suggest new insights valuable for potential translation in therapeutic gene targeting. However, considering our study is the first to apply DL methods in the periodontal disease context, it is expected that further well-designed investigations can validate the model considering other aspects of periodontal disease, where specific and precise associations between clinical parameters and target genes might be identified. One caveat of our study is the lack of phenotype information about the periodontitis cases which were grouped into two distinct immune subtypes. Periodontitis is well recognized as a multifactorial disease, where a disease phenotype may result from multiple factors in a "sufficient cause model" (Heaton and Dietrich, 2012). Distinct "immunotypes" in periodontitis may represent heterogeneity in the core biological mechanisms contributing to disease in different subjects. A more in-depth understanding of these could support precision medicine approaches in the future. Besides, the possible clinical translation of these results may include multiple directions. For instance, the identification of immunosuppression genes may direct the development of improved topical drugs for delivery at diseased periodontal sites, which could avoid side effects inherent to conventional drugs such as antimicrobials. Also, these findings support a hypothesis that manipulation of the identified immunosuppression genes or selection of the drugs targeting immune checkpoints could be protective against periodontal diseases in patients who have had long-time immunosuppressive therapy, such as those with organ transplantation.

CONCLUSION

The DL-based model applied in this study was reliable and robust in predicting immunosuppression genes in periodontitis. An array of pathways and TF-target DEGs were found to be implicated in the immunosuppressive activity during periodontitis. Three "master" immunosuppression genes, PECAM1, FCGR3A, and FOS, were identified as critical to immune suppression occurring during periodontal pathology. Taken together, the DL model revealed novel insights into the molecular mechanisms underpinning periodontitis and identified key candidate genes for further translation in the context of risk profiling and therapeutic development.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

WN conceived of the presented idea, designed the overall study workflow, analyzed the data, prepared the figures and tables, authored and reviewed drafts of the manuscript, and approved the final draft. AA prepared the figures and tables, was involved in the proofreading and deep editing, and approved the final draft. ZS analyzed the data, prepared the figures and tables, and was involved in the discussion of the results. AO analyzed the data, prepared the figures and tables, and was involved in proofreading and deep editing. CL, SH, QO, and MZ were involved in the discussion of the results, and also prepared the figures and tables. XL and YD designed the overall study workflow, analyzed the data, and prepared the figures and tables. RH, DZ, and GS devised the main conceptual idea, supervised the whole work, and approved the final draft. GP, YW, and XH supervised the whole project, and approved the final draft. All authors contributed to the article and approved the submitted version.

REFERENCES

Alvarez, C., Rojas, C., Rojas, L., Cafferata, E. A., Monasterio, G., and Vernal, R. (2018). Regulatory T lymphocytes in periodontitis: a translational view. *Mediators Inflamm.* 2018:7806912. doi: 10.1155/2018/7806912

Ambili, R., Santhi, W. S., Janam, P., Nandakumar, K., and Pillai, M. R. (2005). Expression of activated transcription factor nuclear factor-kappaB in periodontally diseased tissues. *J. Periodontol.* 76, 1148–1153.

Andrukhov, O., Hong, J. S. A., Andrukhova, O., Blufstein, A., Moritz, A., and Rausch-Fan, X. (2017). Response of human periodontal ligament stem cells to IFN-γ and TLR-agonists. *Sci. Rep.* 7:12856. doi: 10.1038/s41598-017-12480-7

Aoyagi, T., Yamazaki, K., Kabasawa-Katoh, Y., Nakajima, T., Yamashita, N., Yoshie, H., et al. (2000). Elevated CTLA-4 expression on CD4 T cells from periodontitis patients stimulated with *Porphyromonas gingivalis* outer membrane antigen: CTLA-4 expression in periodontitis. *Clin. Exp. Immunol.* 119, 280–286. doi: 10.1046/j.1365-2249.2000.01126.x

Bailly, C. (2020). The implication of the PD-1/PD-L1 checkpoint in chronic periodontitis suggests novel therapeutic opportunities with natural products. *Jpn. Dent. Sci Rev.* 56, 90–96. doi: 10.1016/j.jdsr.2020.04.002

Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat. Theory Methods* 3, 1–27. doi: 10.1080/03610927408827101

Cebeci, I., Kantarci, A., Firatli, E., Aygun, S., Tanyeri, H., Aydin, A. E., et al. (1996). Evaluation of the frequency of HLA determinants in patients with gingival overgrowth induced by cyclosporine-A. *J. Clin. Periodontol.* 23, 737–742. doi: 10.1111/j.1600-051X.1996.tb00603.x

Cekici, A., Kantarci, A., Hasturk, H., and Van Dyke, T. E. (2014). Inflammatory and immune pathways in the pathogenesis of periodontal disease: inflammatory and immune pathways in periodontal disease. *Periodontol. 2000* 64, 57–80. doi: 10.1111/prd.12002

Chai, L., Song, Y.-Q., Zee, K.-Y., and Leung, W. K. (2010). SNPs of Fc-gamma receptor genes and chronic periodontitis. *J. Dent. Res.* 89, 705–710. doi: 10.1177/0022034510365444

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

Chicco, D., Sadowski, P., and Baldi, P. (2014). "Deep autoencoder neural networks for gene ontology annotation predictions," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, (Newport Beach, CA: ACM), 533–540. doi: 10.1145/2649387.2649442

Corrado, C., and Fontana, S. (2020). Hypoxia and HIF signaling: one axis with divergent effects. *Int. J. Mol. Sci.* 21:5611. doi: 10.3390/ijms21165611

Cota, L. O. M., Aquino, D. R., Franco, G. C. N., Cortelli, J. R., Cortelli, S. C., and Costa, F. O. (2010). Gingival overgrowth in subjects under immunosuppressive regimens based on cyclosporine, tacrolimus, or sirolimus: risk variables for gingival overgrowth. *J. Clin. Periodontol.* 37, 894–902. doi: 10.1111/j.1600-051X.2010.01601.x

de Souza, J. A. C., Junior, C. R., Garlet, G. P., Nogueira, A. V. B., and Cirelli, J. A. (2012). Modulation of host cell signaling pathways as a therapeutic approach in periodontal disease. *J. Appl. Oral Sci.* 20, 128–138. doi: 10.1590/S1678-77572012000200002

Dutzan, N., Konkel, J. E., Greenwell-Wild, T., and Moutsopoulos, N. M. (2016). Characterization of the human immune cell network at the gingival barrier. *Mucosal. Immunol.* 9, 1163–1172. doi: 10.1038/mi.2015.136

El-Sayed Mohammed Youssef, H., Eldeen Abo-Azma, N. E., and Eldeen Megahed, E. M. (2015). Correlation of hypoxia-inducible factor-1 alpha (HIF-1α) and vascular endothelial growth factor (VEGF) expressions with clinico-pathological features of oral squamous cell carcinoma (OSCC). *Tanta Dent. J.* 12, S1–S14. doi: 10.1016/j.tdj.2015.05.010

Ensminger, S. M., Spriewald, B. M., Steger, U., Morris, P. J., Mak, T. W., and Wood, K. J. (2002). Platelet-endothelial cell adhesion molecule-1 (CD31) expression on donor endothelial cells attenuates the development of transplant arteriosclerosis. *Transplantation* 74, 1267–1273. doi: 10.1097/00007890-200211150-00012

Farrugia, C., Stafford, G. P., Potempa, J., Wilkinson, R. N., Chen, Y., Murdoch, C., et al. (2020). Mechanisms of vascular damage by systemic dissemination of the oral pathogen *Porphyromonas gingivalis*. *FEBS J.* 15486. doi: 10.1111/febs.15486

Galili, T., O'Callaghan, A., Sidi, J., and Sievert, C. (2017). heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* 34, 1600–1602. doi: 10.1093/bioinformatics/btx657

Garcia de Aquino, S., Manzolli Leite, F. R., Stach-Machado, D. R., Francisco da Silva, J. A., Spolidorio, L. C., and Rossa, C. (2009). Signaling pathways associated with the expression of inflammatory mediators activated during the course of two models of experimental periodontitis. *Life Sci.* 84, 745–754. doi: 10.1016/j.lfs.2009.03.00

Gemmell, E., Yamazaki, K., and Seymour, G. J. (2002). Destructive periodontitis lesions are determined by the nature of the lymphocytic response. *Crit. Rev. Oral Biol. Med.* 13, 17–34. doi: 10.1177/154411130201300104

Heaton, B., and Dietrich, T. (2012). Causal theory and the etiology of periodontal diseases. *Periodontol.* 2000, 26–36. doi: 10.1111/j.1600-0757.2011

Huang, F., Chen, M., Chen, W., Gu, J., Yuan, J., Xue, Y., et al. (2017). Human gingiva-derived mesenchymal stem cells inhibit xeno-graft-versus-host disease via CD39–CD73–adenosine and IDO signals. *Front. Immunol.* 8:68. doi: 10.3389/fimmu.2017.00068

Ikeda, F., Nishimura, R., Matsubara, T., Tanaka, S., Inoue, J., Reddy, S. V., et al. (2004). Critical roles of c-Jun signaling in regulation of NFAT family and RANKL-regulated osteoclast differentiation. *J. Clin. Invest.* 114, 475–484. doi: 10.1172/JCI200419657

Ju, Y., Guo, J., and Liu, S. (2015). "A deep learning method combined sparse autoencoder with SVM," in *Proceedings of the 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, (Xi'an: IEEE), 257–260. doi: 10.1109/CyberC.2015.39

Ketherin, K., and Sandra, F. (2018). Osteoclastogenesis in periodontitis: signaling pathway. Synthetic and natural inhibitors. *Mol. Cell. Biomed. Sci.* 2:11. doi: 10.21705/mcbs.v2i1.16

Kim, H., Kwon, M., Kim, B., Jung, H. A., Sun, J.-M., Lee, S.-H., et al. (2020). Clinical outcomes of immune checkpoint inhibitors for patients with recurrent or metastatic head and neck cancer: real-world data in Korea. *BMC Cancer* 20:727. doi: 10.1186/s12885-020-07214-4

Kirkwood, K. L., and Rossa, C. Jr. (2009). The potential of p38 MAPK inhibitors to modulate periodontal infections. *Curr. Drug Metab.* 10, 55–67. doi: 10.2174/138920009787048347

Kobayashi, T., Yamamoto, K., Sugita, N., van der Pol, W. L., Yasuda, K., Kaneko, S., et al. (2001). The Fc gamma receptor genotype as a severity factor for chronic periodontitis in Japanese patients. *J. Periodontol.* 72, 1324–1331.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning – ICML '09*, (Montreal, QC: ACM Press), 1–8. doi: 10.1145/1553374.1553453

Li, Y., Patel, S. P., Roszik, J., and Qin, Y. (2018). Hypoxia-Driven immunosuppressive metabolites in the tumor microenvironment: new approaches for combinational immunotherapy. *Front. Immunol.* 9:1591. doi: 10.3389/fimmu.2018.01591

Maekawa, T., Kulwattanaporn, P., Hosur, K., Domon, H., Oda, M., Terao, Y., et al. (2017). Differential expression and roles of secreted frizzled-related protein 5 and the wingless homolog Wnt5a in periodontitis. *J. Dent. Res.* 96, 571–577. doi: 10.1177/0022034516687248

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* 5, 1315–1316. doi: 10.1097/jto.0b013e3181ec173d

Marelli-Berg, F. M., Clement, M., Mauro, C., and Caligiuri, G. (2013). An immunologist's guide to CD31 function in T-cells. *J. Cell Sci.* 126, 2343–2352. doi: 10.1242/jcs.124099

Meyle, J., and Chapple, I. (2015). Molecular aspects of the pathogenesis of periodontitis. *Periodontol. 2000* 69, 7–17. doi: 10.1111/prd.12104

Molinero, L. L., and Alegre, M.-L. (2012). Role of T cell–nuclear factor κB in transplantation. *Transplant. Rev.* 26, 189–200. doi: 10.1016/j.trre.2011.07.005

O'Shea, J. J., and Plenge, R. (2012). JAK and STAT signaling molecules in immunoregulation and immune-mediated disease. *Immunity* 36, 542–550. doi: 10.1016/j.immuni.2012.03.014

Paul, P., Pedini, P., Lyonnet, L., Di Cristofaro, J., Loundou, A., and Pelardy, M. (2019). FCGR3A and FCGR2A genotypes differentially impact allograft rejection and patients' survival after lung transplant. *Front. Immunol.* 10:1208. doi: 10.3389/fimmu.2019.01208

Pavkovic, M., Petlichkovski, A., Karanfilski, O., Cevreska, L., and Stojanovic, A. (2018). FC gamma receptor polymorphisms in patients with immune thrombocytopenia. *Hematology* 23, 163–168. doi: 10.1080/10245332.2017.1377902

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

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

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

Rodríguez-Carballo, E., Gámez, B., and Ventura, F. (2016). p38 MAPK signaling in osteoblast differentiation. *Front. Cell. Dev. Biol.* 4:40. doi: 10.3389/fcell.2016.00040

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Sanpaolo, E.R., Rotondo, C., Cici, D., Corrado, A., and Cantatore, F.P. (2020). JAK/STAT pathway and molecular mechanism in bone remodeling. *Mol. Biol. Rep.* 47, 9087–9096. doi: 10.1007/s11033-020-05910-9

Scott, D. A., and Krauss, J. (2011). "Neutrophils in periodontal inflammation," in *Frontiers of Oral Biology*, eds D. F. Kinane and A. Mombelli (Basel: KARGER), 56–83. doi: 10.1159/000329672

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

Suda, T., Nagasawa, T., Wara-aswapati, N., Kobayashi, H., Iwasaki, K., Yashiro, R., et al. (2009). Regulatory roles of β-catenin and AP-1 on osteoprotegerin production in interleukin-1α-stimulated periodontal ligament cells. *Oral Microbiol. Immunol.* 24, 384–389. doi: 10.1111/j.1399-302X.2009.00529.x

Tak, P. P., and Firestein, G. S. (2001). NF-κB: a key role in inflammatory diseases. *J. Clin. Invest.* 107, 7–11. doi: 10.1172/jci11830

Tan, J., Ung, M., Cheng, C., and Greene, C. S. (2014). "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders," in *Biocomputing 2015*, (Kohala Coast, HI: World Scientific), 132–143. doi: 10.1142/9789814644730_0014

Taubman, M. A., Valverde, P., Han, X., and Kawai, T. (2005). Immune response: the key to bone resorption in periodontal disease. *J. Periodontol.* 76, 2033–2041. doi: 10.1902/jop.2005.76.11-s.2033

Vasconcelos, R. C., Costa, A. D. L. L., Freitas, R. D. A., Bezerra, B. A. D. A., Santos, B. R. M. D., Pinto, L. P., et al. (2016). Immunoexpression of HIF-1α and VEGF in periodontal disease and healthy gingival tissues. *Braz. Dent. J.* 27, 117–122. doi: 10.1590/0103-6440201600533

Wang, X. X., Chen, Y., and Leung, W. K. (2017). "Role of the hypoxia-inducible factor in periodontal inflammation," in *Hypoxia and Human Diseases*, eds J. Zheng and C. Zhou (Rijeka: Intech Open), 285–302.

Wang, Y., Yao, H., and Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing* 184, 232–242. doi: 10.1016/j.neucom.2015.08.104

Xiao, Y., Wu, J., Lin, Z., and Zhao, X. (2018). A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. *Comput. Methods Programs Biomed.* 166, 99–105. doi: 10.1016/j.cmpb.2018.10.004

Xu, H., Abuduwufuer, A., Lv, W., Zhou, Z., Yang, Y., Zhang, C., et al. (2019). The role of HIF-1α-VEGF pathway in bronchiolitis obliterans after lung transplantation. *J. Cardiothorac. Surg.* 14:27. doi: 10.1186/s13019-019-0832-z

Yang, C., Mai, H., Peng, J., Zhou, B., Hou, J., and Jiang, D. (2020). STAT4: an immunoregulator contributing to diverse human diseases. *Int. J. Biol. Sci.* 16, 1575–1585. doi: 10.7150/ijbs.41852

Yang, X., Yang, X. H., and Zhang, W. Z. (2019). Temporal expression of PD-1 and PD-L1 during the development of experimental periodontitis in rats and its implications. *Shanghai Kou Qiang Yi Xue* 28, 591–596. Chinese.

Yoshida, T., Yamashita, K., Watanabe, M., Koshizuka, Y., Kuraya, D., Ogura, M., et al. (2015). The impact of c-Fos/Activator protein-1 inhibition on allogeneic pancreatic Islet transplantation. *Am. J. Transplant.* 15, 2565–2575. doi: 10.1111/ajt.13338

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

Yun, P. L. W., Decarlo, A. A., Chapple, C. C., and Hunter, N. (2005). Functional implication of the hydrolysis of platelet endothelial cell adhesion molecule 1 (CD31) by gingipains of *Porphyromonas gingivalis* for the pathology of periodontal disease. *Infect. Immun.* 73, 1386–1398. doi: 10.1128/IAI.73.3.1386-1398.2005

Zhao, Z., Li, Y., Wu, Y., and Chen, R. (2019). Deep learning-based model for predicting progression in patients with head and neck squamous cell carcinoma. *Cancer Biomark.* 27, 19–28. doi: 10.3233/CBM-190380

Zhou, L., Bi, C., Gao, L., An, Y., Chen, F., and Chen, F. (2019). Macrophag1996e polarization in human gingival tissue in response to periodontal disease. *Oral Dis.* 25, 265–273. doi: 10.1111/odi.12983

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 March 2021
doi: 10.3389/fgene.2021.645360

Comprehensive Analysis of APA Events and Their Association With Tumor Microenvironment in Lung Adenocarcinoma

Yuchu Zhang[1,2], Libing Shen[3], Qili Shi[4]*, Guofang Zhao[2,5]* and Fajiu Wang[2,5]*

[1] Department of Intensive Care Medicine, HwaMei Hospital, University of Chinese Academy of Sciences, Ningbo, China, [2] Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo, China, [3] Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China, [4] Fudan University Shanghai Cancer Center and Institutes of Biomedical Sciences, Shanghai Medical College, Fudan University, Shanghai, China, [5] Department of Cardiothoracic Surgery, HwaMei Hospital, University of Chinese Academy of Sciences, Ningbo, China

Background: Alternative polyadenylation (APA) is a pervasive posttranscriptional mechanism regulating gene expression. However, the specific dysregulation of APA events and its potential biological or clinical significance in lung adenocarcinoma (LUAD) remain unclear.

Methods: Here, we collected RNA-Seq data from two independent datasets: GSE40419 (n = 146) and The Cancer Genome Atlas (TCGA) LUAD (n = 542). The DaPars algorithm was employed to characterize the APA profiles in tumor and normal samples. Spearman correlation was used to assess the effects of APA regulators on 3′ UTR changes in tumors. The Cox proportional hazard model was used to identify clinically relevant APA events and regulators. We stratified 512 patients with LUAD in the TCGA cohort through consensus clustering based on the expression of APA factors.

Findings: We identified remarkably consistent alternative 3′ UTR isoforms between the two cohorts, most of which were shortened in LUAD. Our analyses further suggested that aberrant usage of proximal polyA sites resulted in escape from miRNA binding, thus increasing gene expression. Notably, we found that the 3′ UTR lengths of the mRNA transcriptome were correlated with the expression levels of APA factors. We further identified that CPSF2 and CPEB3 may serve as key regulators in both datasets. Finally, four LUAD subtypes according to different APA factor expression patterns displayed distinct clinical results and oncogenic features related to tumor microenvironment including immune, metabolic, and hypoxic status.

Interpretation: Our analyses characterize the APA profiles among patients with LUAD and identify two key regulators for APA events in LUAD, CPSF2 and CPEB3, which could serve as the potential prognostic genes in LUAD.

Keywords: alternative polyadenylation, lung adenocarcinoma, immunity, metabolism, miRNA

OPEN ACCESS

Edited by:
Shengli Li,
Shanghai Jiao Tong University, China

Reviewed by:
Zhao Zhang,
University of Texas Health Science Center at Houston, United States
Hang Ruan,
Soochow University, China

****Correspondence:***
Qili Shi
qili.shi@hotmail.com
Guofang Zhao
guofzhao@hotmail.com
Fajiu Wang
wfjwyt@163.com

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

Received: *23 December 2020*
Accepted: *17 February 2021*
Published: *18 March 2021*

Citation:
Zhang Y, Shen L, Shi Q, Zhao G and Wang F (2021) Comprehensive Analysis of APA Events and Their Association With Tumor Microenvironment in Lung Adenocarcinoma. Front. Genet. 12:645360. doi: 10.3389/fgene.2021.645360

INTRODUCTION

Non-small cell lung cancer (NSCLC) is the leading cause of cancer-related mortality worldwide (Herbst et al., 2018). Lung adenocarcinoma (LUAD) is the most prevalent histologic subtype of NSCLC and accounts for approximately 40% of all lung cancer cases (Zappa and Mousa, 2016). The 5-year survival rate for LUAD still remains poor, owing to the dismal prognosis and limited effective treatments. Therefore, elucidating the potential molecular mechanisms underlying LUAD is necessary. Advances in the characterization of alterations in the LUAD transcriptome facilitate interpretations of the complexity of the RNA processing-associated events, such as alternative splicing and polyadenylation, thus providing new perspectives on the oncogenic processes and signaling pathways in cancer development and progression (Esfahani et al., 2019).

Alternative polyadenylation (APA) has been recognized as an important factor regulating gene expression. Approximately, 70% of known human genes contain multiple polyA sites, which produce different lengths of 3′ untranslated regions (3′ UTR), thereby contributing to transcriptome diversity (Derti et al., 2012). 3′ UTR accommodates *cis* elements such as AU-rich elements (Halees et al., 2008) and microRNA (miRNA)-binding sites (Lin et al., 2012), which are involved in various aspects of posttranscriptional RNA processing. Thus, alternative usage of polyA sites can affect mRNA stability, translation, and cellular localization (Tian and Manley, 2017). The polyadenylation of mRNAs is driven by approximately 20 core proteins comprising four complexes: cleavage and polyadenylation specificity factor, cleavage stimulation factor (CstF), cleavage factors I and II, and several single proteins (Gruber and Zavolan, 2019).

Widespread shortening of 3′ UTRs has been identified in multiple types of cancer (Xia et al., 2014; Xiang et al., 2018) and cancer cells (Mayr and Bartel, 2009); this shortening activates oncogenes (Masamha et al., 2014) or represses tumor-suppressor genes (TSGs) in *trans* via disruption of the ceRNA (competing endogenous RNA) network (Park et al., 2018), thus promoting tumorigenesis. Perturbations in the expression level of APA factors have been frequently observed in a variety of cancer types, resulting in aberrant usage of proximal polyA sites (PAS) (Tan et al., 2018; Chu et al., 2019; Fischl et al., 2019; Xiong et al., 2019). Several computational tools utilizing standard RNA-sequencing (RNA-Seq) data for global APA profiling have been developed (Xia et al., 2014; Arefeen et al., 2018; Ye et al., 2018) that facilitate the identification of recurrent and tumor-specific APA events across human cancers (Xia et al., 2014; Xiang et al., 2018; Venkat et al., 2020). Nevertheless, in-depth analysis of specific APA changes in LUAD and their biological or clinical significance in a sufficiently large cohort remain to be determined. To this end, we gathered a large collection of RNA-Seq data from two LUAD cohorts, GSE40419 and TCGA-LUAD, and analyze their differences and similarities in PAS usage. We performed a systematical analysis to reveal the potential regulation and effects of APA in LUAD.

MATERIALS AND METHODS

Data Collection

RNA-Seq data and the corresponding clinical information from two independent LUAD cohorts including tumor and normal samples were downloaded from the TCGA data portal[1] and NCBI Gene Expression Omnibus (GEO) under accession number GSE40419. The numbers of paired samples in those two sets were 57 and 73 for differential analysis. The TCGA dataset contained 484 tumor samples for subsequent analyses. RNAs used in the TCGA dataset were polyA enriched and those in the GSE40419 dataset were unspecified.

Characterization of APA Events

The DaPars algorithm[2] was employed to quantify the relative polyA site usage in 3′ UTR resulting from APA through the Percentage of Distal polyA site Usage Index (PDUI), which indicates lengthening (positive index) or shortening (negative index) of 3′ UTRs (Xia et al., 2014). To identify the differences in 3′ UTRs between tumor and normal samples, we utilized the paired Wilcoxon rank-sum test to determine the significance. The differential APA events were defined by the Benjamini–Hochberg adjusted *p*-value (i.e., false discovery rate) <0.05 and $|\Delta \text{PDUI}| = |\text{PDUI}_{\text{tumor}} - \text{PDUI}_{\text{normal}}| > 0.1$.

Analysis of miRNA-Binding Sites and DEGs

miRNA-predicted targets and binding sites were downloaded from TargetScanHuman 7.2. High-confidence sites were filtered by context + score percentile > 90 (Agarwal et al., 2015). We then applied this genomic feature on the 3′ UTR changes identified by the DaPars algorithm to acquire the number of genes that lost miRNA targets. The R package "EdgeR" (version 3.30.3) was employed to identify differentially expressed genes (DEGs) with a Benjamini and Hochberg adjusted *p*-value < 0.05 (Robinson et al., 2010).

Analysis of APA Core Regulators

Genes in the GO terms associated with mRNA polyadenylation (mRNA polyadenylation, mitochondrial mRNA polyadenylation, regulation of mRNA polyadenylation, negative regulation of mRNA polyadenylation, and positive regulation of mRNA polyadenylation) were considered as APA core regulators. All the somatic mutations of the TCGA-LUAD cohort were obtained from the publicly available TCGA MAF file which includes 562 patients [3]. This dataset along with the copy number variation data were directly downloaded from cBioPortal[3] (Gao et al., 2013). APA regulator genes were expected to control the 3′ UTR lengths of targets. The expression levels of those regulators can be influenced by the copy number variation. Therefore, the transcripts per kilobase million (TPM) values for APA regulators were used to calculate the Spearman correlations between each

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

[2] https://github.com/ZhengXia/DaPars

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

PDUI and the copy number change in those regulators in tumors. A Spearman correlation coefficient | rho| > 0.3 and adjusted p-value < 0.05 were considered significant.

Survival Analysis for APA Events and Their Regulators

The univariate Cox proportional-hazard model implemented in the "coxph" function from the R package "survival" (version 3.1-12) was used for each differential APA event and regulator gene. The expression levels of APA regulators were log2(TPM + 0.01) transformed before analysis. A likelihood ratio test with $p < 0.05$ was considered significantly associated with survival time. Hazard ratios >1 indicated survival risks, whereas those <1 were associated with better outcomes.

Clustering Samples Based on Transcriptional Profiles of APA Regulators

Z-score transformation was performed to normalize the expression of 35 APA regulators. Consensus *K*-means clustering of 512 LUAD samples on the basis of the Euclidean distances of the APA regulators was conducted from $k = 2$ to $k = 9$. For each iteration, 80% of the tumor samples and 100% of the regulators were selected. This process was repeated for 1,000 times. Empirical cumulative distribution CDF plots were generated for each k to identify the k at which the CDF area reached an approximate maximum value. This clustering analysis was performed in the R package "ConsensusClusterPlus" (version 1.52.0) (Wilkerson and Hayes, 2010). Kaplan–Meier survival curve analysis and log-rank tests were used to compare the survival distributions among the four groups identified by consensus clustering in the R package "survival" (version 3.1-12).

Calculation of Immune, Hypoxic, and Metabolic Enrichment Scores

Gene markers of 22 immune cells were downloaded from CIBERSORT[4] (Newman et al., 2015). A 15-gene expression signature was selected for the hypoxia markers because they have been shown to classify hypoxia status at best (Ye et al., 2019). Single sample gene set enrichment analysis (ssGSEA) implemented in the R package "GSVA" (version 1.24.0) (Hanzelmann et al., 2013) was conducted to calculate the normalized enrichment score (NES) for each gene set of the 22 immune cells and the hypoxia status. Genes of 5 metabolic pathways were downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Gene set variation analysis (GSVA) was used to calculate the enrichment score of each metabolic pathway.

RESULTS

Global 3′ UTR Shortening in LUAD

To explore the APA changes between tumor and adjacent normal samples, we analyzed 57 and 73 paired patients with LUAD from the TCGA and Korean cohorts, respectively. Among the events detected in the tumor group or the normal group, less than half of samples (occurrence rate < 50%) were discarded. A total of 4,303 and 7,267 events remained for differential analysis in those two sets. Among those events, 272 and 1,098 from 263 to 1,074 genes significantly differed (adjusted p-value < 0.05 and $|\text{PDUItumor} - \text{PDUInormal}| > 0.1$) in the TCGA and Korean datasets, respectively (**Figures 1A,B** and **Supplementary Figures 1A,B**). Notably, the numbers of shortened 3′ UTR events in tumors far exceeded the numbers of lengthened 3′ UTR events (**Figures 1A,B** and **Supplementary Figures 1A,B**), in agreement with the patterns observed in previous pan-cancer analyses (Xia et al., 2014; Xiang et al., 2018). The significantly changed transcripts showed longer 3′ UTR lengths than were observed below the threshold (**Figure 1C**). Furthermore, we compared the 3′ UTR lengths among oncogenes (Liu et al., 2017), TSGs (Zhao et al., 2016), and other genes (**Figure 1D**). The results indicated that oncogenes tended to have longer 3′ UTR length than TSGs and other genes. Next, to determine the recurrent APA alterations in LUAD, we combined the results from the two studies. As shown in **Figure 1E**. A total 114 transcripts were determined to have changed in both cohorts, thus representing a strongly significant overlap (p-value = 1.66e−52, hypergeometric test). APA-derived 3′ UTRs have been proposed to affect the mRNA and protein location (Berkovits and Mayr, 2015). Therefore, we conducted an overrepresentation analysis of cellular component for those recurrent APA alterations found in LUAD by using the R package "clusterProfiler" (version 3.11) (Yu et al., 2012). Strikingly, the recurrent changed genes were highly enriched in the membrane (**Figure 1F**), thus suggesting that APA may be involved in regulating the localization of membrane proteins (Berkovits and Mayr, 2015) or the subcellular localization of mRNA transcripts for cancer-specific genes. For example, we showed the detailed recurrent alterations in 3′ UTRs located in lysosomal membranes (**Supplementary Figures 1C–F**).

3′ UTR Shortening-Mediated Loss of miRNA-Binding Sites

Independently of mRNA and protein localization, 3′ UTR shortening through APA during tumorigenesis may escape miRNA repression, thus increasing gene expression. Therefore, we calculated the distribution of lost miRNA-binding sites as a result of shortened 3′ UTR lengths in LUAD (**Figure 2A**). As revealed by this analysis, 77.4 and 65.8% of events with shortened 3′ UTRs from the TCGA and Korean cohorts had lost at least one predicted miRNA-binding site (**Figure 2A**). Furthermore, we compared the miRNA-binding sites for 3′ UTR shortened transcripts with those below the threshold. Consistent with a previous pan-cancer analysis (Xia et al., 2014), the results (**Figure 2B**) showed that those shortened events in tumors had overall greater miRNA-binding site density (p-value = 1.34e−10 and 0, Kolmogorov–Smirnov test), suggesting that cancer cells may maximize the mitigation of miRNA binding by preferentially shortening the 3′ UTR, in a process strictly regulated by miRNAs. To examine the effects of miRNA-binding loss mediated by 3′ UTR shortening, we analyzed DEGs between paired normal

[4] https://cibersort.stanford.edu/

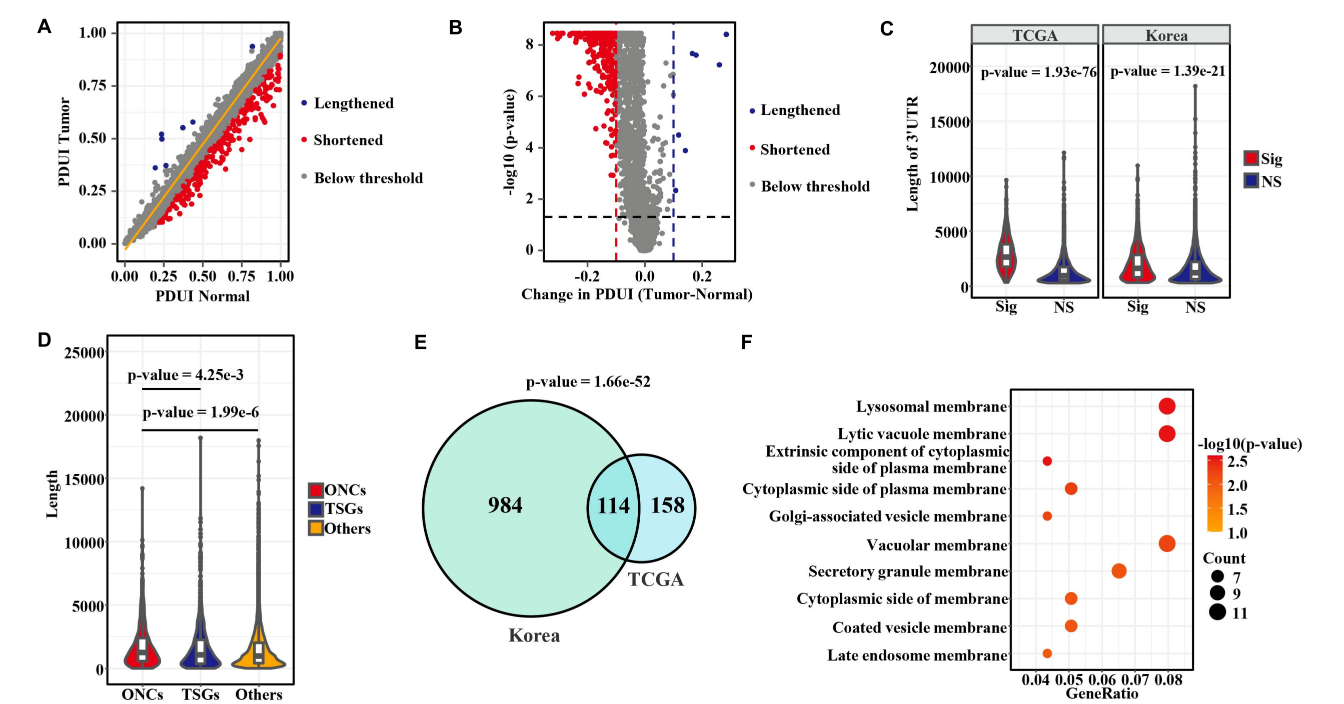


FIGURE 1 | Comprehensive characterization of aberrant APA in LUAD. **(A)** Scatterplot of PDUIs in normal (*x* axis) and tumor (*y* axis) samples from the TCGA cohort. Significantly (adjusted *p*-value < 0.05 and | ΔPDUI| > 0.1) shortened and lengthened transcripts are indicated in red and blue, respectively, whereas those below the threshold are gray. **(B)** Volcano plot showing the significantly altered APA events in the TCGA cohort. **(C)** Comparison of 3′ UTR lengths between significantly changed transcripts in tumors and other transcripts detected in both tumor and normal samples that did not pass the threshold. Here, *p*-values were calculated by the Wilcoxon rank-sum test. **(D)** Comparison of 3′ UTR lengths among oncogenes, tumor-suppressor genes, and other genes annotated in databases. Statistical differences were determined by the Wilcoxon rank-sum test. **(E)** Venn diagram showing the strong overlap of altered APA events between the two datasets. **(F)** Dot plot indicating significantly enriched cellular components of genes with recurrent APA alterations found in both two cohorts.

and tumor tissues. Among genes with shortened 3′ UTR, 103 and 417 were significantly upregulated in the tumors in the two cohorts, possibly as a consequence of escape from miRNA repression (**Figure 2C**). Nevertheless, when compared with all DEGs, the genes with shortened 3′ UTRs did not tend to be more upregulated in LUAD (*p*-value = 0.07 and 1, hypergeometric test). This result is consistent with prior analyses in pancreatic ductal adenocarcinoma (Venkat et al., 2020) and other types of cancer (Xiang et al., 2018), suggesting the presence of other mechanisms in regulating gene expression. In addition, we found three genes, COL5A1, COL1A2, and CP with lengthened 3′ UTR were upregulated in tumors in both the datasets. Several genes have been reported that their longer 3′ UTR isoform can enhance expression through *trans*-regulation mechanism (Allen et al., 2013; Arake et al., 2019).

Regulators of APA Events in LUAD

To investigate potential regulators governing APA alternations in LUAD development, we analyzed the differential expression of 35 genes collected from the GO terms associated with "mRNA polyadenylation" between normal and tumor sample pairs. Most of those genes (TCGA, 26/35, and Korean, 25/32) that were differentially expressed (adjusted *p*-value < 0.05) in the two cohorts were upregulated in tumors (**Figure 3A** and **Supplementary Figure 2A**). Moreover, we found that 19 and 2 APA regulators were both upregulated and downregulated in two datasets (**Figure 3A** and **Supplementary Figure 2A**). For example, CSTF2 has been reported to promote 3′ UTR shortening of cancer-related genes in NSCLC and was upregulated in both the TCGA and Korean cohorts. To further explore genetic alterations in APA regulators in LUAD that may affect their expression levels, we analyzed somatic mutations and copy number variations (CNVs) of these genes in the TCGA cohort. We found that 28.1% (158/562) of the LUAD tumor samples had at least one protein-affecting mutation (**Figure 3B**). Most components of the 3′ end-processing machinery are RNA-binding proteins; the mutation frequency of these factors ranged from 0.2 to 2.8%, a percentage not greater than that for other RNA-binding proteins observed in pan-cancer studies (Sebestyen et al., 2016; Li et al., 2019). Compared with somatic mutations, CNVs were highly recurrent across patients with a range of 32.1–72.8% (**Figure 3C**). We found that 69.7% (23/33) of APA factors were positively correlated (rho > 0.3 and adjusted *p*-value < 0.05) with their mRNA expressions in tumors (**Supplementary Figure 2B**). A total of 14 APA regulators with more than half of CNV gains showed significantly higher expression in tumors (e.g., CDC73 and ZC3H3), whereas the two downregulated factors, CPEB1 and CPEB3, both had more than half of CNV losses (**Figure 3C**). Our data also indicated that widespread 3′ UTR shortening in LUAD might be caused

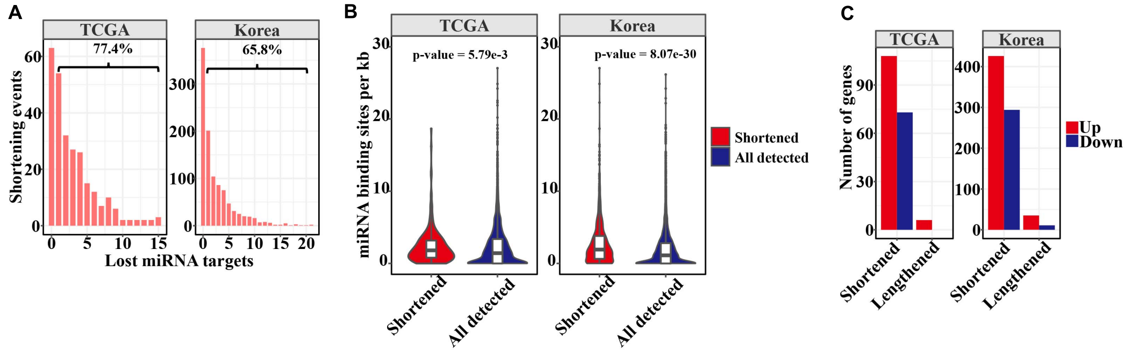


FIGURE 2 | APA mediated loss of miRNA binding sites. **(A)** Barplots showing the distribution of lost miRNA-binding sites resulting from 3′ UTR shortening. The percentage of shortening events losing at least one miRNA-binding site is displayed above the bracket. **(B)** Comparison of miRNA-binding sites between significantly shortened transcripts in tumors and others below the threshold. **(C)** Barplots showing the number of upregulated or downregulated that may be affected by APA in tumors.

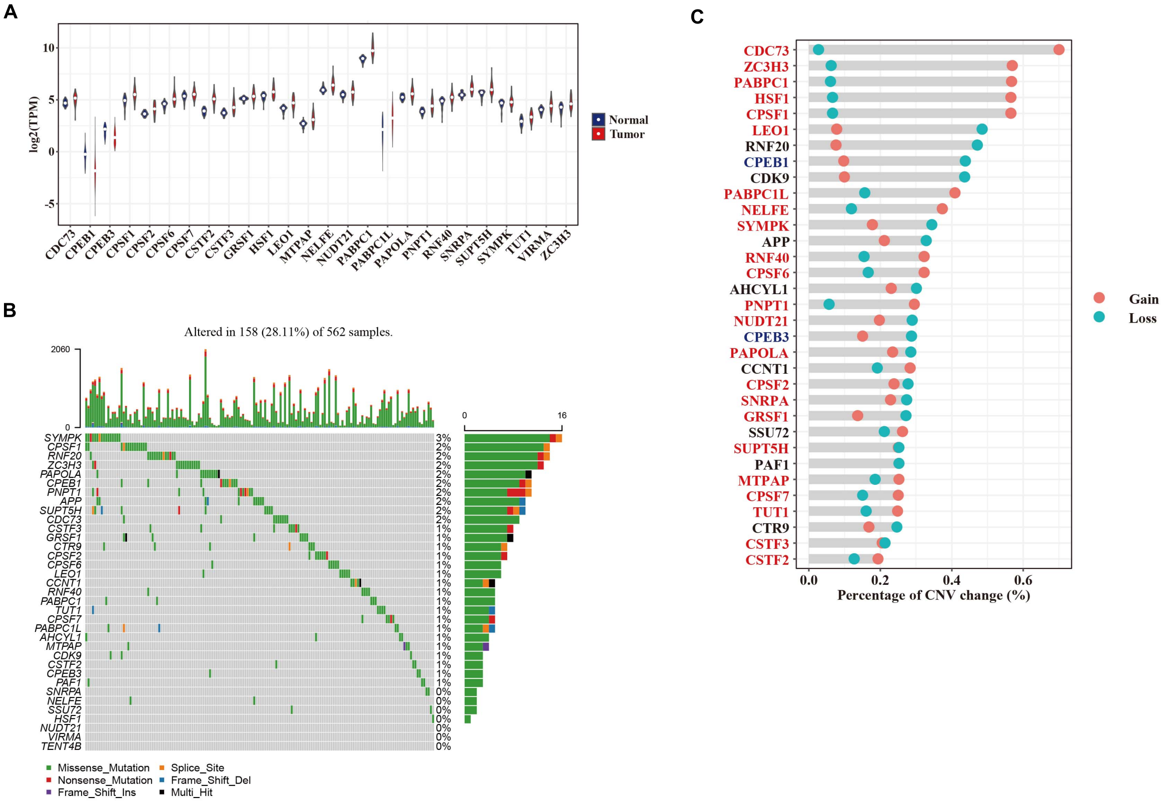


FIGURE 3 | Genetic and expression alterations in APA regulators. **(A)** Violin plot showing the expression of 26 significantly dysregulated APA factors between tumor (red) and adjacent normal (blue) samples in the TCGA cohort. **(B)** The mutation landscape of 35 APA regulators in the TCGA cohort. Top panel shows the tumor mutation rate of each patient. Bottom panel indicates the mutation frequency of individual regulators. Mutation types are shown in the legend at the bottom. **(C)** The CNV variation frequency of APA regulators in the TCGA cohort. Gain and loss of CNV are indicated by red and blue dots, respectively. Upregulated and downregulated APA factors are colored in red and blue.

by the elevated expression of polyadenylation factors through enhanced usage of PAS, consistent with findings from a study in proliferating cells (Elkon et al., 2012).

We further investigated the correlation between APA events and the expression levels of their regulators in tumors. Remarkably, among these APA events 44.2% (3,165/7,163) and 79.4% (5,618/7,072) of them were significantly corrected (|rho| > 0.3 and adjusted p-value < 0.05) with at least one factor in the TCGA and Korea datasets respectively (**Figures 4A,B**). Moreover, we observed strongly negative associations between

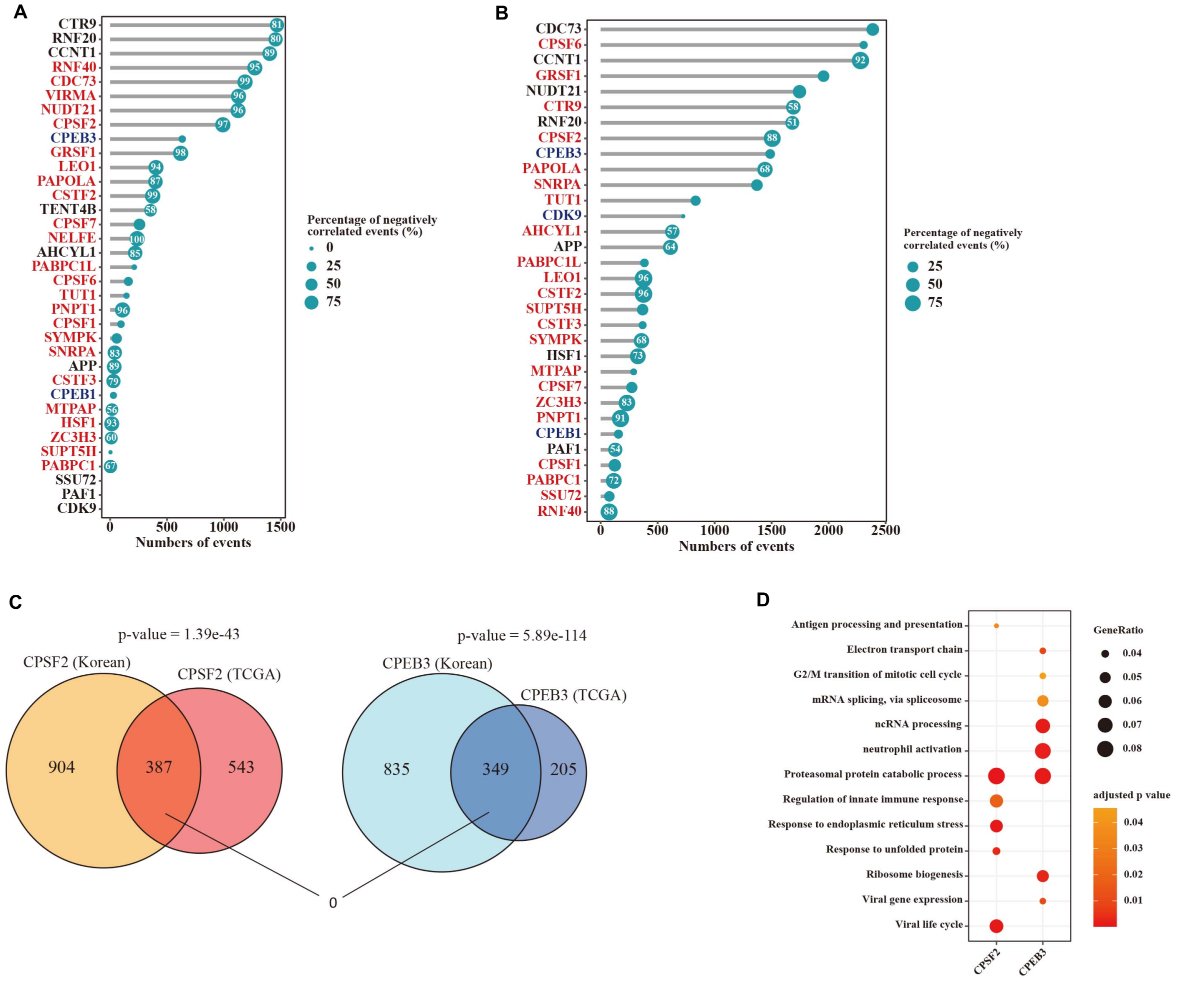


FIGURE 4 | Potential mechanisms for APA regulation in LUAD. **(A)** Lollipop chart indicating the number of significantly correlated APA events with each regulator gene in the TCGA dataset. Dot size is proportional to the percentage of negatively correlated events. Numbers in the dots represent the percentages of negatively correlated events that are greater than 50%. **(B)** Lollipop chart indicating the number of significantly correlated APA events with each regulator gene in the Korea dataset. **(C)** Venn diagram showing the overlap of altered APA events correlated with CPSF2 or CPEB3 between the two datasets. **(D)** Dot plot indicating significantly enriched biological processes of events correlated with CPSF2 or CPEB3 in both two cohorts.

3′ UTR lengths and mRNA expression of those factors in the TCGA dataset (**Figure 4A**). To define certain factors dysregulated in tumors that could primarily be responsible for APA mechanism in LUAD, we filtered APA factors through those upregulated with more than half of negatively correlated events or downregulated with more than half of positively correlated events. Subsequently, CPSF2 and CPEB3 were identified that correlated with more than 500 APA events in both datasets, which could be master regulators of APA in LUAD. Moreover, 387 and 349 genes were determined to be correlated with CPSF2 and CPEB3 in both datasets respectively (**Figure 4C**), showing strongly significant overlaps (p-value = 1.39e−43 and p-value = 5.89e−114, hypergeometric test). Interestingly, no genes were shared by CPSF2- and CPEB3-correlated APA events (**Figure 4C**), suggesting that the two factors may regulate APA alternations in LUAD independently. To test it, we performed an overrepresentation analysis of biological processes for 387 and 349 genes correlated with the two factors. As shown in **Figure 4D**, they can both regulate the proteasomal protein catabolic process through the APA mechanism and most other processes enriched in the two factors were different.

The Prognostic Value of APA Events and Their Regulators in LUAD

Understanding the widespread alterations in APA events and their regulators in LUAD could provide important insights for translational medicine. We performed univariate Cox regression analyses between survival time and 272 transcripts with significant 3′ UTR changes in the TCGA dataset. In total, 51 events significantly associated with survival time were identified. Notably, patients with shortened 3′ UTRs for all those events had poorer clinical outcomes, thus suggesting that use of a PAS may exacerbate LUAD malignancy. The top ten significant events are shown in **Figure 5A**, whose hazard ratios ranged from 0.026 to

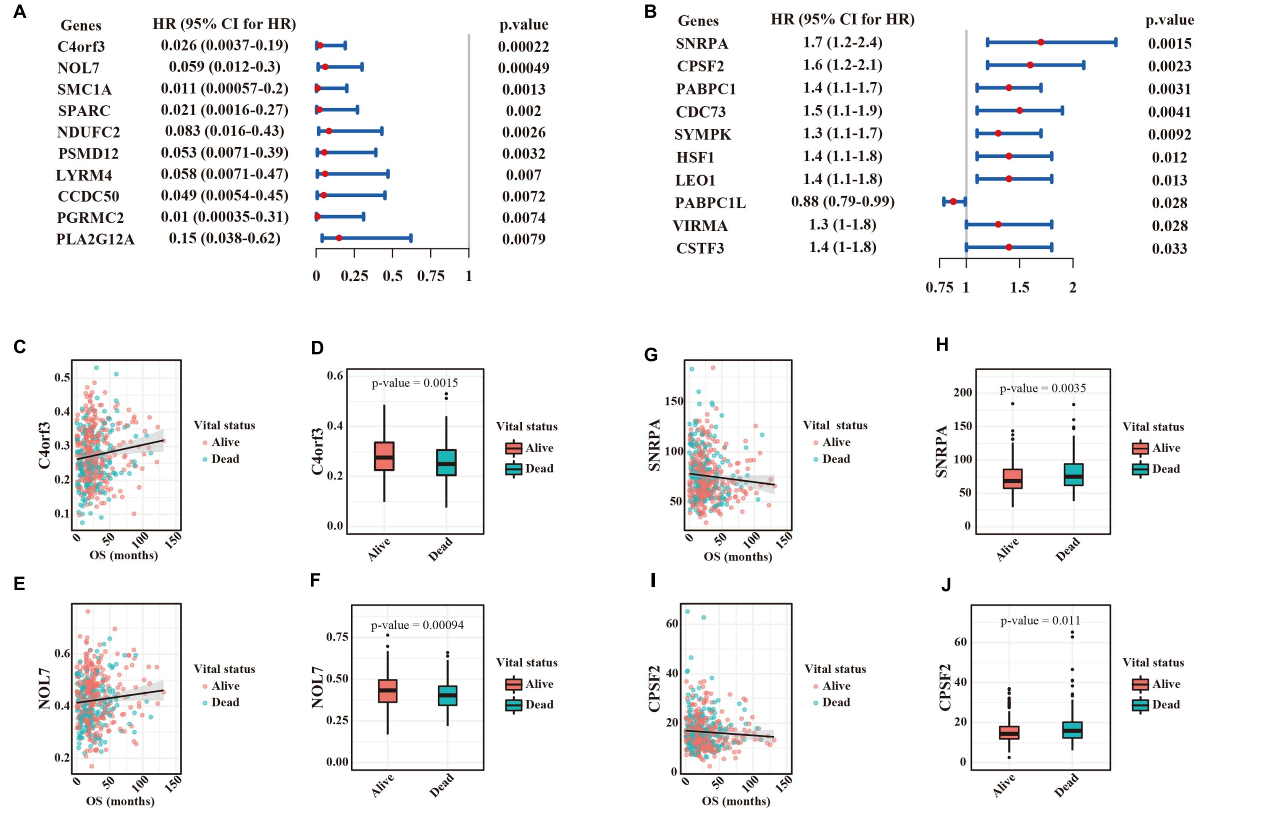


FIGURE 5 | Survival-associated APA events and their regulators in LUAD. **(A)** Ranked list of the top ten survival-associated APA events according to the *p*-values calculated by the likelihood ratio test. Forest plots showing the hazard ratio and its upper and lower boundary of 95% confidence interval. Hazard ratios > 1 indicated survival risks, whereas those <1 were associated with better outcomes. **(B)** Ranked list of the top ten survival-associated APA regulators according to the *p*-values calculated by the likelihood ratio test. **(C)** Scatter plot of C4orf3 PDUI scores (*y* axis) and survival time (*x* axis). Each dot represents a tumor sample. **(D)** Comparison of C4orf3 PDUI scores between alive and dead patients. **(E)** Scatter plot of NOL7 expression (*y* axis) and survival time (*x* axis). **(F)** Comparison of NOL7 PDUI scores between alive and dead patients. **(G)** Scatter plot of SNRPA TPM values (*y* axis) and survival time (*x* axis). **(H)** Comparison of SNRPA expression between alive and dead patients. **(I)** Scatter plot of CPSF2 TPM values (*y* axis) and survival time (*x* axis). **(J)** Comparison of CPSF2 expression between alive and dead patients.

0.15. Scatter plot and box plot (**Figures 5C–F**) further showed the positive association between PDUI scores and survival results (e.g., C4orf3 and NOL7). Furthermore, we focused on the associations between the expression of APA factors and survival results. Ten factors were identified to be significantly correlated with survival time (**Figure 5B**). Importantly, among them, a high expression of nine genes that were upregulated in tumors was associated with poor prognosis of patients with LUAD. Scatter plot and box plot (**Figures 5G–J**) further showed the negative association between the expression levels of most APA factors and survival results (e.g., SNRPA and CPSF2).

Examination of APA Factors Mediating Heterogeneity of Proximal PAS Usage Identifies LUAD Subtypes With Distinct Clinical and Molecular Features

Next, we explored whether the expression of APA factors might contribute to the stratification of LUAD. According to the expression pattern of APA regulators, we identified four subtypes of 512 patients in the TCGA cohort through consensus clustering (**Figure 6A**). The optimal number of subtypes was determined by an empirical CDF plot (**Supplementary Figures 3A,B**). The four subtypes displayed significant differences in overall survival (**Figure 6B**). Among them, subtype 4 consisted of 50 patients with the highest expression of APA factors, who had the worst survival results (**Figures 6B,C**). To investigate APA factors mediating heterogeneity of proximal PAS usage in tumors, we further compared the 3′ UTR differences among the four subtypes. In agreement with the expression levels of APA regulators, subtype 1 showed the greatest usage of distal 3′ UTRs whereas subtype 4 displayed the greatest proximal APAs (**Supplementary Figure 3C**). Moreover, we compared the 3′ UTR lengths and miRNA-binding sites between the events significantly shortened in subtype 4 (adjusted *p*-value < 0.05) and those below the threshold. As with the differentially regulated APA events, 3′ UTR-shortened transcripts in subtype 4 showed longer 3′ UTR lengths and greater miRNA-binding sites (**Supplementary Figures 3D,E**), suggesting that 3′ UTR shortening-mediated loss of miRNA-binding sites was associated with LUAD aggressiveness. To explore whether any distinct APA patterns can be seen among the four subtypes of LUAD,

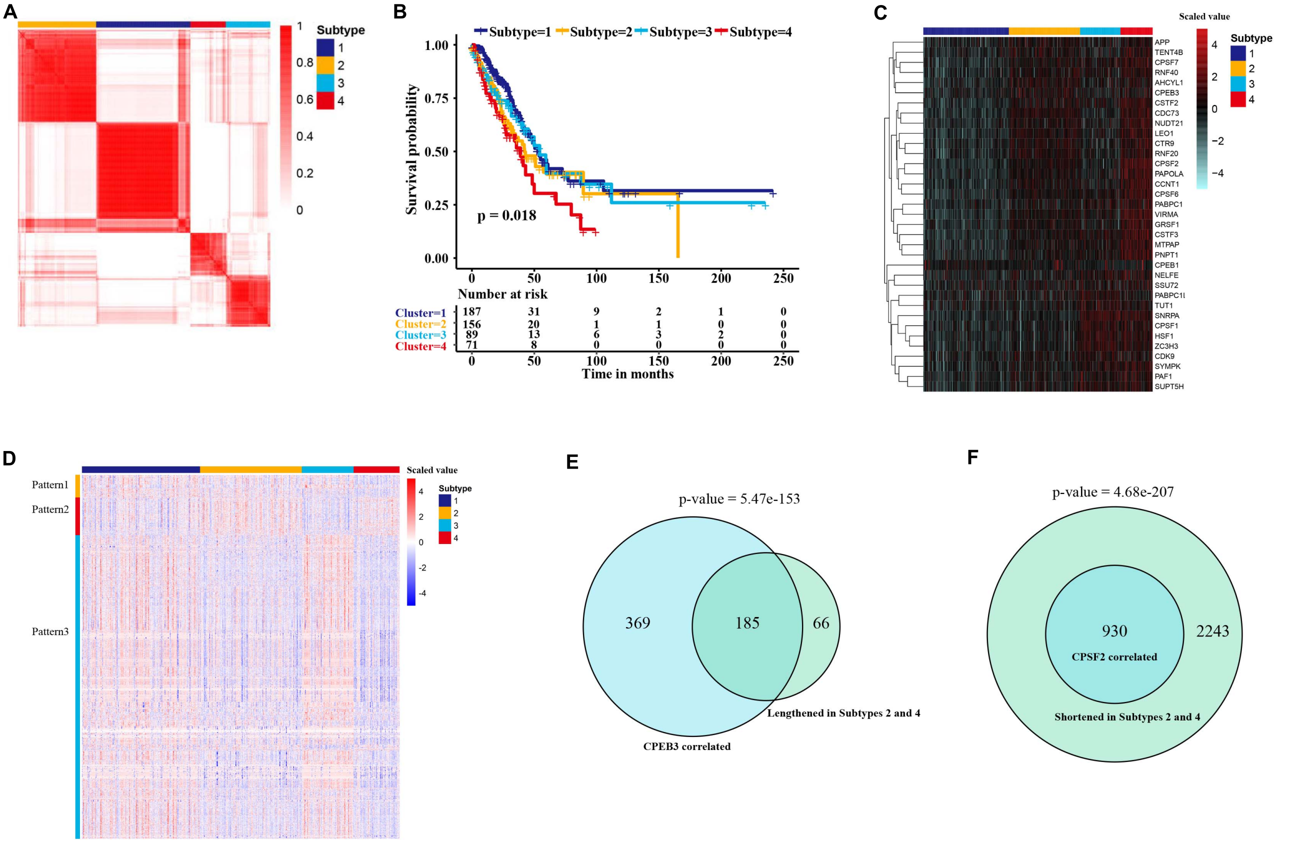


FIGURE 6 | Expression heterogeneity of APA factors reveals LUAD subtypes with distinct APA patterns and clinical features. **(A)** Consensus clustering of patients (n = 512) based on expression of APA factors identifies four subtypes in LUAD. The color from white to red represents the consistency ranging from 0 to 1. **(B)** Kaplan–Meier survival plot of patients grouped by global expression patterns of APA regulators. The survival difference was determined by the log-rank test. **(C)** Heat map of 35 APA factors showing the difference among the four subtypes. The color indicates the scaled expression value (red, high; blue, low). **(D)** Heat map of 3,730 APA events displaying the differences among the four subtypes. The color indicates the scaled PDUI value (red, high; blue, low). **(E)** Venn diagram showing the overlap between altered APA events correlated with CPSF2 and events lengthened in subgroups 2 and 4. **(F)** Venn diagram showing the overlap between altered APA events correlated with CPEB3 and events shortened in subgroups 2 and 4.

we further investigated 3,731 significantly different APA events (Kruskal–Wallis test, adjusted p-value < 0.001). As shown in **Figure 6D**, three distinct patterns that may be regulated by specific factors were observed in four groups. The APA events with pattern 1 showed shorter 3′ UTR subtypes 3 and 4, which may be regulated by the factors upregulated in these two subtypes. Intriguingly, we found that pattern 2 showed longer 3′ UTR in subtypes 2 and 4. We hypothesized that pattern 2 could be caused by CPEB3 that was upregulated in those two subtypes. To test it, we compared this pattern with CPEB3 positively correlated events. As shown in **Figure 6E**, among 251 lengthened genes, 73.7% (185) may depend on the expression of CPEB3. Pattern 3 consisted of the largest numbers of APA events that were shortened in subtypes 2 and 4, which may be caused by the factors upregulated in these two subtypes. In addition, APA events negatively correlated with CPSF2 which we identified as a possible master regulator were all in pattern 3 (**Figure 6F**).

To investigate the impact of APA heterogeneity on gene expression, we focused on subtype 4 exhibiting the greatest APA changes. We found that among 3,669 shortened genes, 2,163 (p-value = 2.24e−6, hypergeometric test) showed a significantly higher expression level in subtype 4 (**Figure 7A**). We further explored the functional implications of the gene expression heterogeneity among the four subtypes mediated by APA events. GO and KEGG pathway enrichment analysis of 2,163 overlapping genes identified several highly enriched GO terms: histone modification, RNA splicing, proteasomal process, and cell cycle (**Figure 7B**). Similar biological processes have been observed in pan-cancer correlation analysis (Xiang et al., 2018), and our results further suggested APA regulation of those functions. Remarkably, we also found enriched pathways related to immune and hypoxia such as NIK/NF-κB, Wnt, and TNF signaling (**Figure 7B**). Thus, the infiltration levels of 22 immune cells and hypoxia status in patients were estimated by using ssGSEA based on the previous reported gene signatures (Newman et al., 2015; Ye et al., 2019). Strikingly, the four subtypes displayed marked differences in immune and hypoxia status (**Figures 7C,D**). Subtype 1, with the greatest distal PAS usage, showed the highest innate and adaptive immune cell infiltration and the lowest hypoxia score (**Figures 7C,D**). Overall,

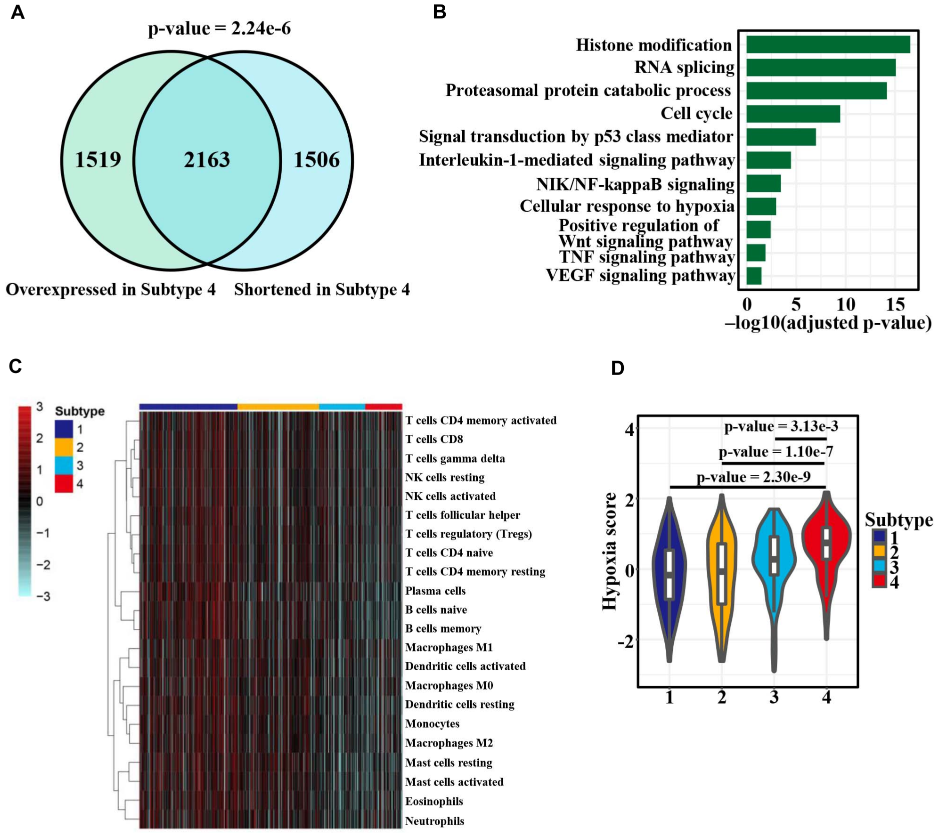


FIGURE 7 | APA heterogeneity contributing to gene expression and TME differences among patients with LUAD. **(A)** Venn diagram comparison between genes with shortened 3′ UTRs and those upregulated in subtype 4 (adjusted *p*-value < 0.05, one-sided Wilcoxon rank-sum test). The color from white to red represents the consistency ranging from 0 to 1. **(B)** Significantly enriched (adjusted *p*-value < 0.05, hypergeometric test) GO terms and KEGG pathways in the 2,163 overlapping genes. **(C)** Heat map of 22 immune cell subsets calculated with ssGSEA, indicating the TME differences among the four subtypes. The color indicates the normalized score value (red, high; blue, low). **(D)** Differences in hypoxia scores among the four groups. Statistical differences were determined by the Wilcoxon rank-sum test.

these results indicated the role of APA in shaping the tumor microenvironment (TME) or vice versa.

Heterogeneity of Proximal PAS Usage of Metabolic Genes in LUAD Patients

We also found that gene expression heterogeneity among LUAD patients mediated by APA events was enriched in five metabolic pathways including citrate cycle (TCA cycle), lysine degradation, cysteine and methionine metabolism, glycolysis (gluconeogenesis), and fructose and mannose metabolism (**Figure 8A**). Therefore, we compared the NESs of five pathways among four subtypes. As expected, subtype 4, with the greatest usage of proximal PAS, showed the highest score in all five metabolic pathways (**Figures 8B–F**). To examine whether this heterogeneity is regulated by APA mechanism, we conducted the correlations between expression levels of genes in the glycolytic pathway and their 3′ UTR lengths. DLAT, PFKM, and PGAM1 were reported can promote cancer cell growth through the glycolytic pathway (Tang et al., 2012; Goh et al., 2015; Huang et al., 2019). As shown in **Figures 8G–I**, DLAT, PFKM, and PGAM1 were all negatively correlated with their PDUI values. Moreover, we found that CPSF2 may regulate APA events of metabolic genes like DLAT (**Figure 8J**).

DISCUSSION

Based on the large-scale RNA-seq data from two cohorts, we provided a systemic and specific portrait of the APA landscape in LUAD. In agreement with previous studies (Xia et al., 2014; Xiang et al., 2018), our analyses revealed global shortening of APA in tumor samples when compared with paired controls. Notably, we found high consistency in APA alterations between the two datasets, a result previously unnoticed in pan-cancer or single-tumor-type analyses. Moreover, genes with significantly changed 3′ UTRs were enriched in locations of cell membrane and some organelle membranes, including those of lysosomes, vacuoles, and late endosomes. A novel mechanism in which alternative 3′ UTR isoforms of membrane genes can determine their subcellular protein localization and function has been identified in a previous study (Berkovits and Mayr, 2015). CD47, a well-established cell surface molecule, can produce alternative 3′ UTR isoforms that localize to

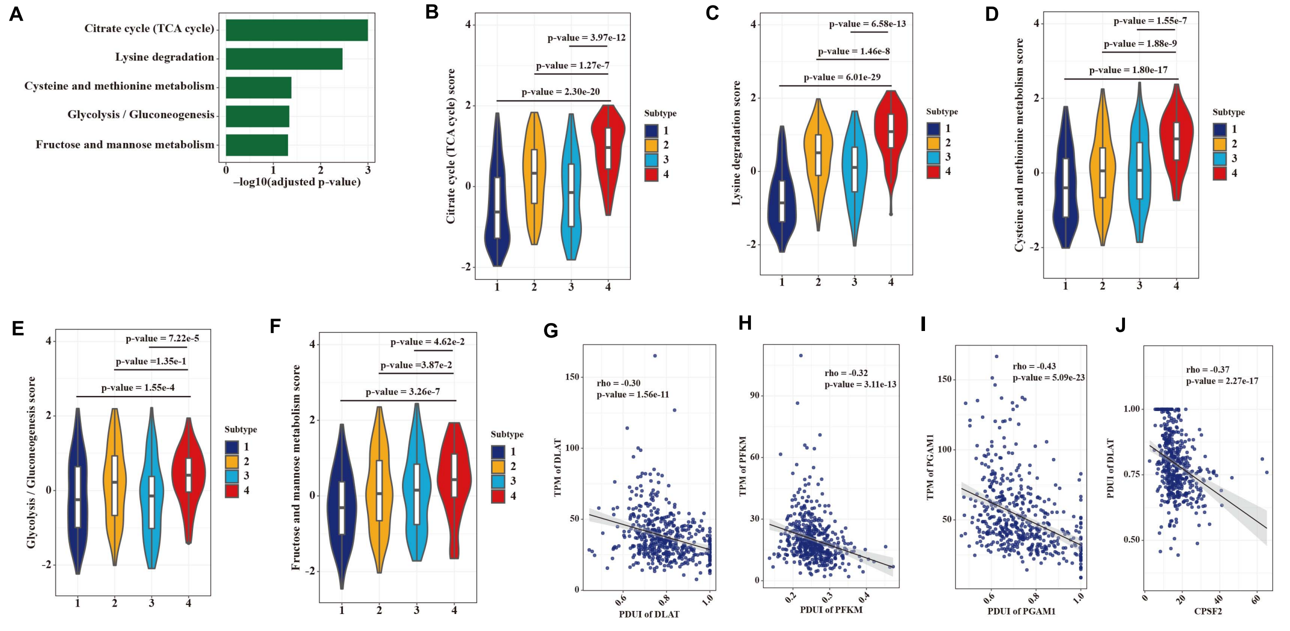


FIGURE 8 | APA heterogeneity contributing to metabolic gene expression differences in LUAD. **(A)** Significantly enriched (adjusted *p*-value < 0.05, hypergeometric test) metabolic pathways in the 2,163 overlapping genes. **(B–F)** Differences in scores of metabolic pathways among the four groups. Statistical differences were determined by the Wilcoxon rank-sum test. **(G)** Correlation between the expression level and the APA event of DLAT. **(H)** Correlation between the expression level and the APA event of PFKM. **(I)** Correlation between the expression level and the APA event of PGAM1. **(J)** Correlation between the expression level of CPSF2 and the APA event of DLAT.

different cellular compartments and show opposite functions in cell survival and cell migration (Berkovits and Mayr, 2015). Our analyses suggest that alternations of 3′ UTR lengths in membrane-associated genes may promote cancer cell growth through APA-dependent protein localization. 3′ UTR shortening-mediated miRNA binding loss has been found to affect the expression levels of these genes (Venkat et al., 2020). We observed a considerable number of genes upregulated in LUAD after shortening of their 3′ UTRs, but this result was not statistically significant when compared with the global pattern of DEGs, which indicates that APA is only one of the multiple mechanisms that govern mRNA expression levels (Venkat et al., 2020).

The regulation of alternative 3′ UTR usage in LUAD remains unclear. Our analyses indicate that most APA factors are overexpressed and negatively correlated with distal PAS usage in LUAD. CSTF2 has been recognized as the key factor that induces 3′ UTR shortening in pan-cancer analysis (Xia et al., 2014) and has been implicated in contributing to carcinogenesis of the bladder (Chen et al., 2018), breast (Akman et al., 2015), and lung (Aragaki et al., 2011). In contrast, we found that several APA factors may act as master regulators in LUAD, such as RNF40, CDC73, and VIRMA, which are not core proteins in the polyadenylation machinery. The methyltransferase component VIRMA facilitates the selection of proximal PAS through preferential m6A mRNA methylation in the 3′ UTR and near the stop codon (Yue et al., 2018). Indeed, depletion of VIRMA or METTL3 elicits global lengthening of APA events in the HeLa cell line (Yue et al., 2018). Our analysis further indicated that a high expression level of VIRMA is associated with poor survival outcomes in patients with LUAD. These findings provide a possibility that VIRMA may serve as an oncogene in LUAD that negatively regulates the 3′ UTR lengths of cancer-associated genes through m6A mRNA methylation to enhance tumorigenicity. In addition to the factors that induce 3′ UTR shortening, a previous study has revealed PABPN1 as a master regulator that promotes distal PAS usage in pan-cancer analyses including LUAD (Xiang et al., 2018). We also found that PABPN1 positively correlates with 17.6% of APA events in the Korean LUAD cohort (data not shown). Our analysis identified two genes, CPEB1 and CPEB3, which were both downregulated in the two datasets. Compared with most upregulated genes, CPEB3 is more positively correlated with APA events in tumors, thus suggesting that its regulation of preferential distal poly(A) site usage may be inhibited in LUAD. We directly calculated correlations between APA events and factors to define the potential regulations of those factors in LUAD. This analysis has a limitation in that identified regulators may be dependent on other co-expressed factors. Therefore, further experimental validation is necessary to explore the molecular mechanisms of APA regulations in LUAD. Together, our results suggest that dysregulated APA factors in LUAD may be considered as potential biomarkers and therapeutic targets, which should be further confirmed through additional experiments.

Several studies have shown the prognostic power of APA events in different cancers (Xia et al., 2014; Venkat et al., 2020). Our analyses further revealed that the patients with shorter 3′ UTR lengths show poor survival in LUAD. Some

APA events from our analyses provided noteworthy biological and clinical insights. SMC1A, a core cohesin gene, has been reported to promote tumor development in some types of human cancers (Pan et al., 2016; Zhou et al., 2017; Sarogni et al., 2019). Our results showed that the 3′ UTR of SMC1A is shortened in tumors and is significantly associated with clinical prognosis, thus providing a potential mechanism through which overexpression of SMC1A in human cancers may be contributed by marked shortening of its 3′ UTR. Expression levels of SPARC in patients with NSCLC are associated with disease diagnosis and prognosis (Koukourakis et al., 2003; Huang et al., 2012; Andriani et al., 2018). Our analyses further indicate that different poly(A) site usage of SPARC may also serve as a diagnostic and prognostic factor.

By stratification of patients with LUAD, we identified that heterogeneity in PAS usage among tumors can be explained by the mRNA expression levels of APA factors. Furthermore, 3′ UTR differences among the four subtypes considerably affected the specific mRNA transcriptome. Previous studies have shown that regulators of 3′ end processing can influence the 3′ UTR of genes in the Wnt/β-catenin and NF-κB signaling pathways, thereby determining the cancer phenotype (Ogorodnikov et al., 2018; Xiong et al., 2019). Consistent with these findings, our data underscore the crucial roles of APA factors in governing the patient-specific APA alternations, a process tightly associated with the activation of oncogenic pathways. Besides demonstrating the influence on the transcriptome in patients, our analyses suggest that 3′ UTR changes strikingly affect tumor immune and hypoxia status or vice versa. Patients with longer 3′ UTRs in global APA characterization showed higher immune and lower hypoxia scores. This finding may provide insights into strategies for potential cancer therapies targeting tumor immunity or hypoxia. Proliferating cells expressing mRNAs with shortened 3′ UTR has long been recognized (Sandberg et al., 2008). Our results suggest that APA may contribute to the altered levels of metabolic genes which in turn create a TME that promote their survival and propagation.

In summary, we presented the comprehensive landscape of 3′ UTR in LUAD and highlighted 113 recurrent APA alterations and specific factors especially two key regulators, CPSF2 and CPEB3, regulating APA patterns. Consistent with previous analyses in other cancer types, 3′ UTR shortening is frequently associated with tumor occurrence in APA events, and it may contribute to elevated gene expression through loss of miRNA-binding sites. Moreover, APA events and their regulators were found to be useful for prognosis and cancer stratification in LUAD. The resources provided herein should be valuable for understanding and exploring alternative 3′ UTR isoforms in LUAD and are expected to promote precision medicine in the future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

FW, GZ, and QS proposed and designed the project. YZ, LS, and QS performed the data collection and analyses. YZ and FW wrote the manuscript. FW and GZ revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Ningbo Health Branding Subject Fund (Grant No. 2018002), China.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.645360/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.

Akman, H. B., Oyken, M., Tuncer, T., Can, T., and Erson-Bensan, A. E. (2015). 3′UTR shortening and EGF signaling: implications for breast cancer. *Hum. Mol. Genet.* 24, 6910–6920.

Allen, M., Bird, C., Feng, W., Liu, G., Li, W., and Perrone-Bizzozero, N. I. (2013). HuD promotes BDNF expression in brain neurons via selective stabilization of the BDNF long 3′UTR mRNA. *PLoS One* 8:e55718. doi: 10.1371/journal.pone.0055718

Andriani, F., Landoni, E., Mensah, M., Facchinetti, F., Miceli, R., and Tagliabue, E. (2018). Diagnostic role of circulating extracellular matrix-related proteins in non-small cell lung cancer. *BMC Cancer* 18:899.

Aragaki, M., Takahashi, K., Akiyama, H., Tsuchiya, E., Kondo, S., and Nakamura, Y. (2011). Characterization of a cleavage stimulation factor, 3′ pre-RNA, subunit 2, 64 kDa (CSTF2) as a therapeutic target for lung cancer. *Clin. Cancer Res.* 17, 5889–5900. doi: 10.1158/1078-0432.ccr-11-0240

Arake, D. T. L., Pulos-Holmes, M. C., Floor, S. N., and Cate, J. (2019). PTBP1 mRNA isoforms and regulation of their translation. *RNA* 25, 1324–1336. doi: 10.1261/rna.070193.118

Arefeen, A., Liu, J., Xiao, X., and Jiang, T. (2018). TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics* 34, 2521–2529. doi: 10.1093/bioinformatics/bty110

Berkovits, B. D., and Mayr, C. (2015). Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature* 522, 363–367. doi: 10.1038/nature14321

Chen, X., Zhang, J. X., Luo, J. H., Wu, S., Yuan, G. J., and Ma, N. F. (2018). CSTF2-induced shortening of the RAC1 3′UTR promotes the pathogenesis of Urothelial Carcinoma of the Bladder. *Cancer Res.* 78, 5848–5862.

Chu, Y., Elrod, N., Wang, C., Li, L., Chen, T., and Routh, A. (2019). Nudt21 regulates the alternative polyadenylation of Pak1 and is predictive in the prognosis of glioblastoma patients. *Oncogene* 38, 4154–4168. doi: 10.1038/s41388-019-0714-9

Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., and Chen, R. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Res.* 22, 1173–1183. doi: 10.1101/gr.132563.111

Elkon, R., Drost, J., van Haaften, G., Jenal, M., Schrier, M., and Oude, V. J. (2012). E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol.* 13:R59.

Esfahani, M. S., Lee, L. J., Jeon, Y. J., Flynn, R. A., Stehr, H., and Hui, A. B. (2019). Functional significance of U2AF1 S34F mutations in lung adenocarcinomas. *Nat. Commun.* 10:5712.

Fischl, H., Neve, J., Wang, Z., Patel, R., Louey, A., and Tian, B. (2019). hnRNPC regulates cancer-specific alternative cleavage and polyadenylation profiles. *Nucleic Acids Res.* 47, 7580–7591. doi: 10.1093/nar/gkz461

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., and Sumer, S. O. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:l1.

Goh, W. Q., Ow, G. S., Kuznetsov, V. A., Chong, S., and Lim, Y. P. (2015). DLAT subunit of the pyruvate dehydrogenase complex is upregulated in gastric cancer-implications in cancer therapy. *Am. J. Transl. Res.* 7, 1140–1151.

Gruber, A. J., and Zavolan, M. (2019). Alternative cleavage and polyadenylation in health and disease. *Nat. Rev. Genet.* 20, 599–614. doi: 10.1038/s41576-019-0145-z

Halees, A. S., El-Badrawi, R., and Khabar, K. S. (2008). ARED Organism: expansion of ARED reveals AU-rich element cluster variations between human and mouse. *Nucleic Acids Res.* 36, D137–D140.

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature* 553, 446–454.

Huang, K., Liang, Q., Zhou, Y., Jiang, L. L., Gu, W. M., and Luo, M. Y. (2019). A Novel allosteric inhibitor of phosphoglycerate mutase 1 suppresses growth and metastasis of non-small-cell lung cancer. *Cell Metab.* 30, 1107–1119. doi: 10.1016/j.cmet.2019.09.014

Huang, Y., Zhang, J., Zhao, Y. Y., Jiang, W., Xue, C., and Xu, F. (2012). SPARC expression and prognostic value in non-small cell lung cancer. *Chin. J. Cancer* 31, 541–548.

Koukourakis, M. I., Giatromanolaki, A., Brekken, R. A., Sivridis, E., Gatter, K. C., and Harris, A. L. (2003). Enhanced expression of SPARC/osteonectin in the tumor-associated stroma of non-small cell lung cancer is correlated with markers of hypoxia/acidity and with poor prognosis of patients. *Cancer Res.* 63, 5376–5380.

Li, Y., Xiao, J., Bai, J., Tian, Y., Qu, Y., and Chen, X. (2019). Molecular characterization and clinical relevance of m(6)A regulators across 33 cancer types. *Mol. Cancer* 18:137.

Lin, Y., Li, Z., Ozsolak, F., Kim, S. W., Arango-Argoty, G., and Liu, T. T. (2012). An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.* 40, 8460–8471. doi: 10.1093/nar/gks637

Liu, Y., Sun, J., and Zhao, M. (2017). ONGene: a literature-based database for human oncogenes. *J. Genet. Genomics* 44, 119–121. doi: 10.1016/j.jgg.2016.12.004

Masamha, C. P., Xia, Z., Yang, J., Albrecht, T. R., Li, M., and Shyu, A. B. (2014). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* 510, 412–416. doi: 10.1038/nature13261

Mayr, C., and Bartel, D. P. (2009). Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138, 673–684. doi: 10.1016/j.cell.2009.06.016

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

Ogorodnikov, A., Levin, M., Tattikota, S., Tokalov, S., Hoque, M., and Scherzinger, D. (2018). Transcriptome 3′end organization by PCF11 links alternative polyadenylation to formation and neuronal differentiation of neuroblastoma. *Nat. Commun.* 9:5331.

Pan, X. W., Gan, S. S., Ye, J. Q., Fan, Y. H., Hong, U., and Chu, C. M. (2016). SMC1A promotes growth and migration of prostate cancer in vitro and in vivo. *Int. J. Oncol.* 49, 1963–1972. doi: 10.3892/ijo.2016.3697

Park, H. J., Ji, P., Kim, S., Xia, Z., Rodriguez, B., and Li, L. (2018). 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat. Genet.* 50, 783–789. doi: 10.1038/s41588-018-0118-8

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

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* 320, 1643–1647. doi: 10.1126/science.1155390

Sarogni, P., Palumbo, O., Servadio, A., Astigiano, S., D'Alessio, B., and Gatti, V. (2019). Overexpression of the cohesin-core subunit SMC1A contributes to colorectal cancer development. *J. Exp. Clin. Cancer Res.* 38:108.

Sebestyen, E., Singh, B., Minana, B., Pages, A., Mateo, F., and Pujana, M. A. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* 26, 732–744. doi: 10.1101/gr.199935.115

Tan, S., Li, H., Zhang, W., Shao, Y., Liu, Y., and Guan, H. (2018). NUDT21 negatively regulates PSMB2 and CXXC5 by alternative polyadenylation and contributes to hepatocellular carcinoma suppression. *Oncogene* 37, 4887–4900. doi: 10.1038/s41388-018-0280-6

Tang, H., Lee, M., Sharpe, O., Salamone, L., Noonan, E. J., and Hoang, C. D. (2012). Oxidative stress-responsive microRNA-320 regulates glycolysis in diverse biological systems. *FASEB J.* 26, 4710–4721. doi: 10.1096/fj.11-197467

Tian, B., and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* 18, 18–30. doi: 10.1038/nrm.2016.116

Venkat, S., Tisdale, A. A., Schwarz, J. R., Alahmari, A. A., Maurer, H. C., and Olive, K. P. (2020). Alternative polyadenylation drives oncogenic gene expression in pancreatic ductal adenocarcinoma. *Genome Res.* 30, 347–360. doi: 10.1101/gr.257550.119

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

Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., and Wagner, E. J. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat. Commun.* 5:5274.

Xiang, Y., Ye, Y., Lou, Y., Yang, Y., Cai, C., and Zhang, Z. (2018). Comprehensive characterization of alternative polyadenylation in human cancer. *J. Natl. Cancer Inst.* 110, 379–389. doi: 10.1093/jnci/djx223

Xiong, M., Chen, L., Zhou, L., Ding, Y., Kazobinka, G., and Chen, Z. (2019). NUDT21 inhibits bladder cancer progression through ANXA2 and LIMK2 by alternative polyadenylation. *Theranostics* 9, 7156–7167. doi: 10.7150/thno.36030

Ye, C., Long, Y., Ji, G., Li, Q. Q., and Wu, X. (2018). APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* 34, 1841–1849. doi: 10.1093/bioinformatics/bty029

Ye, Y., Hu, Q., Chen, H., Liang, K., Yuan, Y., and Xiang, Y. (2019). Characterization of hypoxia-associated molecular features to aid hypoxia-targeted therapy. *Nat. Metab.* 1, 431–444. doi: 10.1038/s42255-019-0045-8

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

Yue, Y., Liu, J., Cui, X., Cao, J., Luo, G., and Zhang, Z. (2018). VIRMA mediates preferential m(6)A mRNA methylation in 3′UTR and near stop codon and associates with alternative polyadenylation. *Cell Discov.* 4:10.

Zappa, C., and Mousa, S. A. (2016). Non-small cell lung cancer: current treatment and future advances. *Transl. Lung Cancer Res.* 5, 288–300. doi: 10.21037/tlcr.2016.06.07

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.

Zhou, P., Xiao, N., Wang, J., Wang, Z., Zheng, S., and Shan, S. (2017). SMC1A recruits tumor-associated-fibroblasts (TAFs) and promotes colorectal cancer metastasis. *Cancer Lett.* 385, 39–45. doi: 10.1016/j.canlet.2016.10.041

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.

ORIGINAL RESEARCH
published: 23 March 2021
doi: 10.3389/fgene.2021.646936

A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data

Ruizhi Xiang[1], Wencan Wang[2], Lei Yang[1], Shiyuan Wang[1], Chaohan Xu[1] and Xiaowen Chen[1]**

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] School of Optometry and Ophthalmology and Eye Hospital, Wenzhou Medical University, Wenzhou, China

Single-cell RNA sequencing (scRNA-seq) is a high-throughput sequencing technology performed at the level of an individual cell, which can have a potential to understand cellular heterogeneity. However, scRNA-seq data are high-dimensional, noisy, and sparse data. Dimension reduction is an important step in downstream analysis of scRNA-seq. Therefore, several dimension reduction methods have been developed. We developed a strategy to evaluate the stability, accuracy, and computing cost of 10 dimensionality reduction methods using 30 simulation datasets and five real datasets. Additionally, we investigated the sensitivity of all the methods to hyperparameter tuning and gave users appropriate suggestions. We found that t-distributed stochastic neighbor embedding (t-SNE) yielded the best overall performance with the highest accuracy and computing cost. Meanwhile, uniform manifold approximation and projection (UMAP) exhibited the highest stability, as well as moderate accuracy and the second highest computing cost. UMAP well preserves the original cohesion and separation of cell populations. In addition, it is worth noting that users need to set the hyperparameters according to the specific situation before using the dimensionality reduction methods based on non-linear model and neural network.

Keywords: single-cell RNA-seq, dimension reduction, benchmark, sequences analysis, deep learning

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Quan Zou,
University of Electronic Science
and Technology of China, China
Kuixi Zhu,
University of Arizona, United States

****Correspondence:***
Xiaowen Chen
hrbmucxw@163.com
Chaohan Xu
chaohanxu@hrbmu.edu.cn

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

Received: *28 December 2020*
Accepted: *19 February 2021*
Published: *23 March 2021*

Citation:
Xiang R, Wang W, Yang L,
Wang S, Xu C and Chen X (2021) A
Comparison for Dimensionality
Reduction Methods of Single-Cell
RNA-seq Data.
Front. Genet. 12:646936.
doi: 10.3389/fgene.2021.646936

INTRODUCTION

The technological advances in single-cell RNA sequencing (scRNA-seq) have allowed to measure the DNA and/or RNA molecules in single cells, enabling us to identify novel cell types, cell states, trace development lineages, and reconstruct the spatial organization of cells (Hedlund and Deng, 2018). Single-cell technology has become a research hotspot. However, such analysis heavily relies on the accurate similarity assessment of a pair of cells, which poses unique challenges such as outlier cell populations, transcript amplification noise, and dropout events. Additionally, single-cell datasets are typically high dimensional in large numbers of measured cells. For example, scRNA-seq can theoretically measure the expression of all the genes in tens of thousands of cells in a single experiment (Wagner et al., 2016). Although whole-transcriptome analyses avoid the bias of using a predefined gene set (Jiang et al., 2015), the dimensionality of such datasets is typically too high for most modeling algorithms to process directly. Moreover, biological systems own the lower intrinsic dimensionality. For example, a differentiating hematopoietic cell can be represented by two or more dimensions: one denotes how far it has progressed in its differentiation toward a particular cell type, and at least another dimension denotes its current cell-cycle stage. Therefore, dimensionality reduction is necessary to project high-dimensional data into low-dimensional space to visualize the cluster structures and development trajectory inference.

Research on data dimension reduction has a long history, and principal component analysis (PCA), which is still widely used, can be traced back to 1901. Since the advent of RNA-seq technology, this linear dimension-reduction method has been favored by researchers. In addition, there are non-linear methods such as uniform manifold approximation and projection (UMAP) and t-distributed stochastic neighbor embedding (t-SNE) to reduce dimension. After the rise of neural network, there are many methods of dimensionality reduction based on neural network such as variational autoencoder (VAE). In addition, there are some new theoretical frameworks such as the multikernel learning [single-cell interpretation *via* multikernel learning (SIMLR)] based on the above methods that have been or are being developed to handle increasingly diverse scRNA-seq data.

In this study, we performed a comprehensive evaluation of 10 different dimensionality reduction algorithms comprising the linear method, the non-linear method, the neural network, model-based method, and ensemble method. These algorithms were run and compared on simulated and real datasets. The performance of the algorithms was evaluated based on accuracy, stability, computing cost, and sensitivity to hyperparameters. This work will be helpful in developing new algorithms in the field. The workflow of the benchmark framework is shown in **Figure 1**.

MATERIALS AND METHODS

Methods for Dimensionality Reduction

To our knowledge, about 10 methods are now available to obtain a low-dimensional representation for scRNA-seq data. In this section, we gave an overview of these 10 methods (**Table 1**).

PCA

As the most widely used dimensionality reduction algorithm, PCA (Jolliffe, 2002) identifies dominant patterns and the linear combinations of the original variables with maximum variance. The basic idea of PCA is to find the first principal component with the largest variance in the data and then seek the second component in the same way, which is uncorrelated with the first component and accounts for the next largest variance. This process repeats until the new component is almost ineffective or reaches the threshold set by users.

ICA

Independent component analysis (ICA) (Liebermeister, 2002), also known as blind source separation (BSS), is a statistical calculation technique used to reveal the factors behind random variables, measured values, and signals. ICA linearly transforms the variables (corresponding to the cells) into independent components with minimal statistical dependencies between them. Unlike PCA, ICA requires the source signal to meet the following two conditions: (1) source signals are independent of each other and (2) the values in each source signal have a non-Gaussian distribution. It assumes that the observed stochastic signal x obeys the model $x = As$, where s is the unknown source signal, its components are independent of each other, and A is an unknown mixing matrix. The purpose of the ICA is to estimate the mixing matrix A and the source signal s by and only by observing x.

ZIFA

The dropout events in scRNA-seq data may make the classic dimensionality reduction algorithm unsuitable. Pierson and Yau (2015) modified the factor analysis framework to solve the dropout problem and provided a method zero-inflated factor analysis (ZIFA) based on an additional zero-inflation modulation layer for reducing the dimension of single-cell gene expression data. Compared with the above two linear methods, employing the zero-inflation model can give ZIFA more powerful projection capabilities but will pay a corresponding cost in computational complexity.

In the statistical model, the expression level of the jth gene in the ith sample y_{ij} (i = 1,. . ., N and j = 1,. . .,D) is described:

$$z_i \sim \text{Normal}\,(0, I)\,,$$

$$x_i|z_i \sim \text{Normal}\,(Az_i + \mu, W)\,,$$

$$h_{ij}|x_{ij} \sim \text{Bernoulli}\left(p_0\right),$$

$$y_{ij} = \begin{cases} x_{ij},\ \text{if}\, h_{ij} = 0 \\ 0,\ \text{if}\, h_{ij} = 1 \end{cases}$$

where z_i is a $K \times 1$ data point in a latent low-dimensional space. A denotes a $D \times K$ factor loadings matrix, H is a $D \times N$ masking matrix, $W = \text{diag}(\sigma_1^2, \cdots, \sigma_D^2)$ a $D \times D$ diagonal matrix, and μ is a $D \times 1$ mean vector. Dropout probability p_0 is a function of the latent expression level, $p_0 = \exp?(-\lambda x_{ij}^2)$, where λ is the exponential decay parameter in the zero-inflation model.

Zero-inflated factor analysis adopted the expectation–maximization (EM) algorithm to infer model parameters $\Theta = (A, \sigma^2, \mu, \lambda)$ that maximize the likelihood $p\,(Y \mid \theta\,)$.

GrandPrix

GrandPrix (Ahmed et al., 2019) is based on the variational sparse approximation of the Bayesian Gaussian process latent variable model (Titsias and Lawrence, 2010) to project data to lower dimensional spaces. It requires only a small number of inducing points to efficiently generate a full posterior distribution. GrandPrix optimizes the coordinate position in the latent space by maximizing the joint density of the observation data, and then establishes a mapping from low-dimensional space to high-dimensional space.

The expression profile of each gene y is modeled as y_g is considered a non-linear function of pseudotime which

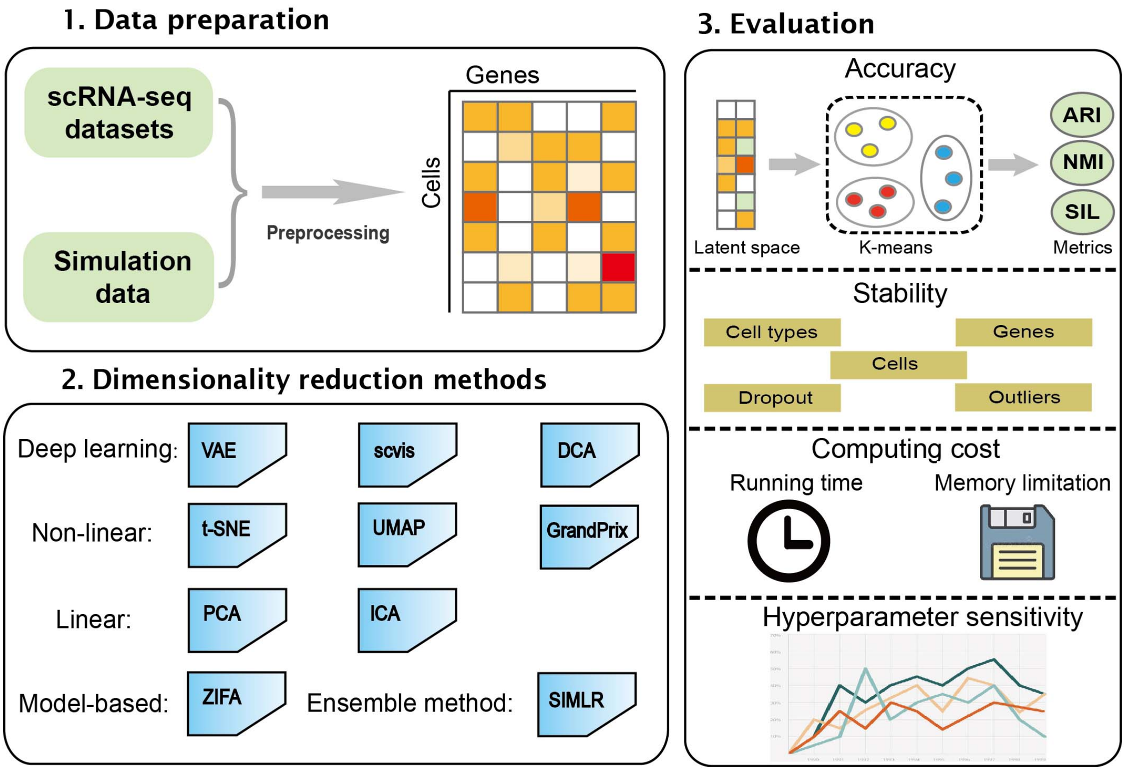


FIGURE 1 | An overview for benchmarking dimensionality reduction methods. The 10 dimensionality reduction methods were evaluated on real scRNA-seq expression datasets and simulation data. k-means was used to cluster low-dimensional latent space. The accuracy, stability, computing cost, and sensitivity to hyperparameters were used to systematically evaluate these methods.

TABLE 1 | Summary of dimensionality reduction methods.

Methods	Year	Method strategy	Platform	Input	Available URL	Version	References
PCA	1987	Linear	R	Counts	R Package Seurat	3.1.0	Jolliffe, 2002
ICA	2001	Linear	R	Counts	R Package Seurat	3.1.0	Liebermeister, 2002
ZIFA	2015	Model-based	Python	Counts	https://github.com/epierson9/ZIFA	0.1	Pierson and Yau, 2015
GrandPrix	2017	Non-linear	Python	1,000 highly genes	https://github.com/ManchesterBioinference/GrandPrix	0.1	Ahmed et al., 2019
t-SNE	2008	Non-linear	R	Counts	R Package Rtsne	0.15	Maaten and Hinton, 2008
UMAP	2018	Non-linear	R/Python	Counts	https://github.com/lmcinnes/umap	0.3.1	McInnes et al., 2018
DCA	2019	Neural network	Python	1,000 Highly genes	https://github.com/theislab/dca	0.2.2	Eraslan et al., 2019
scvis	2018	Neural network	Python	PCA-100	https://bitbucket.org/jerry00/scvis-dev	0.1.0	Ding et al., 2018
VAE	2019	Neural network	Python	Counts	https://github.com/greenelab/CZI-Latent-Assessment/tree/master/single_cell_analysis	NA	Hu and Greene, 2019
SIMLR	2017	Ensemble method	R	Counts	https://github.com/BatzoglouLabSU/SIMLR	1.6.0	Wang et al., 2017

accompanies with some noise ∈:

$$y_g = f_g(t, x) + \in$$

where

$$f_g(t, x) \sim GP(0, \sigma^2 k((t, x), (t, x)^*))$$

$\in \sim \mathrm{N}(0, \sigma^2_{\text{noise}})$ is a Gaussian distribution with variance σ^2_{noise}, x is the extra latent dimension, σ^2 is the process variance, and $k(t, t^*)$ is the covariance function between two distinct pseudotime points t and t^*. GrandPrix employed the variational free energy (VFE) approximation for inference.

t-SNE

t-Distributed stochastic neighbor embedding is a state-of-the-art dimensionality reduction algorithm for non-linear data representation that produces a low-dimensional distribution of high-dimensional data (Maaten and Hinton, 2008; Van Der Maaten, 2014). It excels at revealing local structure in high-dimensional data. t-SNE is based on the SNE (Hinton and Roweis, 2002), which starts from converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. The main idea and the modifications of t-SNE are (1) the symmetric version of SNE and (2) using a Student's *t* distribution to compute the similarity between two points in the low-dimensional space.

UMAP

Uniform manifold approximation and projection is a dimension reduction technique that can be used not only for visualization similarly to t-SNE but also for general non-linear dimension reduction. Compared with t-SNE, UMAP retains more global structure with superior run-time performance (McInnes et al., 2018; Becht et al., 2019).

The algorithm is based on three assumptions about the data: (a) the data are uniformly distributed on the Riemannian manifold; (b) the Riemannian metric is locally constant (or can be approximated); and (c) the manifold is locally connected. According to these assumptions, the manifold with fuzzy topology can be modeled. The embedding is found by searching the low-dimensional projection of the data with the closest equivalent fuzzy topology. In terms of model construction, UMAP includes two steps: (1) building a particular weighted k-neighbor graph using the nearest-neighbor descent algorithm (Dong et al., 2011) and (2) computing a low-dimensional representation which can preserve desired characteristics of this graph.

DCA

Deep count autoencoder (DCA) can denoise scRNA-seq data by deep learning (Eraslan et al., 2019). It extends the typical autoencoder approach to solve denoising and imputation tasks in in one step. The autoencoder framework of DCA is composed by default of three hidden layers with neurons of 64, 32, and 64, respectively, with zero-inflated negative binomial (ZINB) loss functions (Salehi and Roudbari, 2015), learning three parameters of the negative binomial distribution: mean, dispersion, and dropout. The inferred mean parameter of the distribution represents the denoised reconstruction and the main output of DCA. The deep leaning framework enables DCA to capture the complexity and non-linearity in scRNA-seq data. Additionally, DCA can be applied to datasets with more than millions of cells. DCA is parallelizable through a graphics processing unit (GPU) to increase the speed.

Scvis

Scvis is a statistical model to capture the low-dimensional structures in scRNA-seq (Ding et al., 2018). The assumption of scvis is a high-dimensional gene expression vector x_n of cell n which can be generated by drawing a sample from the distribution $p(x|z, \theta)$. Here, z is a low-dimensional latent vector which follows a simple distribution, e.g., a two-dimensional standard normal distribution. The data-point-specific parameters θ are the output of a feedforward neural network. To better visualize the manifold structure of an scRNA-seq dataset, scvis applies t-SNE objective function on the latent z distribution as a constraint to make cells with similar expression profiles to be close in the latent space. In addition, scvis also provides log likelihood ratio to measure the quality of embedding, which can potentially be used for outlier detection.

VAE

Variational autoencoder is a data-driven, unsupervised model for dimension reduction using an autoencoding framework, built in Keras with a TensorFlow backend (Hu and Greene, 2019). Comparing with a traditional autoencoder, VAE determined non-linear explanatory features over samples through learning two different latent representations: a mean and standard deviation vector encoding.

The model is mainly composed of two connected neural networks, encoder and decoder. The scRNA-seq data are compressed by the encoder and reconstructed by the decoder. The variable probability $Q(z|X)$ is used to approximate the posterior distribution $P(z|X)$, and it is optimized to minimize the Kullback–Leibler divergence between $Q(z|X)$ and $P(z|X)$ and reconstruction loss. Here, the encoder network is designed as a zero- to two-layer fully connected neural network to generate the mean and variance of a Gaussian distribution $q_\theta(z|X)$, and then the representative latent space z is sampled from this distribution. The decoder is also a zero- to two-layer fully connected neural network to reconstruct the count matrix.

SIMLR

Single-cell interpretation *via* multikernel learning performs dimension reduction through learning a symmetric matrix $S_{N \times N}$ that captures the cell-to-cell similarity from the input scRNA-seq data (Wang et al., 2017). The assumption of SIMLR is that $S_{N = N}$ should have an approximate block-diagonal structure with C blocks if the input cells have C cell types. SIMLR learns proper weights for multiple kernels, which are different measures of cell-to-cell distances, and constructs a symmetric similarity matrix.

Specifically, developers first define the distance between cell i and cell j as$D\left(c_i, c_j\right)$:

$$D\left(c_i, c_j\right) = 2 - 2\sum_l w_l K_l\left(c_i, c_j\right), \ \sum_l w_l = 1, \ w_l \geq 0,$$

where each linear weight w represents the importance of each kernel K, which is an expression function for cell i and cell j. In addition, SIMLR applies the following optimization framework to compute cell-to-cell similarity S:

$$\underset{S,\, L,\, W}{\text{minimize}} - \sum_{i,j,l} w_l Kl\left(c_i, c_j\right) S_{ij} + \beta ||S||_F^2 + \gamma \cdot tr\left(L^T\left(I_N - S\right)L\right)$$

$$+\rho \sum_l w_l log w_l$$

subject to

$$L^T L = I_C \sum_j w_l = 1, w_l \geq 0, \ \sum_j S_{ij} = 1 \text{ and} S_{ij} = 0$$

where I_N and I_C are $N \times N$ and $C \times C$ identification matrices, respectively, and β and γ are non-negative tuning parameters; L denotes an auxiliary low-dimensional matrix enforcing the low rank constraint on S, $tr(.)$ denotes the matrix trace, and $|S|_F$ represents the Frobenius norm of S. The optimization problem has three variables: the similarity matrix S, the weight vector w, and an $N \times C$ rank-enforcing matrix L. SIMLR solves the optimization problem through updating each variable and fixing the other two variables.

Single-cell interpretation *via* multikernel learning used the stochastic neighbor embedding (SNE) method (Maaten and Hinton, 2008) to dimension reduction based on the cell-to-cell similarity S learned from the above optimization model. However, the objective function of SIMLR involves large-scale matrix multiplication, which leads to a large amount of calculation; thus, it is difficult to extend to high-dimensional datasets.

Simulated scRNA-seq Datasets

To investigate the sensitivity of some characteristics of scRNA-seq datasets including cell type number, the number of cells and genes, outliers, and dropout event, we generated simulated datasets using the *Splatter* R package (Zappia et al., 2017). Function *splatSimulate()* is used to generate simulations, and *setParams()* is used to set specific parameters. First, we initialized the number of cell types as 5, the cell number as 2,000, the gene numbers as 5,000, and the probability of expression outlier as 0.05. When generating the simulated scRNA-seq data, we updated each parameter and fixed other parameters. Specifically, we generated the simulated data with variable numbers of cell types (5, 7, 9, 11, 13), cells (100, 500, 1,000, 2,000, 5,000, 10,000, 20,000, 30,000, 40,000, 50,000), genes (10,000, 20,000, 30,000, 40,000, 50,000), and probabilities of expression outliers (0.1, 0.2, 0.3, 0.4, 0.5). In addition, considering the impact of dropout, we also simulated datasets with five different levels of dropout (dropout.mid = −1, 0, 1, 2, 3, the larger the parameter, the more the points will be marked as 0); other parameters are set as default. Here, the probability of zero value in the data is 41, 53, 62, 71, and 80%, respectively. The detailed parameters are provided in **Supplementary Table 1**. In total, we created 30 simulated scRNA-seq datasets. The raw expression count matrices of these datasets are generated and normalized to suit for each investigated method.

Real scRNA-seq Datasets

This study analyzed five real scRNA-seq datasets, all of which were downloaded from the publicly available EMBL or GEO databases (**Supplementary Table 2**). They are derived from different species and organs, covering a variety of cell types and data dimensions. Cell types of every dataset provided in original experiments were used as a gold standard to evaluate dimension reduction methods. The descriptions of all the scRNA-seq datasets are as follows:

1. Deng dataset: isolated cells from F1 embryos from oocyte to blastocyst stages of mouse preimplantation development with six cell types were collected and sequenced by Smart-Seq2 (Deng et al., 2014).
2. Chu dataset: single undifferentiated H1 cells and definitive endoderm cells (DECs) from human embryonic stem cells sequenced by SMARTer (Chu et al., 2016).
3. Kolodziejczyk dataset: mouse embryonic stem cells from different culture conditions with three cell types (Kolodziejczyk et al., 2015). Each library was sequenced by SMARTer.
4. Segerstolpe dataset: human pancreatic islet cells with 15 cell types obtained by Smart-Seq2 (Segerstolpe et al., 2016).

Additionally, we use PBMCs from a healthy human (PBMC68k dataset) (Zheng et al., 2017) generated by the 10X Genomics platform to assess the scalability of methods.

Evaluation Metrics

To compare different dimension reduction methods, we performed the iterative k-means clustering on the low-dimensional representation of scRNA-seq data. Taking into account the randomness of k-means clustering when setting the initial cluster centroids, we performed k-means clustering 50 times to obtain a stable metric, and then set the cluster number k to the true cell type number. The evaluation metrics comparing the results to the true cell types are adjusted rand index (ARI), normalized mutual information (NMI), and Silhouette score.

Adjusted rand index (Santos and Embrechts, 2009) is a widely used metric which calculates the similarity between the two clustering results, which ranges from 0 to 1. A larger score means that two clusters are more consistent with each other. Conversely, when the clustering results are randomly generated, the score should be close to zero. Given two clustering X and Y,

$$\mathrm{ARI} = \frac{\binom{n}{2}\left(a+d\right) - \left[\left(a+b\right)\left(a+c\right) + \left(c+d\right)\left(b+d\right)\right]}{\binom{n}{2} - \left[\left(a+b\right)\left(a+c\right) + \left(c+d\right)\left(b+d\right)\right]}$$

where a is the number of objects in a pair placed in the same group in X and in the same group in Y; b is the number of objects in a pair placed in the same group in X and in different groups in Y; c is the number of objects in a pair placed in the same group in Y and in different groups in X; and d is the number of objects in a pair placed in the different groups in Y and in different groups in X.

Normalized mutual information (Emmons et al., 2016) is used to estimate the concordance between the obtained clustering and the true labels of cells. NMI value is from 0 to 1. A higher NMI refers to higher consistency with the golden standard.

Specifically, given two clustering results X and Y on a dataset, $\mathrm{NMI} = I(X, Y/max\{H\left(U\right), H(V)\}$, where

$$I\left(X, Y\right) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

$$U\left(X, Y\right) = \frac{2 \cdot I\left(X, Y\right)}{H\left(X\right), H(Y)}$$

$$H(X) = \sum_{i=1}^{n} p(x_i) I(x_i) = \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

Silhouette coefficient (Aranganayagi and Thangavel, 2007) measures how well each cell lies with its own cluster, which indicates the separability of each individual cluster. The value of Silhouette coefficient s (i) is between −1 and 1; 1 means that the cell is far away from its neighboring clusters, whereas −1 means that the cell is far away from points of the same cluster.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance from cell i to other cells in the same cluster and $b(i)$ is the average distance from cell i to all cells in other clusters. Average $s(i)$ over all the cells indicates how separable each cell type in the low-dimensional representation, which we call the Silhouette score.

Computing Cost

Computing cost of each method is estimated by monitoring the running time and peak memory usage. We analyzed the PBMC68k datasets from 10X Genomics. The raw count matrix was downsampled to 100, 500, 1,000, 2,000, 5,000, 10,000, 20,000, 30,000, 50,000, and 68,579 cells with 1,000 highly variable genes. All methods were run on the 10 downsampled datasets. We use the command *pidstat* from the sysstat tool to return the peak memory usage of the process in operation. When calculating the running time, we used the function *system.time()* in R. In this step, only the running time of the model is considered, and other processes such as data loading are excluded.

Overall Performance Score

To rank methods, the overall scores of the methods were calculated through aggregating accuracy, stability, and computing cost (Zhang et al., 2020). After k-means clustering, we used the known cell populations to calculate the ARI, NMI, and Silhouette scores for simulated data and real data, respectively. For accuracy, scaled mean ARI, scaled NMI, and scaled Silhouette scores obtained from real data were aggregated to the accuracy score. For stability, aggregated scaled scores across different simulation datasets were denoted as the stability score of each method. For the computing cost, we first scale the running time and memory usage to get a value ranging from 0 to 1. Then, we averaged scaled running time and memory usage to obtain the computing cost. Finally, we integrated the accuracy, stability, and computing cost with a ratio of 40:40:20 into the overall performance score of each method.

RESULTS

We benchmarked a total of 10 methods on 30 simulated and five real datasets. We normalized scRNA-seq data based on the corresponding method, and then performed dimensionality reduction to obtain 2D latent space. k-Means clustering method was used to perform cluster analysis. Finally, the methods were compared using accuracy, stability, computing cost, and sensitivity to hyperparameters (**Figure 1**).

Evaluation of Stability

We used 30 simulated datasets to assess the stability of the 10 dimensionality reduction methods with respect to the number of cell type, cells and genes, outliers, and dropout event.

First, we investigated the effect of cell type numbers to the approaches. We fixed the cell number (n = 2,000), gene number (n = 5,000), and probability of outliers (p = 0.05), and then changed the cell type number from 5 to 13 stepped by 2. As the number of cell types increased, the performance of PCA, ICA, and GrandPrix descended faster (**Figure 2A**). While the performance of ZIFA, VAE, SIMLR, scvis, and DCA decreased slightly, UMAP and t-SNE fluctuated. Generally, ZIFA, VAE, SIMLR, scvis, DCA, UMAP, and t-SNE have better stability with respect to cell type number than PCA, ICA, and GrandPrix, since their standard deviation is relatively small.

Second, we changed the cell number from 100 to 50,000 and fixed other factors. It was found that too many or too few cells are not conducive to the construction of low-dimensional space of single-cell RNA-seq data. All the methods' performance fluctuated greatly except for PCA and UMAP. PCA and UMAP have strong adaptability to cell number change based on standard deviation (**Figure 2B**). All of the methods obtained the best performance between 1,000 and 10,000 cells. It is worth noting that SIMLR has a high computational complexity as it involves large matrix operations which could not perform dimensionality reduction on data with a cell count greater than or equal to 10,000. Additionally, all the methods except PCA and ZIFA have good stability with respect to gene number (**Figure 2C**).

To investigate the effect of the complex cell mixtures to methods, we simulated expression outliers; it was found that the performance of all the methods is stable to expression outliers (**Figure 3A**). Finally, we randomly dropped expressed genes in each cell to investigate the ability of methods to deal with datasets with various library sizes. Generally, ZIFA, VAE, UMAP, t-SNE, SIMLR, and GrandPrix showed a stable performance, whereas the performance of scvis, PCA, ICA, and DCA decreased remarkably with the increase in the dropout ratio (**Figure 3B**).

We found that the stability of each method is different with respect to the number of cell types, cells and genes, outliers, and dropout rate. To evaluate the overall stability of each method, we aggregated all the metrics across simulation datasets to obtain the overall stability score (see section "Materials and Methods"). In summary, the overall stability scores showed that the performance of UMAP has shown more stability than the other methods. Conversely, ICA has poor stability (**Figure 4**). It is worth mentioning that the Silhouette score of UMAP is significantly higher than the other methods in all simulation tests, indicating that it better separated distinct cell types.

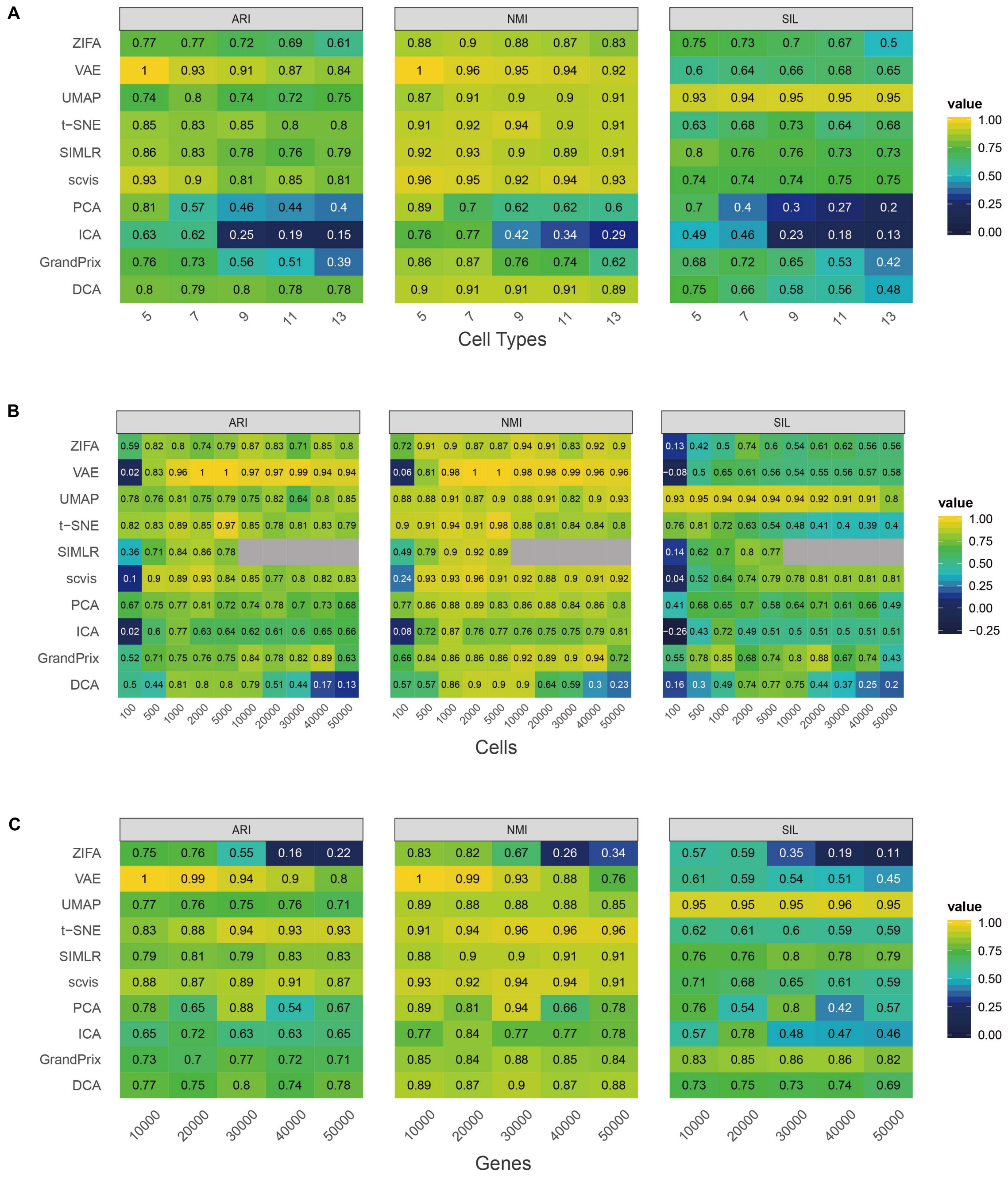


FIGURE 2 | Evaluation stability of the 10 dimensionality reduction methods on simulated scRNA-seq data with respect to the number of cell type **(A)**, cell number **(B)**, or gene number **(C)**. The performance is measured by ARI, NMI, and Silhouette score (SIL). Gray indicates that the SIMLR cannot run on data with more than 10,000 cells.

Evaluation of Accuracy

We applied the 10 dimensionality reduction methods to the four real data and performed k-means cluster analysis based on the low-dimensional representation and calculated the evaluation metrics. No single method dominated on all of these datasets, indicating that there is no "one-size-fits-all" method that works well on every dataset. Regarding the ARI and NMI measures, PCA and t-SNE were ranked in the top five performers on all the four datasets (**Figures 5A,B**). VAE was ranked in the top five performers on the three datasets. Consistent with the simulation dataset, UMAP can separate each individual cluster very well based on the Silhouette score, compared with other methods (**Figure 5C**). In addition, the dataset of Segerstolpe et al. has the lowest evaluation metrics compared with the other three datasets,

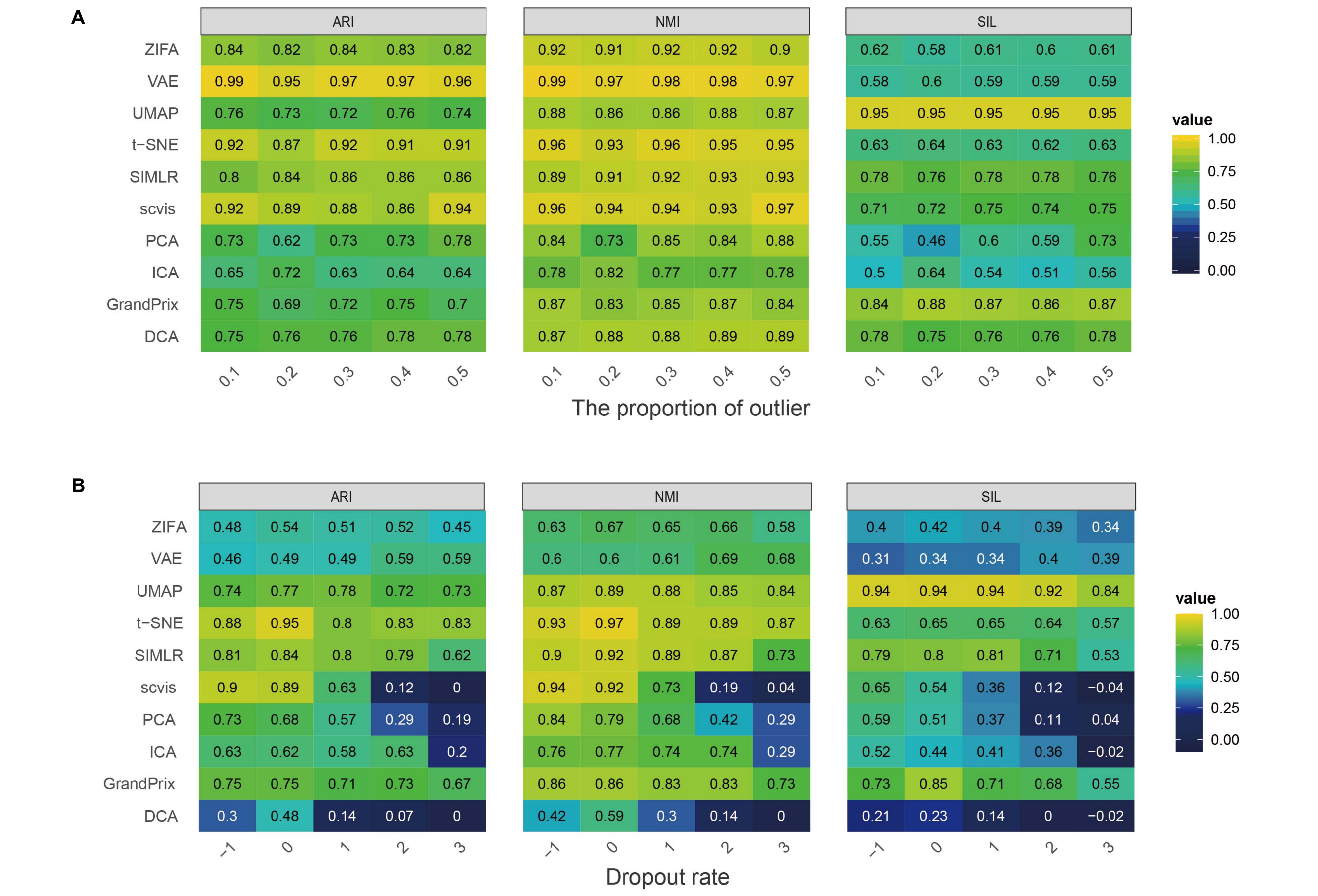


FIGURE 3 | Evaluation stability of the 10 dimensionality reduction methods on simulated scRNA-seq data with respect to the proportion of outlier **(A)** or dropout rate **(B)**. The performance is measured by ARI, NMI, and SIL.

indicating that the dimensionality reduction method should be improved for the heterogeneous dataset with more cell types. We also visualized the low-dimensional reductions of all the methods on the four datasets (**Supplementary Figures 1–4**). The ability to separate different cell types of each method is consistent with the above metrics. Aggregating all the three metrics across datasets, t-SNE has the best accuracy, followed by VAE (**Figure 4**).

Sensitivity of Methods to Hyperparameters

The hyperparameters play a crucial part of the dimension reduction algorithm, especially the deep machine learning model. Therefore, we examined the effect of the hyperparameter settings on the dimensionality reduction in order to guide the user in making a reasonable choice. Among all the 10 algorithms discussed, there are seven methods whose developers have added parameter settings. PCA and ICA are based on linear transformations, so do not require hyperparameter adjustment. In addition, DCA implements an automatic search that could identify a set of hyperparameters in minimizing errors. To decrease time consumption, we used the datasets of Deng to investigate the effect of the hyperparameters to the performance of these seven methods. Detailed evaluation parameters are shown in **Supplementary Table 3**. Using grid search strategy, we found that ZIFA is insensitive to their respective hyperparameters, and the evaluation metrics have little change in different settings (**Figure 6A**). The evaluation metrics of t-SNE and SIMLR increased when their hyperparameters increased from 2 to 5, after that ARI and NMI tend to be stable. Silhouette scores are largely reduced when the hyperparameters are larger than 20 (**Figures 6B,C**). For those methods with multiple adjustable hyperparameters including GrandPrix, scvis, UMAP, and VAE, we noticed a dramatic change in the results when choosing different hyperparameter settings (**Figures 6D–G**). Therefore, we recommend that users consider the impact of hyperparameter settings before using these four methods.

Data Preprocessing of All Methods

For the arithmetic design adapting to different algorithms, we performed the corresponding normalization process for one raw single-cell RNA-seq data based on the description of the algorithm. First, PCA, ICA, t-SNE, UMAP, ZIFA, and SIMLR used the original count

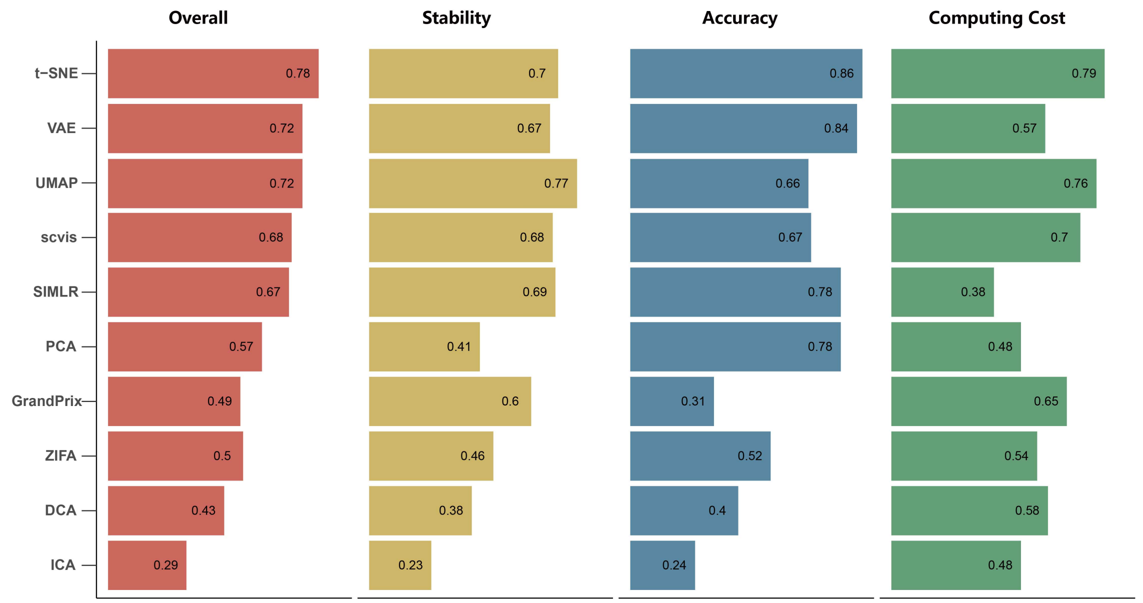


FIGURE 4 | The overall performance of the 10 dimensionality reduction algorithms. The methods are sorted by overall performance score, which is a weighted integration of accuracy, stability, and computing cost. The accuracy and stability are the average value of scaled ARI, scaled NMI, and scaled SIL in real data and simulated data, respectively. Running time and memory are scaled to a value in [0,1] before averaged as computing cost.

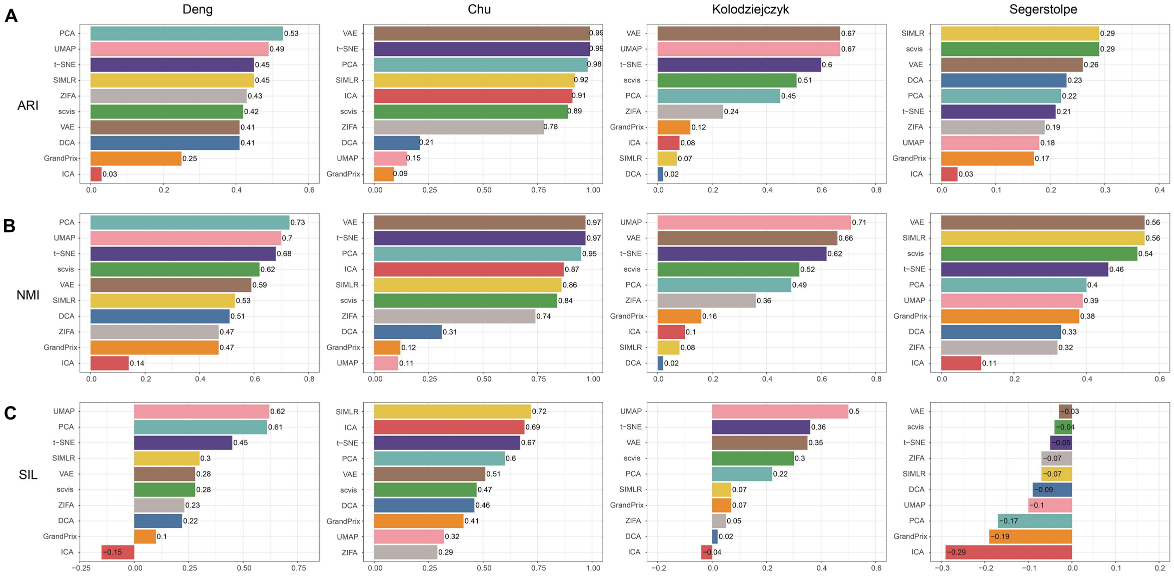


FIGURE 5 | Evaluation accuracy of the 10 dimensionality reduction methods on real scRNA-seq data measured by **(A)** ARI, **(B)** NMI, and **(C)** SIL.

matrix of scRNA-seq data as the input. For DCA and GrandPrix, the input is a feature matrix with all the cells and 1,000 highly variable genes. Scvis used PCA as a preprocessing for noise reduction to project the cells into a 100-dimensional space.

The Outputs of All Methods

For some methods, in addition to the low-dimensional representation of the data, other useful information is also provided. Specifically, scvis, DCA, and VAE were developed based on deep learning; thus, a trained model is saved in the corresponding output folder, containing the loss parameters and validation for models. Furthermore, being used as a process of noise reduction, DCA provides an output file which represents the mean parameter of the ZINB distribution which has the same dimensions as the input file. Detailed workflows and explanations are available in the original publications.

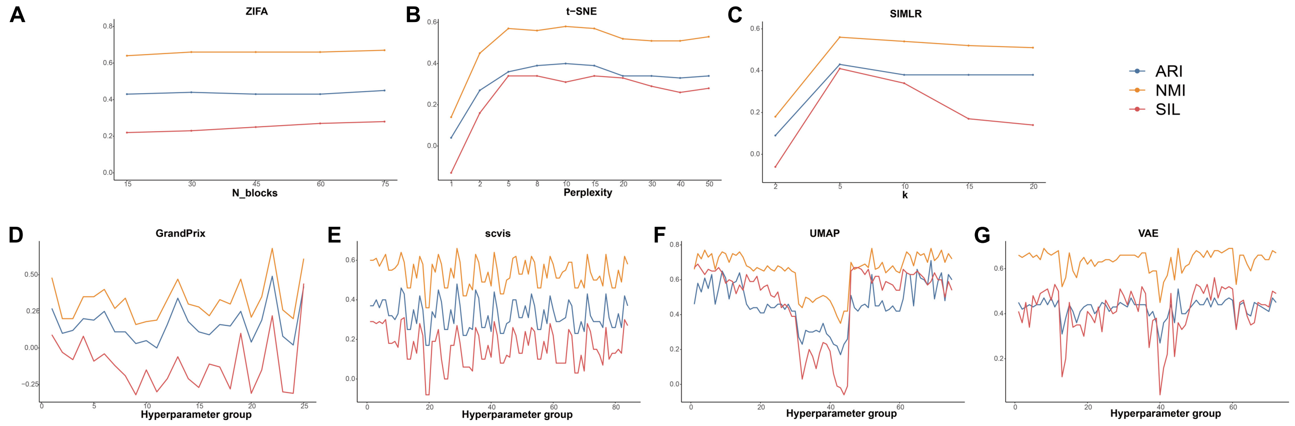


FIGURE 6 | The effect of hyperparameters to the performance of dimensionality reduction methods. **(A)** ZIFA. **(B)** t-SNE. **(C)** SIMLR. **(D)** Grandprix. **(E)** Scvis. **(F)** UMAP. **(G)** VAE.

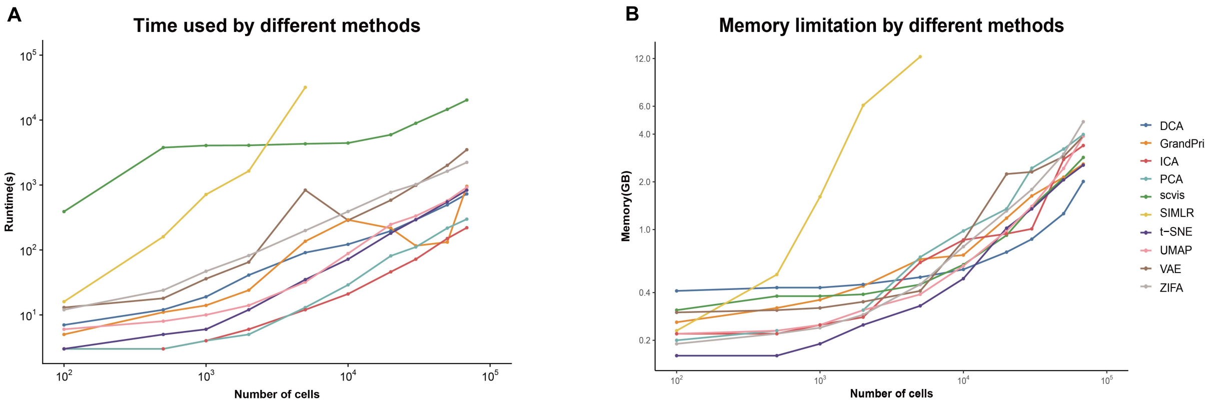


FIGURE 7 | Evaluation computing cost for each method on metrics of **(A)** running time and **(B)** memory limitation. The analyses were run on computing equipment with Inter i7 4790@3.60 GHz CPU and 16G running memory.

Computing Cost Overview

The current scRNA-seq analysis methods are expected to cope with hundreds of thousands of cells as the number of cells profiling by the current protocols increases. We estimated the computational efficiency of each method using running time and memory usage. We generated ten datasets containing different number of cells through downsampling the PBMC68k data. Overall, the running time and memory usage of all methods are positively correlated with the cell number. Most methods except SIMLR and scvis can be completed in 30 min even using all the cells of PBMC68k dataset (**Figure 7A**). Most methods except SIMLR and ZIFA can complete all the processes within 4 GB (**Figure 7B**). We noted that SIMLR is difficult to be performed on the dataset with more than 10,000 cells due to its unique multikernel matrix operation. In general, ICA took the shortest time (3.7 min) and t-SNE had the lowest memory requirements (2.5 GB) when the number of cells is 68k. Overall, t-SNE has the best computing cost (**Figure 4**).

Overall Performance

By integrating three metrics from measurement of accuracy, stability, and computing cost, we obtained the overall performance score for each method (**Figure 4**). We found that t-SNE achieved the best overall performance score with the highest accuracy and computing cost. Meanwhile, UMAP exhibited the highest stability, as well as moderate accuracy and the second highest computing cost. However, the performance score of these methods is different across evaluation criteria. For example, SIMLR and PCA performed better than UMAP based on accuracy, while SIMLR showed weaker computing cost and PCA showed weaker stability.

DISCUSSION

Since 2015, the emergence of 10X Genomics, Drop-seq, Micro-well, and Split-seq technologies has completely reduced the cost of single-cell sequencing. This technology has been widely used in basic scientific and clinical research. An important application of single-cell sequencing is to identify and characterize new cell types and cell states. In this process, the key question is how to measure the similarity of the expression profiles of a set of cells, whereas, such similarity analysis can be improved after reducing dimensionality, which can help in noise reduction.

Here, we performed a comprehensive evaluation of 10 dimensionality reduction methods using simulation and real dataset to examine the stability, accuracy, computing cost, and sensitivity to hyperparameters. Taken together, we observed that the summarized performance of t-SNE outperformed the performance of other methods. UMAP has the highest stability and can separate distinct cell types very well. Although, both methods are not specifically designed for single-cell expression data. However, the performance of most methods decreased as cell number and dropout rate increased. Therefore, new algorithms will likely be needed to effectively deal with dropout rate and millions of cells. In addition, the dataset from Segerstolpe et al. containing the lower evaluation metrics showed that the dimensionality reduction method should be improved for the heterogeneous dataset with more cell types. We suggested that users adjust the hyperparameters when using these non-linear and neural network methods. Finally, basic linear methods such as PCA and ICA have shown to be most time saving but perform worse in highly heterogeneous data.

To conclude, we provide a new procedure for comparing single-cell dimensionality reduction methods. We hope that this will be useful in providing and giving method users and algorithm developers an exhaustive evaluation of different data and appropriate recommendation guidelines. At the same time, new dimensionality reduction methods are being developed which will become more robust and standardized. These developments will deepen further exploration and comprehensive understanding of single-cell RNA-seq applications.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

RX performed data analysis, data visualization, and manuscript writing. WW, LY, and SW participated in the discussion. CX and XC supervised the project and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Heilongjiang Postdoctoral Research Startup Foundation (Grant Nos. LBH-Q19116 and LBH-Q19118) and Heilongjiang Natural Science Fund Project (Grant No. LH2019C087). This work was also supported in part by the National Natural Science Foundation of China (Grant Nos. 31871336, 61573122, and 31701159), the Wu Lien-Teh Youth Science fund project of Harbin Medical University (Grant No. WLD-QN1407), the Health Department Science Foundation of Heilongjiang Province (Grant No. 2013128), the Education Department Science Foundation of Heilongjiang Province (Grant No. 12541415), and the Postdoctoral project of Heilongjiang Province (Grant No. LBH-Z14130).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.646936/full#supplementary-material

REFERENCES

Ahmed, S., Rattray, M., and Boukouvalas, A. (2019). GrandPrix: scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics* 35, 47–54. doi: 10.1093/bioinformatics/bty533

Aranganayagi, S., and Thangavel, K. (2007). “Clustering categorical data using silhouette coefficient as a relocating measure,” in *Proceeding of The International Conference on Computational Intelligence And Multimedia Applications (ICCIMA 2007)*, (New York, NY: IEEE), 13–17. doi: 10.1109/ICCIMA.2007.328

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A. I, Kwok, W., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314

Chu, L. F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17:173. doi: 10.1186/s13059-016-1033-x

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343, 193–196. doi: 10.1126/science.1245316

Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* 9:2002. doi: 10.1038/s41467-018-04368-5

Dong, W., Moses, C., and Li, K. (2011). “Efficient k-nearest neighbor graph construction for generic similarity measures,” in *Proceedings of the 20th International Conference on World Wide Web*, (Hyderabad: WWW), 577–586. doi: 10.1145/1963405.1963487

Emmons, S., Kobourov, S., Gallant, M., and Borner, K. (2016). Analysis of network clustering algorithms and cluster quality metrics at scale. *PLoS One* 11:e0159161. doi: 10.1371/journal.pone.0159161

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* 10:390. doi: 10.1038/s41467-018-07931-2

Hedlund, E., and Deng, Q. (2018). Single-cell RNA sequencing: technical advancements and biological applications. *Mol. Aspects Med.* 59, 36–46. doi: 10.1016/j.mam.2017.07.003

Hinton, G., and Roweis, S. T. (2002). Stochastic neighbor embedding. *NIPS* 15, 833–840.

Hu, Q., and Greene, C. S. (2019). Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *Pac. Symp. Biocomput.* 24, 362–373. doi: 10.1101/385534

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

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York, NY: Springer, doi: 10.1007/b98835

Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C., Ilicic, T., Henriksson, J., Natarajan, K. N., et al. (2015). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17, 471–485. doi: 10.1016/j.stem.2015.09.011

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 18, 51–60. doi: 10.1093/bioinformatics/18.1.51

Maaten, L.v.d, and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. doi: 10.1007/s10846-008-9235-4

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv* [Preprint] arXiv:1802.03426. https://arxiv.org/abs/1802.03426,

Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16:241. doi: 10.1186/s13059-015-0805-z

Salehi, M., and Roudbari, M. (2015). Zero inflated poisson and negative binomial regression models: application in education. *Med. J. Islam. Repub. Iran* 29:297.

Santos, J. M., and Embrechts, M. (2009). *On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification*. Berlin: Springer, 175–184. doi: 10.1007/978-3-642-04277-5_18

Segerstolpe, A., Palasantza, A., Eliasson, P., Andersson, E. M., Andreasson, A. C., Sun, X., et al. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* 24, 593–607. doi: 10.1016/j.cmet.2016.08.020

Titsias, M., and Lawrence, N. D. (2010). "Bayesian Gaussian process latent variable model," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, (Norfolk MA: JMLR), 844–851.

Van Der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* 15, 3221–3245.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34, 1145–1160. doi: 10.1038/nbt.3711

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14, 414–416. doi: 10.1038/nmeth.4207

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 18:174. doi: 10.1186/s13059-017-1305-0

Zhang, Y., Ma, Y., Huang, Y., Zhang, Y., Jiang, Q., Zhou, M., et al. (2020). Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput. Struct. Biotechnol. J.* 18, 2953–2961. doi: 10.1016/j.csbj.2020.10.007

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049. doi: 10.1038/ncomms14049

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 March 2021
doi: 10.3389/fgene.2021.653209

Exploring the Imbalance of Periodontitis Immune System From the Cellular to Molecular Level

*Longfei He[1,2], Lijuan Liu[2], Ti Li[2], Deshu Zhuang[1,3], Jiayin Dai[1], Bo Wang[1] and Liangjia Bi[1]**

[1] Department of Stomatology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China, [2] Department of Stomatology, Weifang People's Hospital, Weifang, China, [3] Department of Oral Biological and Medical Sciences, Faculty of Dentistry, University of British Columbia, Vancouver, BC, Canada

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Yi Zhang,
Dana-Farber Cancer Institute,
United States
Chong Jin,
University of Pennsylvania,
United States
Tingfang Chen,
University of Pennsylvania,
United States

***Correspondence:**
Liangjia Bi
biliangjia66@163.com

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

Received: 14 January 2021
Accepted: 08 March 2021
Published: 26 March 2021

Citation:
He L, Liu L, Li T, Zhuang D, Dai J,
Wang B and Bi L (2021) Exploring
the Imbalance of Periodontitis
Immune System From the Cellular
to Molecular Level.
Front. Genet. 12:653209.
doi: 10.3389/fgene.2021.653209

Periodontitis is a common chronic inflammatory disease of periodontal tissue, mostly concentrated in people over 30 years old. Statistics show that compared with foreign countries, the prevalence of periodontitis in China is as high as 40%, and the prevalence of periodontal disease is more than 90%, which must arouse our great attention. Diagnosis and treatment of periodontitis currently rely mainly on clinical criteria, and the exploration of the etiologic criteria is relatively lacking. We, therefore, have explored the pathogenesis of periodontitis from the perspective of immune imbalance. By predicting the fraction of 22 immune cells in periodontitis tissues and comparing them with normal tissues, we found that multiple immune cell infiltration in periodontitis tissues was inhibited and this feature can clearly distinguish periodontitis from normal tissues. Further, protein interaction network (PPI) and transcription regulation network have been constructed based on differentially expressed genes (DEGs) to explore the interaction function modules and regulation pathways. Three functional modules have been revealed and top TFs such as EGR1 and ETS1 have been shown to regulate the expression of periodontitis-related immune genes that play an important role in the formation of the immunosuppressive microenvironment. The classifier was also used to verify the reliability of periodontitis features obtained at the cellular and molecular levels. In conclusion, we have revealed the immune microenvironment and molecular characteristics of periodontitis, which will help to better understand the mechanism of periodontitis and its application in clinical diagnosis and treatment.

Keywords: periodontitis, DEGs, crosstalk gene, PPI, immune system

INTRODUCTION

Periodontitis is a chronic inflammatory disease with complex pathogenesis. It will gradually cause the loss of periodontal ligament and alveolar bone, and eventually cause tooth loss (Hajishengallis, 2015; Hajishengallis and Korostoff, 2017). As one of the most prevalent chronic inflammatory diseases in the world, periodontitis directly affects more than 11% of the global population. According to the National Health and Nutrition Examination Survey of the United States, nearly half of American adults suffer from periodontitis, which is a huge number and seriously affects the quality of life of individuals (Eke et al., 2015). Recent studies have shown that periodontitis not only affects the periodontal area, it is also the cause of other systemic diseases, such as rheumatoid arthritis, atherosclerosis and cerebrovascular diseases (Genco and Van Dyke, 2010; Kebschull et al., 2010; Lundberg et al., 2010). In addition, studies have found that as many as one-third of the periodontitis mutations in the population are caused by genetic factors, and the more severe the periodontitis, the stronger the heritability (Nibali et al., 2019).

Studies have confirmed that infection of external microbial flora is an important factor in causing periodontitis. Earlier, Porphyromonas gingivalis was considered to be the cause of periodontitis. But with the advancement of science, we have found that periodontitis induced by Porphyromonas gingivalis requires the presence of symbiotic flora (Hajishengallis et al., 2011). Although with the study of the etiology of periodontitis is more detailed, the most important is the local microbiota and host immune response (Hajishengallis, 2014a). Under normal physiological conditions, the host periodontal local immune response and microbes are in a delicate balance state, realizing routine monitoring of the flora (Graves et al., 2019). However, once the pathogen colonizes the periodontal area, it will significantly increase the number and destructiveness of the microbial flora, breaking the original dynamic balance (Hajishengallis et al., 2012). Under this condition, the immunity will be over-activated and immune invasion will occur, thereby destroying the activity of periodontal tissues. Different from the immune evasion of other pathogens (Cyktor and Turner, 2011), the periodontitis flora interacts with the immune response to improve its adaptability and use the tissues destroyed by inflammation to obtain nutrients (Hajishengallis, 2014a,b).

After all, the process of periodontitis is caused by the dynamic imbalance of local immunity and microbial community. Immune invasion will cause the activation of osteoclasts, which will resorb alveolar bone (Belibasakis and Bostanci, 2012). The abnormality of cytokines in the host immune response has been revealed in previous studies (Pan et al., 2019). Cytokines are key regulators of local tissue homeostasis and inflammatory processes, playing a role in the first wave of the host's response to pathogens and stimuli, and connect tissue cells with lymphocytes and helper cell populations to work together (Graves, 2008). The immune imbalance of periodontitis leads to systemic inflammation (Hajishengallis, 2015), and a large number of studies on the pathogenesis of periodontitis involve changes in host immunity. But so far, no scholar has fully revealed the immune imbalance of periodontitis from cells to molecules. In this study, we will reveal the new pathogenesis of periodontitis and the abnormal molecular mechanism through protein interaction analysis and targeted regulation analysis of related immune genes.

MATERIALS AND METHODS

Data Collection

The expression profile and sample annotation of periodontitis diseases was downloaded from the GEO database[1], including three series GSE10334 (183 periodontitis and 64 normal), GSE16134 (241 periodontitis and 69 normal) and GSE23586 (3 periodontitis and 3 normal, **Table 1**). Next, we download all immunosuppressive-related genes from DisGeNET (Pinero et al., 2017)[2] and HisgAtlas (Liu et al., 2017)[3]. In addition, we searched for drugs related to immunosuppressive agents from Drugbank (Wishart et al., 2018)[4] obtained 311 immunosuppressive-related drugs, and then downloaded immunosuppressive-related genes. We merged the immunosuppressant-related genes obtained from the above three databases, and obtained a total of 1,332 genes. We started from BIND (Gilbert, 2005), BioGRID (Oughtred et al., 2019)[5], MINT (Chatr-aryamontri et al., 2007)[6], HPRD (Goel et al., 2012)[7], IntAct (Kerrien et al., 2012)[8], and OPHID (Brown and Jurisica, 2005)[9] database to download protein interaction data, and integrate these data. We also downloaded immune-related genes from the InnateDB (Breuer et al., 2013)[10] database. The transcription factor (TF) and target gene relationship from the relevant transcription regulation databases TRRUST v2 (Yang et al., 2018)[11] and ORTI (Vafaee et al., 2016)[12].

Immune Cell Distribution Analysis

We have preprocessed the expression matrices of the three series of GSE10334, GSE16134, and GSE23586 and extracted the expression profiles of immunosuppressant-related genes in periodontitis diseases for immune invasion analysis. CIBERSORT (Newman et al., 2015) could be used to predict the infiltrating immune cells that are highly related to periodontitis disease. Here, we used the R version of CIBERSORT instead of the web version, taking into account the user-friendly operation. CIBERSORT has four parameters including the reference set that can be downloaded at https://cibersort.stanford.edu/download.php, the expression matrix we prepared, perm that is the number of permutations when calculating the *p*-value and is set to 1,000, and QN that is

TABLE 1 | Description of microarray profiles in gingival tissue.

GEO series	Periodontitis Sample	Normal sample	Tissues	Platforms	Citation (PMID)
GSE10334	183	64	Gingival	Affymetrix; GPL570	18980520
GSE16134	241	69	Gingival	Affymetrix; GPL570	19835625, 24646639
GSE23586	6	6	Gingival	Affymetrix; GPL570	21382035

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

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

[3] http://biokb.nb.org/HisgAtlas/

[4] https://www.drugbank.ca/

[5] http://thebiogrid.org/

[6] http://mint.bio.uniroma2.it/mint/

[7] http://www.hprd. org/

[8] https://www.ebi.ac.uk/intact/

[9] http://ophid.utoronto.ca/ophidv2.204/

[10] https://www.innatedb.ca/

[11] https://www.grnpedia.org/trrust/

[12] http://orti.sydney.edu.au/about.html

whether to perform quantile normalization and is set to TRUE taking into account the microarray expression data. In order to see more group differences in the fraction of cell types other than plasma cells, we further transformed the raw cell fractions into the log ratio of log (plasma_cell_fraction + 1e-3)/log (cell_fraction + 1e-3). We also combined previous studies on periodontitis clustering to explore the differences in the immune microenvironment between periodontitis subtypes.

Differential Expression Analysis and Functional Enrichment Analysis

We consider the sample size of each series in the downloaded data, so we only perform differential expression analysis on the downloaded sample data of GSE10334 and GSE16134. In data preprocessing, missing values of the expression matrix were filled by zero value. Further, the gene expression values were log2-transformed to be suitable for differential expression analysis. The limma package was used to measure gene expression variation between periodontitis and normal samples. We defined the cutoff of gene *p*-value as 0.05 and the cutoff of fold-change as 1.5 (Demmer et al., 2008), which filtered out differentially expressed genes (DEGs). The clinical variables were not include in the DEG identification pipeline Next, we integrate the significant DEGs of these two series of samples into a multi-gene set list, and use the compareCluster_go() function of the latest clusterProfiler package of the R language to perform GO function and KEGG enrichment on the data set, and set threshold $p < 0.05$.

Construction of PPI Network and Transcriptional Regulatory Network

We extract gene pairs that interact with DEGs from the PPI data, and use the network rendering tool Cytoscape to map the differential gene PPI data. Further, the MCODE module of Cytoscape were used to screen the significant function modules in the DEG PPI network (parameter selection: Degree cutoff: 5, Node score cutoff: 0.2, K-core: 2, and Max. depth: 100), and used the network analysis tool to analyze the topological properties of the network (Degree, Average Shortest Path Length, Betweenness Centrality, Closeness Centrality, Clustering Coefficient, Topological Coefficient). We use differentially expressed immune genes as crosstalk genes, and extract the PPI relationship pairs of these crosstalk genes, and use Cytoscape to construct the crosstalk gene PPI network. We defined the modules identified in the PPI network of immune-related genes that were masked in the PPI network constructed directly using DEGs as New-module of immune function. We extracted the TF-target relationship pairs related to the crosstalk gene and constructed the TF-target network using Cytoscape software. We then analyzed the topological properties of the network, and extracted the top 10 genes of outdegree and indegree, respectively, as key periodontitis related genes.

Build the Classifier

We constructed periodontitis disease classifiers with significantly different infiltration of immune cells as the characteristic and New-module functional gene in the crosstalk gene PPI network as the characteristic. The former uses the fraction of immune cell identified by CIBERSORT and the latter uses gene expression data. Here, we consider two classification algorithms, including decision tree and SVM, to build the model. We randomly select 70% of the samples in GSE10334 as the training set, and the remaining 30% as the test set, and use the data of the GSE16134 and GSE23586 series as the validation sets. Further, we combine the possibility provided by the classifier and the true sample label to measure the performance of the classifier. In order to understand the generalization ability of the model, we introduced fivefold cross-validation. We use the pROC package and plot function of the R language to display the ROC curve to evaluate the effectiveness of the model.

RESULTS

Immune System Imbalance at the Cellular Level

Immune Cell Infiltration in Periodontitis

We developed a computational pipeline to analyze the gene expression profile of periodontitis disease (**Figure 1A**). In this study, we selected microarray profiles of the GSE10334 and GSE16134 series with sufficient periodontitis and normal samples for immuno-infiltration analysis of gingival tissue. After quality control and normalization, we obtained two processed expression profiles. Here, we used the CIBERSORT method to predict the infiltration of immune cells in periodontitis disease. We obtained the fraction of 22 immune cell types in these samples. We further transformed the raw cell fractions in order to see more group differences in the fraction of cell types (**Figures 1B,C** and **Supplementary Table S1**). We found decreased levels of immune infiltration during the malignant transformation of normal tissue to periodontitis that was verified in both series of samples, which indicates that periodontitis tissue undergoes immunosuppressive microenvironment. By combining this with previous studies (Kebschull et al., 2014), we found that the level of immune infiltration in type 1 periodontitis was superior to that in type 2 periodontitis (**Figure 1B**), indicating that type 1 periodontitis may be more suitable for immune targeted therapy. We found that the fraction of CD4+/CD8+ T cells in periodontitis tissue was significantly depressed (**Figure 1D**), which might be one of the factors contributing to the suppression of the immune microenvironment in periodontitis tissues.

Construct a Classifier Based on Immune Cells

In order to consider whether immune cells with significant changes in fraction can represent the overall difference between periodontitis and normal patients, we constructed a classifier based on the significantly different distribution of immune cells. The two machine learning methods, including Decision tree and SVM, were used to build the classifier model, and the training set, test set, and validation set were also scientifically allocated. In the model constructed by the decision tree, dendritic cell, neutrophils, and CD4+/CD8+T cell were used as important screening indicators to control sample filtering (**Figure 2A**). In order to predict the accuracy of the model, the data of the test set

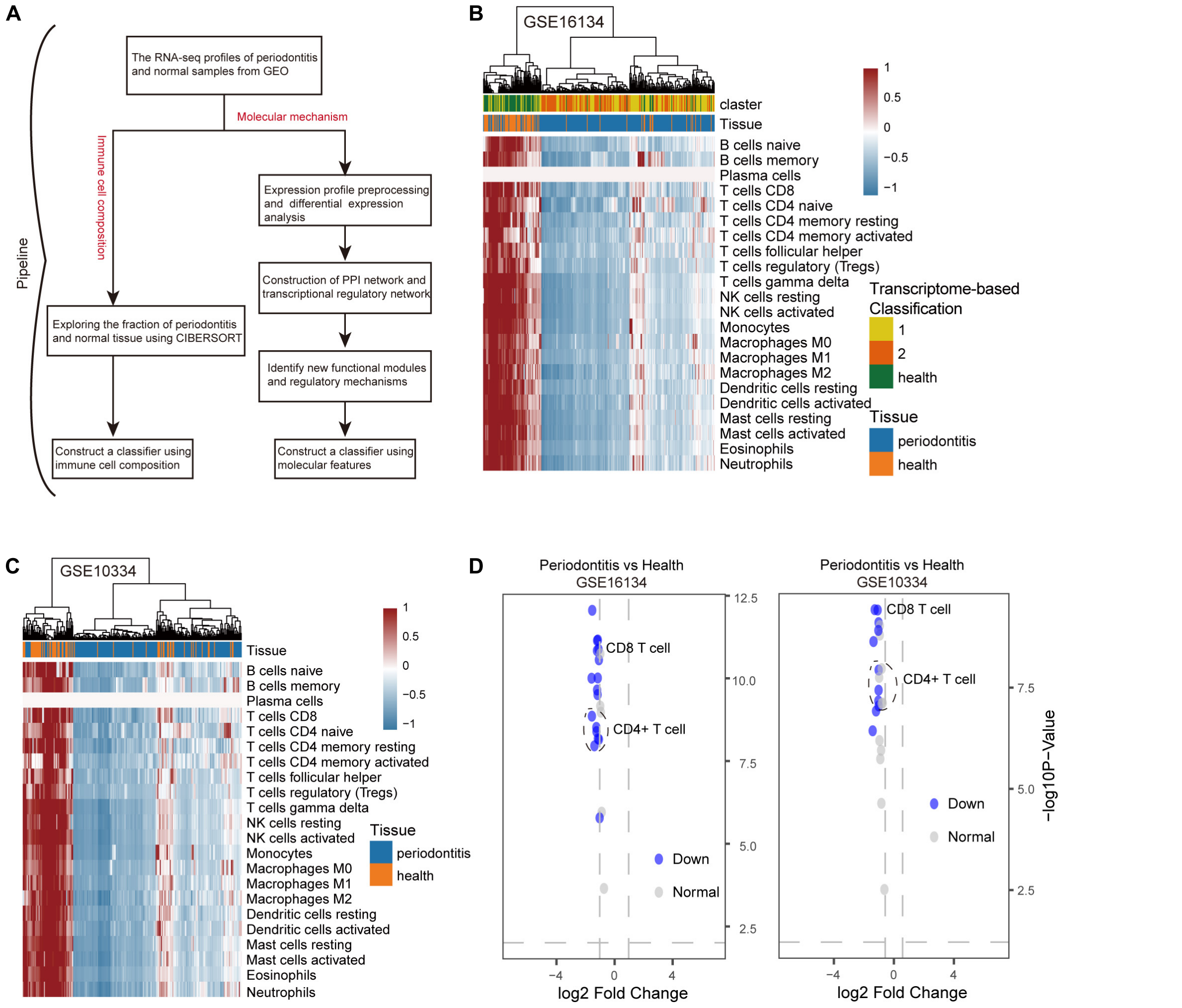


FIGURE 1 | The distribution of 22 types of immune cells in periodontitis and healthy samples. **(A)** Diagram of the multiple components and workflows of pipeline. **(B)** The heatmap represents the fraction of immune cells for the GSE16134 series. The horizontal axis is the immune cell type and the vertical axis is the sample. **(C)** The same as in **(B)** but for GSE10334. **(D)** The volcano plot represents the immune cells with significantly different gene expression levels between periodontitis and healthy samples for the GSE10334 and GSE16134 series.

and the validation set were verified by a trained classifier, and the prediction results are output. Then we use the pROC package and plot function of the R language to display the ROC curve of the data set to evaluate the effectiveness of the model.

After the construction of the classifier and the evaluation of the classification efficiency, we found that the classifier constructed by the SVM algorithm has a slight advantage over the classifier constructed by the Decision tree algorithm (**Figures 2B,C**). In order to measure the generalization ability of the support vector machine model, we introduced fivefold cross-validation. We found that the AUC value of the fivefold verification result is stable (**Supplementary Figure S1**), indicating that the choice of hyperparameters of the model is excellent. We obtained excellent results in differentiating periodontitis from normal tissue from the perspective of immune cells, suggesting that the disruption of the immune microenvironment of the gingival tissue is an important cause of periodontitis. Further, the exploration of the molecular mechanisms underlying the formation of the immunosuppressive microenvironment in periodontitis is crucial.

Immune System Imbalance at the Molecular Level

Statistical Analysis of Gene Expression Matrix

First, we performed statistical tests on the expression profile data of the GSE10334 and GSE16134 series with abundant sample sizes, and calculated two test indicators *P*-value and Fold Change. We obtained 1,571 and 1,680 DEGs from the two series of GSE10334 and GSE16134, respectively (**Figures 3A,B**). From the results, we found that there are a large number of DEGs between periodontitis samples and normal samples. In order to evaluate the reliability of the experimental data, we tested the overlap levels of the up-regulated and down-regulated

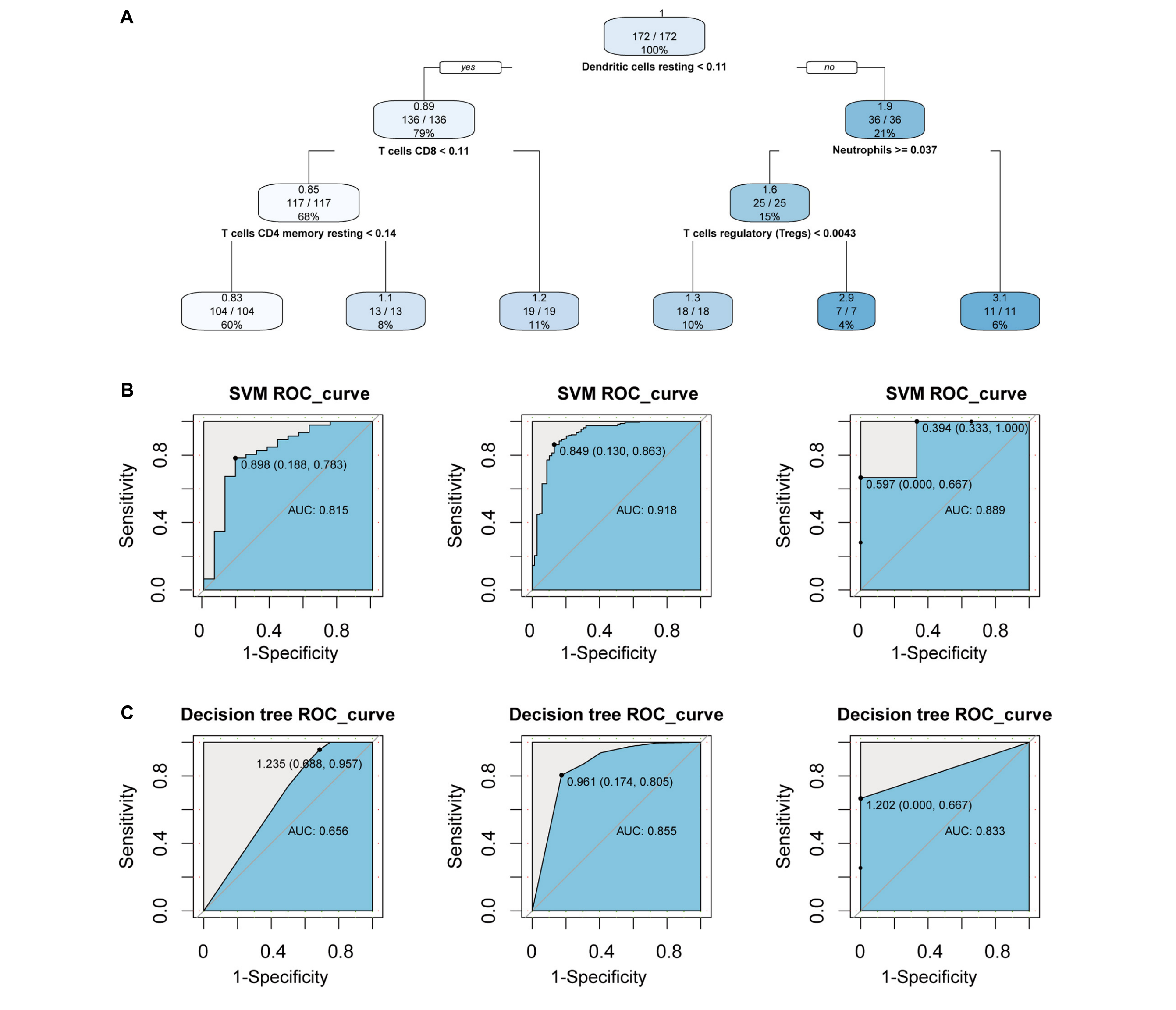


FIGURE 2 | Construct a classifier with significantly different distribution of immune cells. **(A)** This picture is the decision tree diagram of the decision tree classifier. **(B)** The ROC curve represents the area under curve (AUC) of the test set and validation set for SVM classifiers. **(C)** The same as in **(B)** but the Decision tree.

genes in GSE10334 and GSE16134, respectively. We found that the up-regulated and down-regulated genes in GSE10334 and GSE16134 have significant overlap, indicating that the DEGs we obtained from the analysis of experimental data are reliable (**Figure 3C**). Further, there are 1,424 DEGs shared by GSE10334 and GSE16134.

Next, we conduct preliminary statistics on the functional effects of DEGs. These two series of DEGs are integrated into a multi-gene set list, which is used for multi-gene set GO function enrichment and KEGG pathway enrichment, and the functional pathway with $p < 0.05$ is selected as the significant function. We use dotplot and emapplot to display 15 functional nodes and pathways in the results of function and pathway enrichment (**Figure 3D**). Since the DEGs of the two series of samples have a large overlap, they are very similar in function and pathway enrichment. We can see from the enrichment results that periodontitis disease has significant enrichment in cell growth and related immune functions. And which DEGs interact and regulate relationships deserve further analysis.

PPI Network of DEGs

Building a protein interaction network (PPI) is a common method to reveal the interaction relationships and functional modules between genes, so we constructed a PPI network of DEGs (**Figure 4A**). First, merge these two series of DEGS to obtain a total of 1,822 DEGs, and then extract the corresponding interaction relationship pairs to draw the PPI network. In the biological network, the node with the higher degree plays a bigger role in the network and has important functions. Therefore, we extracted the top 30 degree-ranked genes as

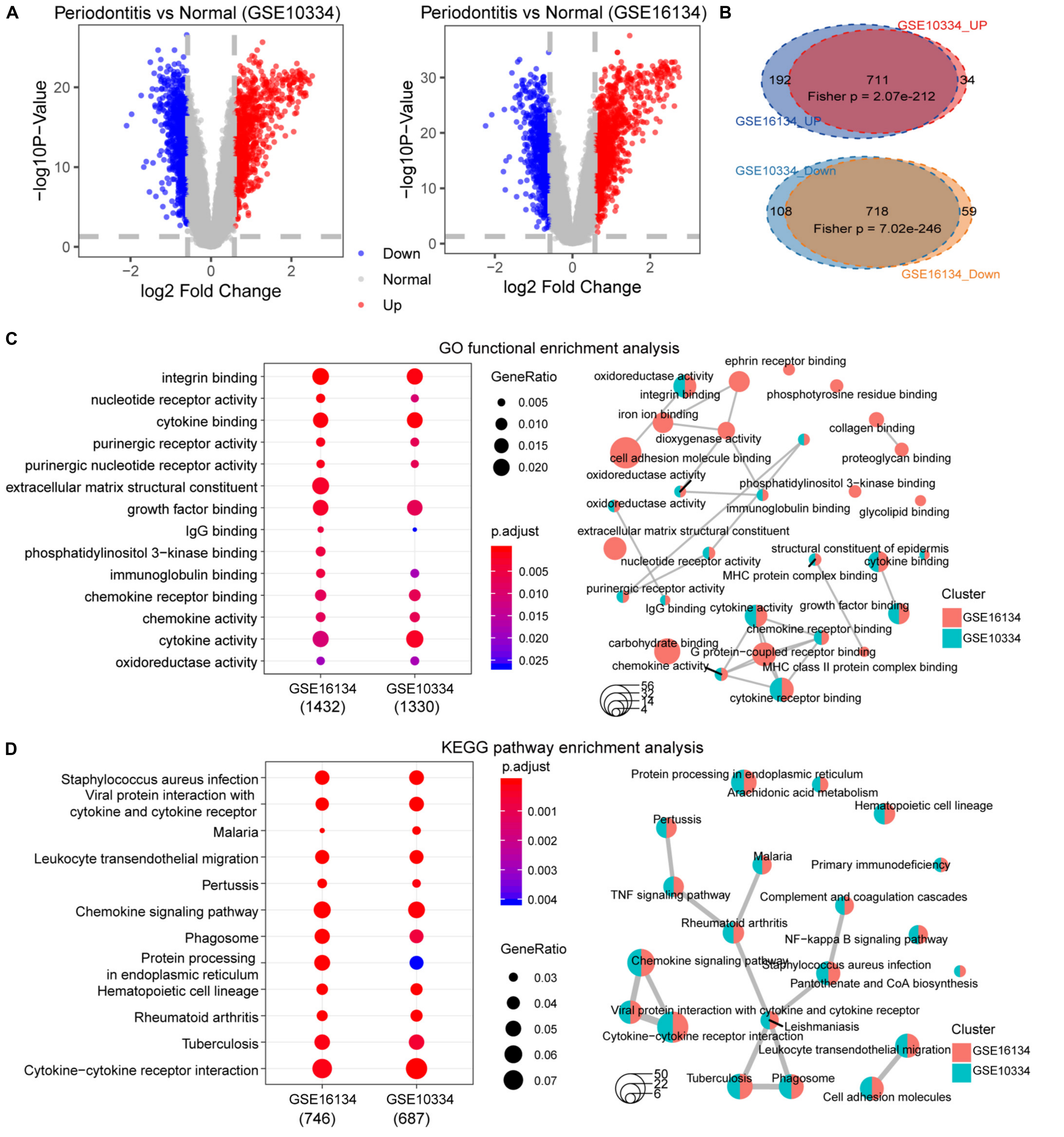


FIGURE 3 | Differential expression analysis and functional enrichment analysis between periodontitis and normal samples. **(A)** This picture represents the volcano map of DEGs for the GSE10334 series. **(B)** This venn diagram describe the intersection of the up- and down-regulated genes in the GSE10334 and GSE16134 series. Fisher's exact test is used to measure the significance level of overlap. **(C–D)** This picture represents the dotplot and emapplot of the GO function enrichment node of DEGs in the GSE10334 and GSE16134 series of samples. e represents the dotplot and emapplot of the DEGs KEGG pathway enrichment in GSE10334 and GSE16134 series samples.

important periodontitis disease-related genes (**Supplementary Table S2**). The results show that genes such as FYN, LYN, LCK, Critical Assessment of Techniques for Protein Structure Prediction experiment (CASP3), arrestin beta 2 (ARRB2) are the central node genes with high connectivity in the PPI network. Among them, FYN, LYN, and LCK are all members of the protein tyrosine kinase (PTK) family, and they are non-receptor PTKs. Studies have shown that most proto-oncogenes have PTK

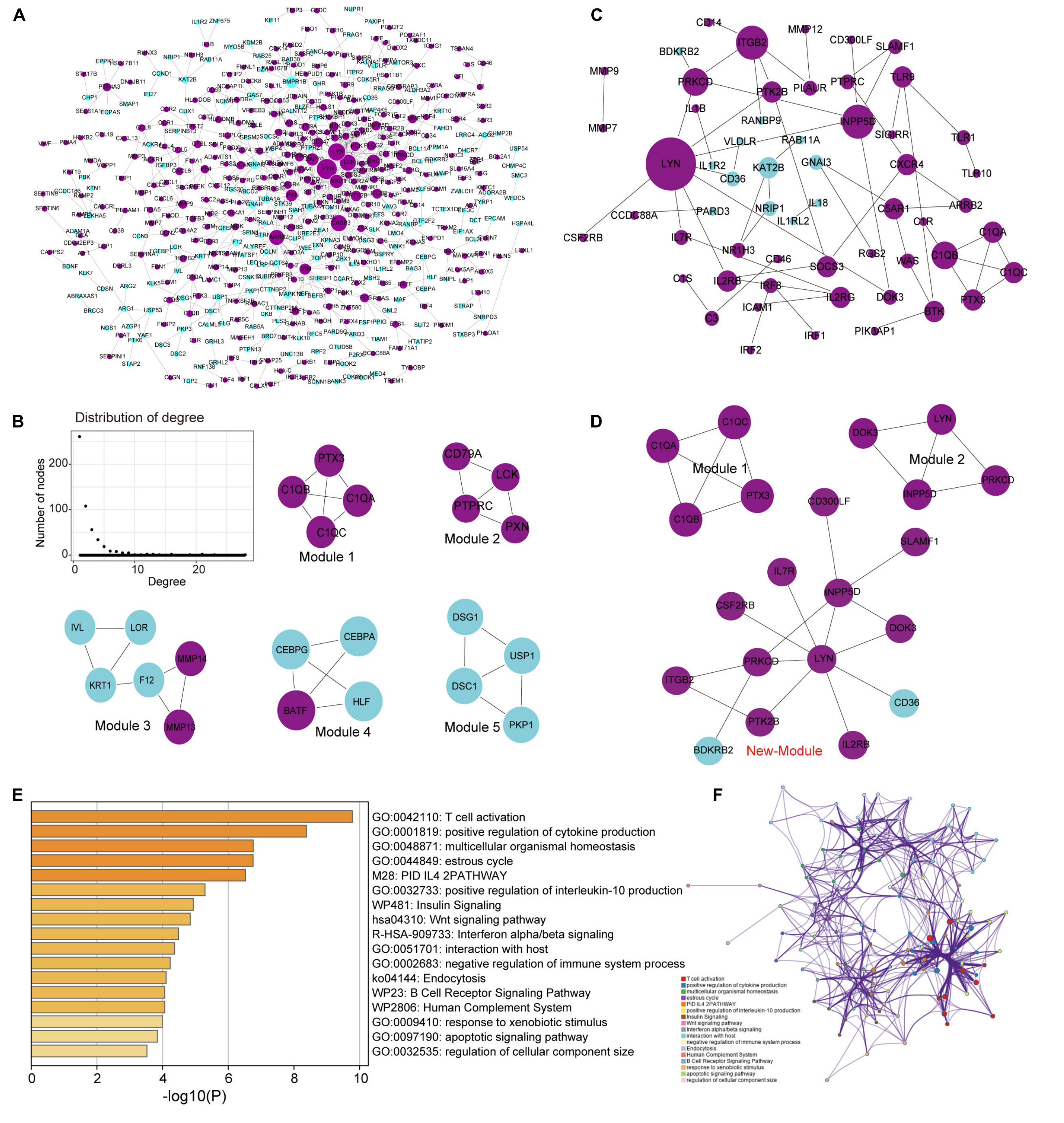

A
B
Distribution of degree
Number of nodes
200
100
0
0
10
20
Degree
PTX3
C1QB
C1QA
C1QC
Module 1
CD79A
LCK
PTPRC
PXN
Module 2
IVL
LOR
KRT1
F12
MMP14
MMP13
Module 3
CEBPG
CEBPA
BATF
HLF
Module 4
DSG1
USP1
DSC1
PKP1
Module 5
C
CD14
ITGB2
MMP12
CD300LF
SLAMF1
BDKRB2
MMP9
PRKCD
PTK2B
PLAUR
PTPRC
TLR9
MMP7
IL1B
RANBP9
INPP5D
LYN
VLDLR
RAB11A
SIGIRR
TLR1
IL1R2
KAT2B
GNAI3
CXCR4
TLR10
CD36
CCDC88A
PARD3
NRIP1
IL18
C5AR1
ARRB2
CSF2RB
IL7R
IL1RL2
C1R
C1QA
NR1H3
RGS2
CD46
SOCS3
WAS
C1QB
C1S
IL2RB
IRF8
C1QC
IL2RG
DOK3
PTX3
ICAM1
BTK
C3
IRF1
PIK3AP1
IRF2
D
C1QC
C1QA
Module 1
PTX3
C1QB
LYN
DOK3
Module 2
INPP5D
PRKCD
CD300LF
SLAMF1
IL7R
INPP5D
CSF2RB
DOK3
LYN
PRKCD
ITGB2
CD36
PTK2B
BDKRB2
New-Module
IL2RB
E
GO:0042110: T cell activation
GO:0001819: positive regulation of cytokine production
GO:0048871: multicellular organismal homeostasis
GO:0044849: estrous cycle
M28: PID IL4 2PATHWAY
GO:0032733: positive regulation of interleukin-10 production
WP481: Insulin Signaling
hsa04310: Wnt signaling pathway
R-HSA-909733: Interferon alpha/beta signaling
GO:0051701: interaction with host
GO:0002683: negative regulation of immune system process
ko04144: Endocytosis
WP23: B Cell Receptor Signaling Pathway
WP2806: Human Complement System
GO:0009410: response to xenobiotic stimulus
GO:0097190: apoptotic signaling pathway
GO:0032535: regulation of cellular component size
0
2
4
6
8
10
-log10(P)
F
T cell activation
positive regulation of cytokine production
multicellular organismal homeostasis
estrous cycle
PID IL4 2PATHWAY
positive regulation of interleukin-10 production
Insulin Signaling
Wnt signaling pathway
Interferon alpha/beta signaling
interaction with host
negative regulation of immune system process
Endocytosis
Human Complement System
B Cell Receptor Signaling Pathway
response to xenobiotic stimulus
apoptotic signaling pathway
regulation of cellular component size

FIGURE 4 | Analysis of the topological properties and functional modules of the PPI network of DEGs and crosstalk genes. **(A)** This picture represents the protein interaction network of two series of integrated DEGs. There are 647 relationship pairs and 515 nodes in the network. **(B)** This picture is a moderate topological analysis of the PPI network of DEGs and the five functional modules in the network. **(C)** This picture shows the PPI network of crosstalk gene, which has 58 relational pairs and 57 nodes. **(D)** This picture is the three modules in the PPI network of crosstalk gene. **(E)** Bar graph of enriched terms across TF and target genes associated with immune pathways, colored by p-values. **(F)** Network of enriched terms colored by cluster ID, where nodes that share the same cluster ID are typically close to each other.

activity, and their abnormal expression will lead to disorders of cell proliferation and eventually tumorigenesis (Drake et al., 2014). Non-receptor PTK-mediated signal transmission plays an important role in the activation of T cells, B cells, NK cells and granulocytes, and the abnormality of its gene structure or gene expression is the cause of certain immunodeficiency diseases and immunoproliferative diseases (Vivier et al., 2004; Vasquez et al., 2019). This means that FYN, LYN and LCK, which are highly expressed, play an important role in the imbalance of the immune system of periodontitis. In addition, we selected five important functional modules from the PPI network (**Figure 4B**), all of which play an important role in cellular immunity (Module 1) and cell growth and proliferation. In order to further study the relationship between immunity and periodontitis disease, we extracted genes related to immunity among DEGs and conducted a series of analysis and research.

Crosstalk Gene in Immune Imbalance

Since crosstalk occurs when TFs regulate multitude of immune-related genes in periodontitis disease, it is intriguing to explore the regulatory mechanisms of immune-related genes (Friedlander et al., 2016; Grah and Friedlander, 2020). We extracted the immune-related genes from the DEGs and defined them as crosstalk genes. Then, we obtained 159 crosstalk genes, which are immune-related genes differentially expressed in periodontitis diseases. We extracted the PPI relationship pairs of these crosstalk genes to draw a PPI interaction network, and analyzed the functional modules and topological properties of the network (**Figure 4C**). We obtained 3 functional modules including a new immune function module (New-module) which was not recognized in the previous PPI network (**Figure 4D**).

As we all know, TFs can control gene expression and expression efficiency (Lambert et al., 2018). Therefore, the analysis of transcription regulation relationship helps us understand the process of several gene expression changes. We collected TF-target relationships from TRRUST and ORTI database which identify TF-target regulations from small-scale experimental studies and interrogating gene expression data. These TF-target relationships were mapped to the transcriptional regulatory network of DEGs associating with crosstalk genes (**Supplementary Figure S2**). There were 19 TFs in this transcriptional regulatory network, of which 14 were up-regulated and 5 were down-regulated. A total of 5 TFs were crosstalk genes that had unbinding event with known target genes, and they were all up-regulated in expression, including early growth response 1 (EGR1), ETS proto-oncogene 1 (ETS1), interferon regulatory factor 4 (IRF4), RUNX family transcription factor 3 (RUNX3), and X-box binding protein 1 (XBP1). We combined immune-related genes on the basis of transcriptional regulatory network to explore the functions of TFs in the immune microenvironment according to Metascape (Zhou et al., 2019). We found that these TFs and their targeted genes are closely related to the activity of T cells (**Figures 4E,F**), which may lead to the formation of periodontitis immunosuppressive microenvironment. By analyzing the topological properties of the network (**Table 2** and **Supplementary Table S3**), we found that EGR1, ETS1, RUNX3, and XBP1 were associating with multiple genes. We also found that most of the up-regulated genes in the New-module functional module of the cross-talk gene PPI network are regulated by ETS1 and EGR1.

Explore the Immune Function of New-Module

As an important and novel functional module, New-module is worthy of our in-depth exploration. We extracted the up-regulated genes in New-module as a gene set, and analyzed their biological pathways (BP) and functional pathways, where ont = 'BP' was set in enrichGO, and $p < 0.05$ was set uniformly. Through enrichment analysis of the up-regulated target genes in module3, we have obtained significantly enriched functional pathways. For the large number of BPs, we used dotplot and cnetplot to show only the top 30 BPs terms (**Figures 5A,B**). These BPs are mainly related to immune cell invasion and activity. In the cnetplot, we found that these biological pathways mainly involve 7 genes, including INPP5D, LYN, PRKCD, PTK2B, ITGB2, SLAMF1, and IL2RB. These genes are only significantly enriched in one pathway, namely the Chemokine Signaling pathway (hsa04062; chemokine signaling pathway), in which three genes including LYN, PRKCD and PTK2B are involved (**Figures 5C,D**). Studies have found that chemokines play a basic role in the transport and activation of monocytes and lymphocytes in the inflammation site. For example, this mechanism can perpetuate local inflammation in the joints of RA patients (Zhang et al., 2015). So, in periodontitis disease, it was possible to believe that the production and persistence of inflammation caused by immunosuppressive microenvironment is achieved through the influence on chemokine signaling pathways.

We then used boxplot to show the relationship between these genes and the expression of TFs, and we found that the expression changes of TF and target genes are consistent, which is in line with the transcription regulation relationship (**Figures 5E,F**). The TFs involved are the two high-outdegree TFs, ETS1 and EGR1, which reveals that the TFs ETS1 and EGR1 play a crucial role

TABLE 2 | Top 10 outdegree genes in the transcriptional regulatory network as key genes.

Symbol	Out degree	Average shortest path Length	Betweenness centrality	Closeness centrality	Regulatory_type	EXP_type
ETS1	859	1.022	0.001	0.979	TF_Target	Down
EGR1	41	1	3.23E-05	1	TF_Target	Down
RUNX3	8	1	1.55E-05	1	TF_Target	Down
XBP1	6	1	1.29E-06	1	TF_Target	Down
CEBPA	5	1	0	1	TF	Down
IRF1	2	1	8.60E-07	1	TF_Target	Up
IRF2	2	1	8.60E-07	1	TF_Target	Up
POU2F2	2	2.018	1.29E-06	0.495	TF_Target	Down
STAT4	2	1.333	0	0.750	TF_Target	Down
IRF4	1	1	3.87E-06	1	TF_Target	Up

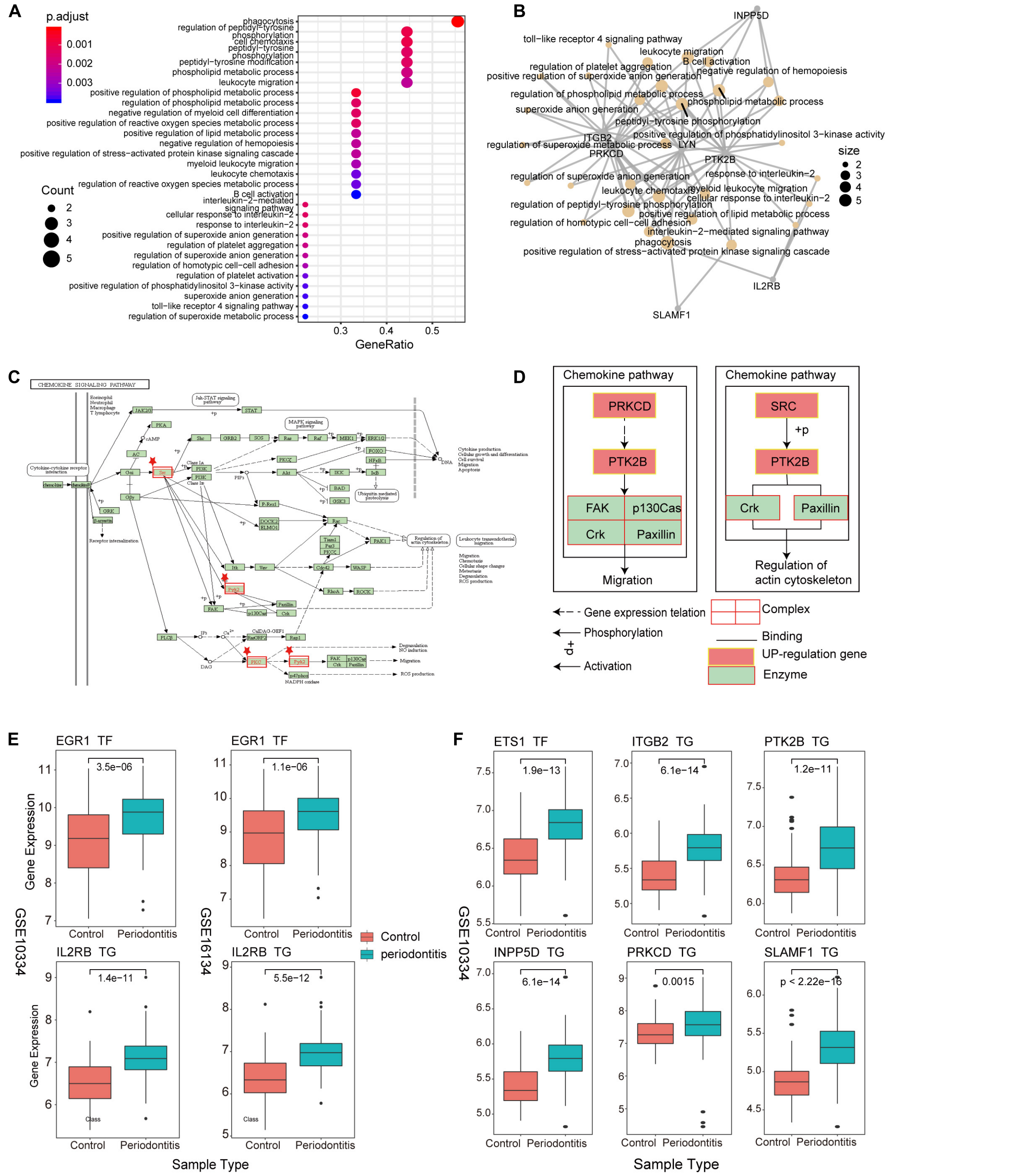


FIGURE 5 | New-module function and pathway analysis **(A)**. The dotplot of enriched biological pathways (BP) across up-regulated genes in the New-module. GeneRatio is the number of enriched genes/number of all genes of a GO term. **(B)** Network of enriched terms, where nodes that share the same genes are typically link to each other. The size of the dot represents the counts of gene. **(C)** The pathway diagram is one of the functional pathways enriched by up-regulated genes in the New-module gene. **(D)** The mechanism of the New-module up-regulated genes on the chemokine signaling pathway. **(E)** The boxplot represents the transcriptional regulatory relationship of the up-regulated genes in the New-module for GSE10334 and GSE16134 series. **(F)** The same as in **(E)** but only for GSE10334 series.

in the invasion and activity of immune cells in periodontitis. Further, we explored whether these TFs played driver roles in TF-target relationships by using Chromatin Immunoprecipitation Sequencing (ChIP-seq) data from ENCODE (v112). Enriched sequencing read peaks of these TFs have been found in the transcription factor binding site (TFBS) regions of downstream target genes. For example, the EGR1-IL2RB relationship of **Figure 5E** has been supported by multiple ChIP-seq datasets

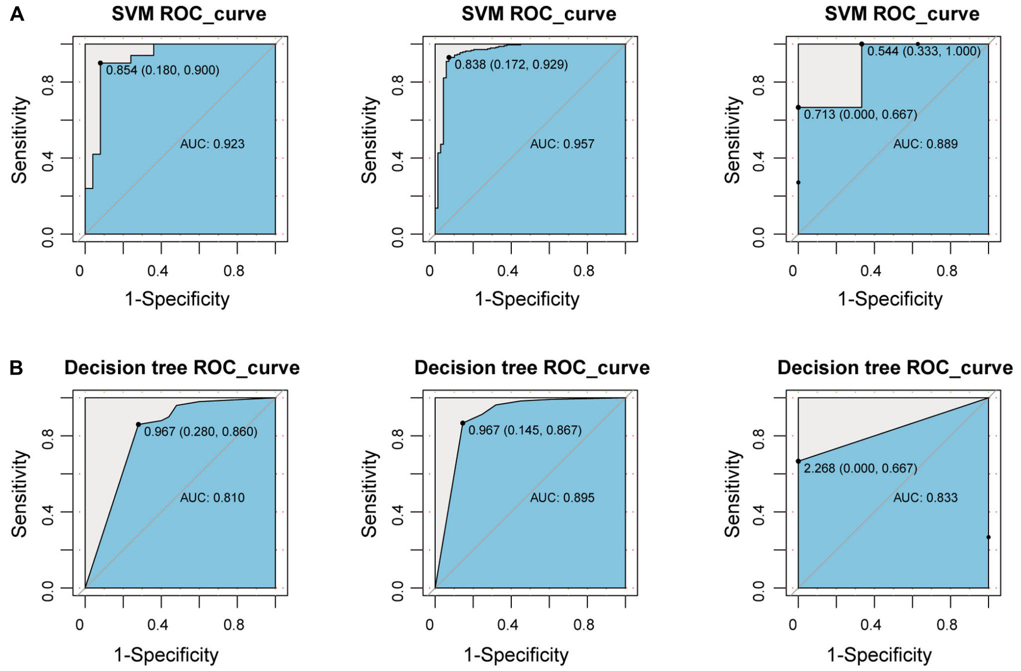


FIGURE 6 | Construct a classifier based on the New-module gene. **(A)** The ROC curves of the test set and validation set for SVM algorithm constructed with the New-module functional module gene in the crosstalk gene PPI network. **(B)** The same as in **(A)** but for the decision tree algorithm.

(**Supplementary Figure S3**). The ETS1-target relationships of **Figure 5F** has also been supported by multiple ChIP-seq datasets (**Supplementary Figures S4–S8**). Since the immune function module New-module plays an important role in periodontitis disease, we decided to rebuild the classifier using the gene of this module as features and compare the performance of the previous classifiers.

Construct a Classifier Based on New-Module

To explore whether the new-module can accurately define periodontitis and normal tissue, we constructed classifiers using the genes in the module as features. Considering that the expression values of the GSE10334, GSE16134, and GSE23586 series are of different magnitudes, we normalized them to make them consistent. We built two classifiers based on decision tree and SVM and used the classifier to predict the test set and the validation set (see section "Materials and Methods"). We found that the classifier constructed with SVM is the best here, and the AUC values of the test set and the two validation sets are 0.923, 0.957, and 0.889, respectively (**Figure 6**). The lower AUC value of GSE23586 as the test set is caused by the small sample size. Generally speaking, the effect of the classifier is better.

Then we compared the performance evaluation results of this classifier with the previous ones (**Table 3**). From the comparison results, we can clearly see that the effect of constructing a classifier based on the new-module functional module is better than based on the different content of immune cells. All these suggesting that although there are differences in the fraction of immune cells between periodontitis samples and normal samples, the differences will be more significant at the level of molecular level.

DISCUSSION

In this study, we systematically analyzed the immune imbalance of periodontitis from the cellular to molecular level. Measuring the fraction of immune cells between periodontitis and normal tissues was used to determine the feature and role of immune cells in periodontitis. Statistical analysis of gene expression profiles is used to reveal abnormally expressed genes in periodontitis. The PPI was constructed to explore potential functional modules and reveal new molecular mechanisms of immune imbalance in periodontitis. We have reconstructed the PPI network base on immune genes and discovered a new immune function module named New-module. By integrating TF-target relationships and

TABLE 3 | Comparison table of performance evaluation of two classifiers successively.

Classification features	Series number	data sets	SVM AUC	Decision tree AUC
Immune cells	GSE10334	Test set	0.815	0.656
	GSE16134	Validation set	0.918	0.855
	GSE23586	Validation set	0.889	0.833
Important crosstalk genes	GSE10334	Test set	0.923	0.810
	GSE16134	Validation set	0.957	0.895
	GSE23586	Validation set	0.889	0.833

ChIP-seq data, we found that EGR1, ETS1, RUNX3, and XBP1 were key TFs that regulate the expression of genes that participate in the formation of the immunosuppressive microenvironment. The up-regulated genes are mainly regulated by EGR1 and ETS1 in New-module. In addition, New-module not only plays an important role in the imbalance of the immune system, but is also closely related to the occurrence and persistence of periodontal tissue inflammation.

Periodontitis is mainly a chronic inflammation of periodontal tissue caused by pathogens, which has the characteristics of complicated pathogenesis and long duration. Previous studies have shown that the imbalance of the immune system caused by pathogen colonization is an important factor in the occurrence and development of periodontitis. The majority of work has focused on the external pathogenic factors and clinical treatment of periodontitis, with limited documentation of indications that the changes in the molecular mechanism of the immune system of patients with periodontitis. In addition, more and more studies have demonstrated the significance of the imbalance of the immune system for periodontitis, including the abnormality of cytokines in the host immune response (Pan et al., 2019), and the immune imbalance of periodontitis leads to systemic inflammation (Hajishengallis, 2015). Exploring the disease tissue microenvironment at single-cell resolution is a popular direction, but the lack of high-throughput data for periodontitis has forced us to consider other approaches. In order to be able to further explore the tissue microenvironment and epigenetic characteristics of periodontitis in future research, TOAST (Li et al., 2019, 2020; Li and Wu, 2019) tool that offers functions for detecting cell-type specific differential expression (csDE) and differential methylation (csDM) brings convenience to our research. In the current study, we comprehensively assessed the immune system imbalance of periodontitis from the cellular to molecular level, which gained a new insight in protein interaction and transcriptional regulation.

During the construction of PPI networks, usage of immune genes only will lose many other pathway signals. Our purpose is to explore the molecular mechanism of the immune microenvironment reprogramming of periodontitis disease. Although our selection of immune genes will ignore other signaling pathways, the formation mechanism of the immunosuppressive microenvironment of periodontitis disease is important. In the future, we will integrate more genes into the PPI networks and perform functional analysis to characterize periodontitis disease comprehensively.

We successfully determined the immunosuppressive microenvironment of periodontitis in the measurement of immune cell distribution. Notably, we measured the distribution of immune cells and differential gene expression in two series with rich samples, which can effectively avoid the false negative problem faced in the research. Our data may discover previously overlooked pathogenic genes and molecular mechanisms, adding a new blueprint for periodontitis research. In addition, we also used a machine learning algorithm to build a classifier model to consider the reliability and pros and cons of the statistically obtained disease characteristics. Periodontitis is mainly a local inflammation caused by pathogen-induced immune invasion. Therefore, investigation and interpretation of the immune system would provide novel and useful insights into the mechanisms underlying the functions of these molecules in periodontitis. In our further work, we will perform experiments *in vitro* to validate key regulators identified from our results. The experimental strategy will measure the expression levels of risk genes using qRT-PCR in normal and disease tissues. Further, siRNAs will be used to knockdown their expression and study gene functions with cell proliferation assay, wound healing assay.

CONCLUSION

In summary, we provide a comprehensive view of the imbalance mechanism of the periodontitis immune system from the cellular to the molecular level. Our findings expand existing knowledge about immunosuppressive associated with periodontitis. The integration of multi-platform data comprehensively reveal that the immune system imbalance mechanism of periodontitis patients enhances the interpretability of the pathogenesis of periodontitis, which may help the development of new periodontitis treatments.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LB and LH conceived and designed the experiments. LL, TL, and DZ analyzed the data. JD and BW collected the data. LH and LL validated the method and data. LH wrote this 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.2021.653209/full#supplementary-material

REFERENCES

Belibasakis, G. N., and Bostanci, N. (2012). The RANKL-OPG system in clinical periodontology. *J. Clin. Periodontol.* 39, 239–248. doi: 10.1111/j.1600-051X.2011.01810.x

Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., et al. (2013). InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res.* 41, D1228–D1233. doi: 10.1093/nar/gks1147

Brown, K. R., and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics* 21, 2076–2082. doi: 10.1093/bioinformatics/bti273

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., et al. (2007). MINT: the molecular INTeraction database. *Nucleic Acids Res.* 35, D572–D574. doi: 10.1093/nar/gkl950

Cyktor, J. C., and Turner, J. (2011). Interleukin-10 and immunity against prokaryotic and eukaryotic intracellular pathogens. *Infect. Immun.* 79, 2964–2973. doi: 10.1128/IAI.00047-11

Demmer, R. T., Behle, J. H., Wolf, D. L., Handfield, M., Kebschull, M., Celenti, R., et al. (2008). Transcriptomes in healthy and diseased gingival tissues. *J. Periodontol.* 79, 2112–2124. doi: 10.1902/jop.2008.080139

Drake, J. M., Lee, J. K., and Witte, O. N. (2014). Clinical targeting of mutated and wild-type protein tyrosine kinases in cancer. *Mol. Cell. Biol.* 34, 1722–1732. doi: 10.1128/MCB.01592-13

Eke, P. I., Dye, B. A., Wei, L., Slade, G. D., Thornton-Evans, G. O., Borgnakke, W. S., et al. (2015). Update on prevalence of periodontitis in adults in the United States: NHANES 2009 to 2012. *J. Periodontol.* 86, 611–622. doi: 10.1902/jop.2015.140520

Friedlander, T., Prizak, R., Guet, C. C., Barton, N. H., and Tkacik, G. (2016). Intrinsic limits to gene regulation by global crosstalk. *Nat. Commun.* 7:12307. doi: 10.1038/ncomms12307

Genco, R. J., and Van Dyke, T. E. (2010). Prevention: reducing the risk of CVD in patients with periodontitis. *Nat. Rev. Cardiol.* 7, 479–480. doi: 10.1038/nrcardio.2010.120

Gilbert, D. (2005). Biomolecular interaction network database. *Brief. Bioinform.* 6, 194–198. doi: 10.1093/bib/6.2.194

Goel, R., Harsha, H. C., Pandey, A., and Prasad, T. S. (2012). Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Mol. Biosyst.* 8, 453–463. doi: 10.1039/c1mb05340j

Grah, R., and Friedlander, T. (2020). The relation between crosstalk and gene regulation form revisited. *PLoS Comput. Biol.* 16:e1007642. doi: 10.1371/journal.pcbi.1007642

Graves, D. (2008). Cytokines that promote periodontal tissue destruction. *J. Periodontol.* 79, 1585–1591. doi: 10.1902/jop.2008.080183

Graves, D. T., Correa, J. D., and Silva, T. A. (2019). The oral microbiota is modified by systemic diseases. *J. Dent. Res.* 98, 148–156. doi: 10.1177/0022034518805739

Hajishengallis, G. (2014a). Immunomicrobial pathogenesis of periodontitis: keystones, pathobionts, and host response. *Trends Immunol.* 35, 3–11. doi: 10.1016/j.it.2013.09.001

Hajishengallis, G. (2014b). The inflammophilic character of the periodontitis-associated microbiota. *Mol. Oral Microbiol.* 29, 248–257. doi: 10.1111/omi.12065

Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation. *Nat. Rev. Immunol.* 15, 30–44. doi: 10.1038/nri3785

Hajishengallis, G., Darveau, R. P., and Curtis, M. A. (2012). The keystone-pathogen hypothesis. *Nat. Rev. Microbiol.* 10, 717–725. doi: 10.1038/nrmicro2873

Hajishengallis, G., and Korostoff, J. M. (2017). Revisiting the Page & Schroeder model: the good, the bad and the unknowns in the periodontal host response 40 years later. *Periodontol. 2000* 75, 116–151. doi: 10.1111/prd.12181

Hajishengallis, G., Liang, S., Payne, M. A., Hashim, A., Jotwani, R., Eskan, M. A., et al. (2011). Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement. *Cell Host Microbe* 10, 497–506. doi: 10.1016/j.chom.2011.10.006

Kebschull, M., Demmer, R. T., Grun, B., Guarnieri, P., Pavlidis, P., and Papapanou, P. N. (2014). Gingival tissue transcriptomes identify distinct periodontitis phenotypes. *J. Dent. Res.* 93, 459–468. doi: 10.1177/0022034514527288

Kebschull, M., Demmer, R. T., and Papapanou, P. N. (2010). "Gum bug, leave my heart alone!"–epidemiologic and mechanistic evidence linking periodontal infections and atherosclerosis. *J. Dent. Res.* 89, 879–902. doi: 10.1177/0022034510375281

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40, D841–D846. doi: 10.1093/nar/gkr1088

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172, 650–665. doi: 10.1016/j.cell.2018.01.029

Li, Z., Guo, Z., Cheng, Y., Jin, P., and Wu, H. (2020). Robust partial reference-free cell composition estimation from tissue expression. *Bioinformatics* 36, 3431–3438. doi: 10.1093/bioinformatics/btaa184

Li, Z., and Wu, H. (2019). TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol.* 20:190. doi: 10.1186/s13059-019-1778-0

Li, Z., Wu, Z., Jin, P., and Wu, H. (2019). Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics* 35, 3898–3905. doi: 10.1093/bioinformatics/btz196

Liu, Y., He, M., Wang, D., Diao, L., Liu, J., Tang, L., et al. (2017). HisgAtlas 1.0: a human immunosuppression gene database. *Database (Oxford)* 2017:bax094. doi: 10.1093/database/bax094

Lundberg, K., Wegner, N., Yucel-Lindberg, T., and Venables, P. J. (2010). Periodontitis in RA-the citrullinated enolase connection. *Nat. Rev. Rheumatol.* 6, 727–730. doi: 10.1038/nrrheum.2010.139

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

Nibali, L., Bayliss-Chapman, J., Almofareh, S. A., Zhou, Y., Divaris, K., and Vieira, A. R. (2019). What is the heritability of periodontitis? A systematic review. *J. Dent. Res.* 98, 632–641. doi: 10.1177/0022034519842510

Oughtred, R., Stark, C., Breitkreutz, B. J., Rust, J., Boucher, L., Chang, C., et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47, D529–D541. doi: 10.1093/nar/gky1079

Pan, W., Wang, Q., and Chen, Q. (2019). The cytokine network involved in the host immune response to periodontitis. *Int. J. Oral Sci.* 11:30. doi: 10.1038/s41368-019-0064-z

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

Vafaee, F., Krycer, J. R., Ma, X., Burykin, T., James, D. E., and Kuncic, Z. (2016). ORTI: an open-access repository of transcriptional interactions for interrogating mammalian gene expression data. *PLoS One* 11:e0164535. doi: 10.1371/journal.pone.0164535

Vasquez, A., Baena, A., Gonzalez, L. A., Restrepo, M., Munoz, C. H., Vanegas-Garcia, A., et al. (2019). Altered recruitment of Lyn, Syk and ZAP-70 into lipid rafts of activated B cells in systemic lupus erythematosus. *Cell. Signal.* 58, 9–19. doi: 10.1016/j.cellsig.2019.03.003

Vivier, E., Nunes, J. A., and Vely, F. (2004). Natural killer cell signaling pathways. *Science* 306, 1517–1519. doi: 10.1126/science.1103478

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Yang, F., Qin, Y., Wang, Y., Li, A., Lv, J., Sun, X., et al. (2018). LncRNA KCNQ1OT1 mediates pyroptosis in diabetic cardiomyopathy. *Cell. Physiol. Biochem.* 50, 1230–1244. doi: 10.1159/000494576

Zhang, L., Yu, M., Deng, J., Lv, X., Liu, J., Xiao, Y., et al. (2015). Chemokine signaling pathway involved in CCL2 expression in patients with rheumatoid arthritis. *Yonsei Med. J.* 56, 1134–1142. doi: 10.3349/ymj.2015.56.4.1134

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523. doi: 10.1038/s41467-019-09234-6

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

frontiers
in Genetics

ORIGINAL RESEARCH
published: 12 May 2021
doi: 10.3389/fgene.2021.670787

Systemic Bioinformatic Analyses of Nuclear-Encoded Mitochondrial Genes in Hypertrophic Cardiomyopathy

Zhaochong Tan[1]†, Limeng Wu[1]†, Yan Fang[1], Pingshan Chen[2], Rong Wan[3], Yang Shen[3], Jianping Hu[3], Zhenhong Jiang[3] and Kui Hong[1,3]**

[1] *Department of Cardiovascular Medicine, The Second Affiliated Hospital of Nanchang University, Nanchang, China,*
[2] *Department of Science and Technology, The Second Affiliated Hospital of Nanchang University, Nanchang, China,*
[3] *Jiangxi Key Laboratory of Molecular Medicine, The Second Affiliated Hospital of Nanchang University, Nanchang, China*

OPEN ACCESS

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Renxiang Yan,
Fuzhou University, China
Zhou Zhou,
Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, China

***Correspondence:**
Zhenhong Jiang
jxzhjiang@163.com
Kui Hong
hongkui88@163.com

†These authors have contributed equally to this work

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

Received: *22 February 2021*
Accepted: *08 April 2021*
Published: *12 May 2021*

Citation:
Tan Z, Wu L, Fang Y, Chen P, Wan R, Shen Y, Hu J, Jiang Z and Hong K (2021) Systemic Bioinformatic Analyses of Nuclear-Encoded Mitochondrial Genes in Hypertrophic Cardiomyopathy.
Front. Genet. 12:670787.
doi: 10.3389/fgene.2021.670787

Hypertrophic cardiomyopathy (HCM) is an autosomal dominant disease and mitochondria plays a key role in the progression in HCM. Here, we analyzed the expression pattern of nuclear-encoded mitochondrial genes (NMGenes) in HCM and found that the expression of NMGenes was significantly changed. A total of 316 differentially expressed NMGenes (DE-NMGenes) were identified. Pathway enrichment analyses showed that energy metabolism-related pathways such as "pyruvate metabolism" and "fatty acid degradation" were dysregulated, which highlighted the importance of energy metabolism in HCM. Next, we constructed a protein-protein interaction network based on 316 DE-NMGenes and identified thirteen hubs. Then, a total of 17 TFs (transcription factors) were predicted to potentially regulate the expression of 316 DE-NMGenes according to iRegulon, among which 8 TFs were already found involved in pathological hypertrophy. The remaining TFs (like GATA1, GATA5, and NFYA) were good candidates for further experimental verification. Finally, a mouse model of transverse aortic constriction (TAC) was established to validate the genes and results showed that DDIT4, TKT, CLIC1, DDOST, and SNCA were all upregulated in TAC mice. The present study represents the first effort to evaluate the global expression pattern of NMGenes in HCM and provides innovative insight into the molecular mechanism of HCM.

Keywords: hypertrophic cardiomyopathy, microarrays, bioinformatics analysis, nuclear-encoded mitochondrial genes, transcription factors

INTRODUCTION

Hypertrophic cardiomyopathy (HCM) is an autosomal dominant genetic disease that is mainly characterized by ventricular hypertrophy with asymptomatic or serious complications such as sudden cardiac death (SCD), heart failure, and thrombosis (Marian and Braunwald, 2017). The prevalence of HCM in the general population was estimated to be 1/500 (Gersh et al., 2011), which was underestimated due to the limited HCM diagnostic technology. HCM is considered a leading cause of SCD in younger people and the leading cause of heart failure in cardiac diseases originating primarily from the myocardium (Weissler-Snir et al., 2019).

Normal myocardial energy metabolism from mitochondria is also an important material basis for keeping the normal heart tissue structure and the internal environment stable. Cardiac function will inevitably be impaired by mitochondrial dysfunction. Clinical and experimental studies have shown that the myocardial energy source switching from fatty acid oxidation to glycolysis is a common event in HCM (Tian, 2003). Mutations in a wide spectrum of nuclear-encoded mitochondrial genes (NMGenes) have been reported to be able to cause HCM characterized by impaired mitochondrial function (Marin-Garcia and Goldenthal, 2002b). For example, mutations in *ELAC2* (*ElaC* ribonuclease *Z 2*) encoding a short form of RNase Z were found to be associated with HCM (Saoura et al., 2019). Mitochondrial function depends on proteins encoded by both mitochondrial DNA (mtDNA) and nuclear DNA (nDNA). The mitochondrial proteome has been estimated to contain approximately 1000–1500 proteins, more than 99% of which are encoded by nuclear DNA (nDNA), while mtDNA refers to only 13 protein-coding genes (Pfanner et al., 2019). Considering the importance of mitochondria in HCM and the fact that functional proteins in mitochondria are encoded mainly by nDNA genes, exploring the function of NMGenes in HCM would help us better understand the novel role of mitochondria in the development of HCM.

With the development of genetic studies, high-throughput omics technologies (such as DNA microarrays and next-generation sequencing) that investigate gene function and expression at the genome-wide level have been widely used in basic research, clinical diagnosis, drug research and other fields. As a powerful technique, gene expression microarray-based bioinformatics analyses have also been widely used to identify HCM-related genes or noncoding RNAs, possible molecular mechanisms, and biological pathways (Lim et al., 2001; Yang et al., 2015; Hu et al., 2019; Li et al., 2019; Liu et al., 2019). For example, microarray analysis was performed to explore the expression pattern of lncRNAs (long noncoding RNAs) and mRNAs (messenger RNAs) in HCM, which identified hundreds of differentially expressed lncRNAs and genes (Yang et al., 2015). A recent study systemically analyzed RNA-seq data from 28 HCM patients and 9 healthy controls and identified 43 potential pathogenic variants in 19 genes and four subnetworks with significant roles in the progression of HCM (Gao et al., 2020). Although previous studies have highlighted the importance of integrative gene expression analysis in exploring the molecular mechanism of HCM, a systemic analysis of the expression pattern of NMGenes in HCM patients has never been reported.

To investigate the potential role of NMGenes in the pathogenesis of HCM, in this study, we performed a computational systems biology analysis based on large-scale HCM-related transcriptional data. A total of 316 differentially expressed NMGenes (DE-NMGenes) were identified. Based on these DE-NMGenes, gene ontology (GO) and pathway enrichment analyses were performed, and 17 KEGG-dysregulated pathways were identified. We also constructed a PPI (protein-protein interaction) network that consisted of 215 DE-NMGenes and 440 interactions. Finally, a total of 17 TFs (transcription factors) were predicted to potentially regulate the expression of the 316 DE-NMGenes. We provided a systematic view of the roles of mitochondrial genes in HCM and revealed some available candidates for future experimental verification.

RESULTS

Nuclear-Encoded Mitochondrial Genes Are Significantly Changed in HCM

The normalized gene expression dataset GSE36961 was downloaded from the GEO (Gene Expression Omnibus) database[1], which included 107 HCM samples and 40 control samples (Clough and Barrett, 2016). Differentially expressed genes (DEGs) between HCM and the corresponding control samples were detected using the "Limma" package from R software (Ritchie et al., 2015). By keeping genes with a BH (Benjamini-Hochberg)-corrected *p*-value less than 0.01 and fold change (FC) larger than 1.2, we obtained 2927 DEGs, 1499 of which were upregulated and the remaining 1428 were downregulated (**Figure 1A** and **Supplementary Table 1**). To explore the expression pattern of NMGenes in HCM, we collected 1943 mitochondrial genes from the MitoCarta (Calvo et al., 2016), MitoMiner (Smith and Robinson, 2019), IMPI and UniProt databases (UniProt, 2019) (**Figure 1B**, see section "Materials and Methods" for details). After removing 13 mtDNA-encoded genes, 1930 NMGenes were retained for further analysis. Among these genes, 1562 genes were detected on microarray, and 316 genes were differentially expressed. Compared with the overall genes detected on the microarray, the proportion of DEGs in NMGenes was significantly higher (**Figure 1C**, Fisher's exact test, p-value $< 2.20 \times 10^{-16}$). The extensive expression changes of NMGenes in HCM indicate that mitochondria play critical roles in the progression of HCM. **Table 1** lists the top ten upregulated and downregulated NMGenes in HCM. Among these genes, four upregulated genes [namely, *PDK4* (*thpyruvate dehydrogenase kinase isozyme 4*), *STAT3* (*Signal Transducer and Activator of Transcription 3*), *HCLS1* (*Hematopoietic Cell-Specific Lyn Substrate 1*) and *FKBP11* (*FKBP Prolyl Isomerase 11*)], and four downregulated genes [namely, *GATM* (*Glycine Amidinotransferase*), *ATPIF1* (*ATP Synthase Inhibitory Factor Subunit 1*), *CPT1B* (*Carnitine Palmitoyltransferase 1B*), and *GJA1* (*Gap Junction Protein Alpha 1*)] have already been proven to play important roles in pathological hypertrophy (summarized in **Table 1**).

Downregulated DE-NMGenes Are More Functionally Diverse Than Upregulated DE-NMGenes

GO biological process (BP) and KEGG pathway enrichment analyses for 316 DE-NMGenes were performed using DAVID (Database for Annotation, Visualization, and Integrated Discovery) (da Huang et al., 2009). Although the numbers of

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

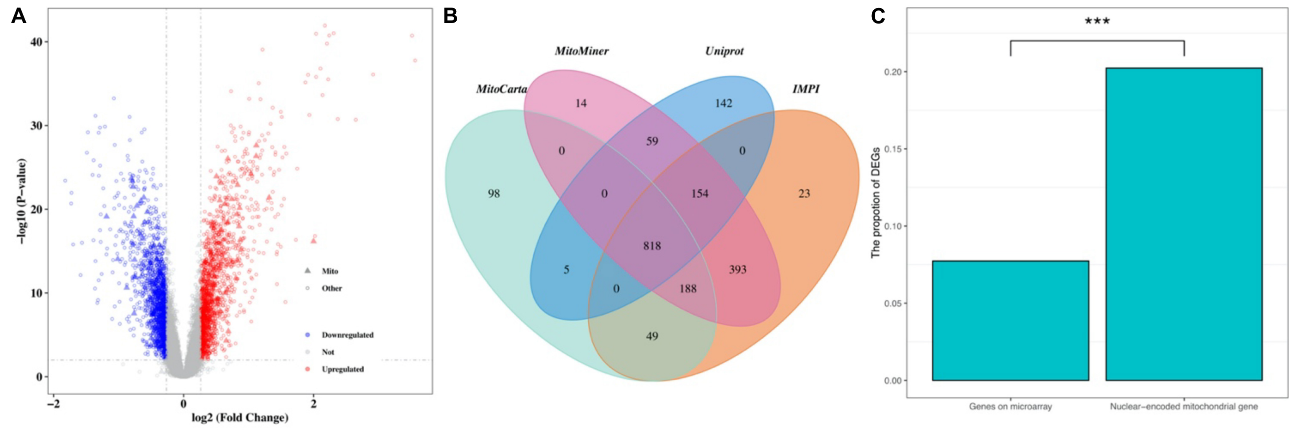


FIGURE 1 | The expression pattern of NMGenes in HCM. **(A)** A volcano plot shows DEGs in HCM detected from GSE36961. The vertical lines correspond to 1.2-fold up and down, respectively, and the horizontal line represents a BH-corrected *p* value of 0.05. Therefore, red and blue dots represent upregulated and downregulated DEGs, respectively. Triangular nodes and circle nodes represent NMGenes and the other genes, respectively. **(B)** A four-way Venn diagram shows the number of NMGenes collected from the MitoCarta, MitoMiner, IMPI, and UniProt databases. **(C)** A bar plot shows the proportion of DEGs in all genes on microarray and NMGenes. The percentage of DEGs in NMGenes was significantly larger than the percentage of DEGs in overall genes detected on microarray (Fisher's exact test, *p*-value $< 2.20 \times 10^{-16}$). *** represents *P*-value < 0.001.

upregulated and downregulated DE-NMGenes were similar, downregulated DE-NMGenes were more functionally diverse than upregulated DE-NMGenes. By keeping terms with BH-corrected *p*-values less than 0.05, we obtained 4 GO BP terms and 4 KEGG pathways for 141 upregulated DE-NMGenes and 16 GO BP terms and 17 KEGG pathways for 175 downregulated DE-NMGenes (**Figures 2A,B** and **Supplementary Table 2**). The top 3 enriched GO terms in downregulated DE-NMGenes were "oxidation-reduction process," "branched-chain amino acid catabolic process" and "fatty acid beta-oxidation." The GO terms "oxidation-reduction process" and "translation" were both enriched in 141 upregulated and 175 downregulated DE-NMGenes. KEGG pathway enrichment analysis showed that downregulated DE-NMGenes were significantly enriched in energy metabolism-related pathways such as "Carbon metabolism" (15 genes, BH-corrected *p*-value = $1.94{*}10^{-9}$), "Pyruvate metabolism" (7 genes, BH-corrected *p*-value = $1.25{*}10^{-4}$), "Fatty acid metabolism" (7 genes, BH-corrected *p*-value = $3.24{*}10^{-4}$), and "Citrate cycle" (5 genes, BH-corrected *p*-value = $4.35{*}10^{-3}$). For upregulated DE-NMGenes, the KEGG pathways "Biosynthesis of antibiotics" (15 genes, BH-corrected *p*-value = $1.75{*}10^{-5}$) and "Biosynthesis of amino acids" (7 genes, BH-corrected *p*-value = $8.88{*}10^{-3}$) were significantly enriched.

A Group of 215 DE-NMGenes Are Biologically Connected to Form a Network

The 316 DE-NMGenes were analyzed together to construct a PPI network. Consequently, a PPI network including 440 interactions and 215 nodes was obtained by using STRING (Search Tool for the Retrieval of Interacting Genes/proteins database) (Szklarczyk et al., 2019), with parameters including a minimum required interaction score larger than 0.7 (high confidence) and only query proteins being displayed. Thus, 215 out of the 316 DE-NMGenes were included in the final PPI network (**Figure 3A**). The 316 DE-NMGenes had significantly more interactions than would be expected (*p*-value $< 2.2{*}10^{-16}$) from a randomly chosen set of proteins of the same size drawn from the genome. In a PPI network, highly connected nodes are called hubs, which are expected to play an important role in understanding the biological mechanism of disease (Barabasi and Oltvai, 2004). Then, we calculated the degree for each node and selected genes with the degree ranked in the top 5% as hubs. Of the 215 nodes in the PPI network, 13 nodes were ranked in the top 5% and selected as hubs (**Table 2**). *DLD* (*dihydrolipoamide dehydrogenase*) was the hub gene with the largest degree and interacted with 19 proteins in the PPI network.

The MCODE (Molecular Complex Detection) plugin in Cytoscape was used to detect network modules from the PPI network (Bader and Hogue, 2003). A module is a group of closely related proteins that act in concert to perform specific biological functions through a PPI network that occurs in time and space (Lin et al., 2015). A total of 12 modules were extracted from the PPI network, of which five modules (modules 1–5) had nodes ≥ 5 (**Figure 3B** and **Supplementary Table 3**). The associated BPs for module 1, module 2, module 4 and module 5 were "mitochondrial respiratory chain complex I assembly," "mitochondrial translation," "protein N-linked glycosylation via asparagine" and "folic acid-containing compound biosynthetic process," respectively, but no-GO term was significantly enriched in module 3.

TFs Potentially Regulating DE-NMGenes Play Key Roles in HCM

iRegulon (Janky et al., 2014), available as a Cytoscape plugin, was used to predict TFs potentially regulating the 316 DE-NMGenes. A total of 17 TFs were obtained with the

TABLE 1 | Top 10 upregulated and top 10 downregulated DE-NMGenes.

Gene symbol	logFC	*P*-value	Roles in pathological hypertrophy
PDK4	2	6.79E-17	ANG II induced cardiac hypertrophy was associated with a marked upregulation of *PDK4* (Mori et al., 2012)
DDIT4 (DNA-damage-inducible transcript 4)	1.32	3.94E-22	–
STAT3	1.12	2.38E-28	Pharmacologic inhibition of *STAT3* with WP1066 could suppress Ang II-induced myocyte hypertrophy (Chen et al., 2017).
HCLS1	1.12	4.52E-27	Changes in phospholipid metabolism occur in mammalian hypertrophied myocardium (Reibel et al., 1986).
TKT (Transketolase)	1.04	6.31E-25	–
CLIC1 (Chloride Intracellular Channel 1)	0.88	6.85E-16	–
ACTB (Actin Beta)	0.87	4.59E-17	–
DDOST (Dolichyl-Diphosphooligosaccharide—Diphosphooligosaccharide-Protein Glycosyltransferase Non-Catalytic Subunit)	0.86	7.14E-21	–
FKBP11	0.85	1.11E-18	*FKBP11* was strongly and acutely induced in cardiac hypertrophy induced by TAC (Wang et al., 2019).
TUBB (Tubulin Beta Class I)	0.83	1.26E-23	–
GATM	−1.19	7.50E-20	In *GATM*-deficient mouses, hypertrophic marker *NPPA* expression was significantly upregulated (Jensen et al., 2020).
SNCA (Synuclein Alpha)	−1.01	8.68E-15	–
CASQ1 (Calsequestrin 1)	−0.87	2.36E-11	–
LYPLAL1 (Lysophospholipase Like 1)	−0.8	2.01E-24	–
ATPIF1	−0.79	4.15E-24	The knockout of *ATPIF1* protected the heart from myocardial hypertrophy induced by transverse aortic constriction or isoproterenol infusion (Yang et al., 2017).
SDSL (Serine Dehydratase Like)	−0.78	2.14E-23	–
KLHDC9 (Kelch Domain Containing 9)	−0.77	7.07E-20	–
DPYSL4 (Dihydropyrimidinase Like 4)	−0.76	2.80E-08	–
CPT1B	−0.75	1.07E-14	*CPT1B* deficiency could cause lipotoxicity in the heart under pathological stress, leading to exacerbated cardiac pathology (He et al., 2012).
GJA1	−0.75	1.28E-12	In HCM patients with valvular aortic stenosis, compensated hypertrophy had increased levels and increased lateral *CJA1* expression (Fontes et al., 2012).

Genes with verified roles in pathological hypertrophy are marked in bold.

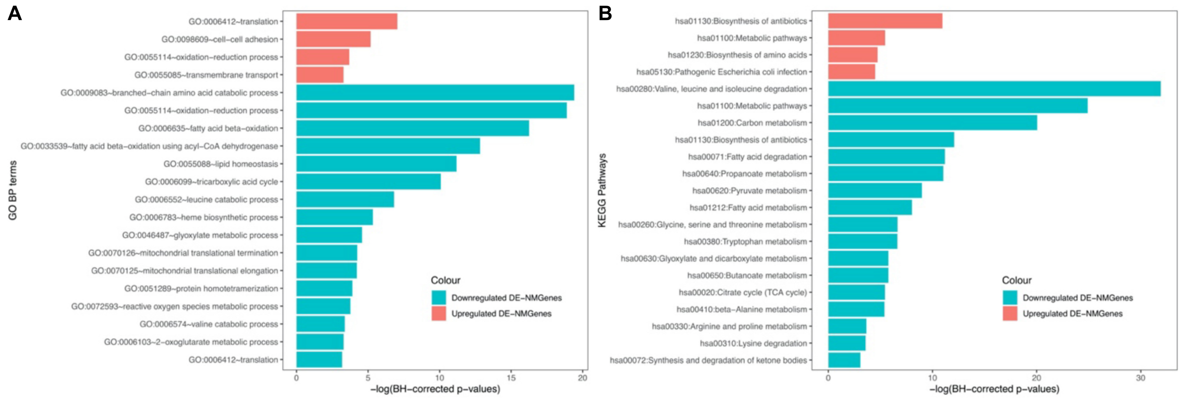


FIGURE 2 | GO and KEGG enrichment results of 141 upregulated DE-NMGenes and 175 downregulated DE-NMGenes. Bar graph of the significantly enriched BP **(A)** and KEGG **(B)**. Blue and red bars represent the enriched BP terms or KEGG pathways for upregulated and downregulated DE-NMGenes, respectively.

minimum normalized enrichment score >3 and the FDR on motif similarity <0.001 (**Table 3**). The TF with the largest number of targets is *PBX3* (*pre-B*-cell leukemia transcription factor 3), which regulates 145 DE-NMGenes. Eight of the 17 TFs, namely, *BACH1* (*BTB Domain and CNC Homolog 1*), ATF3 (*Activating Transcription Factor 3*), *XBP1* (*X-Box*

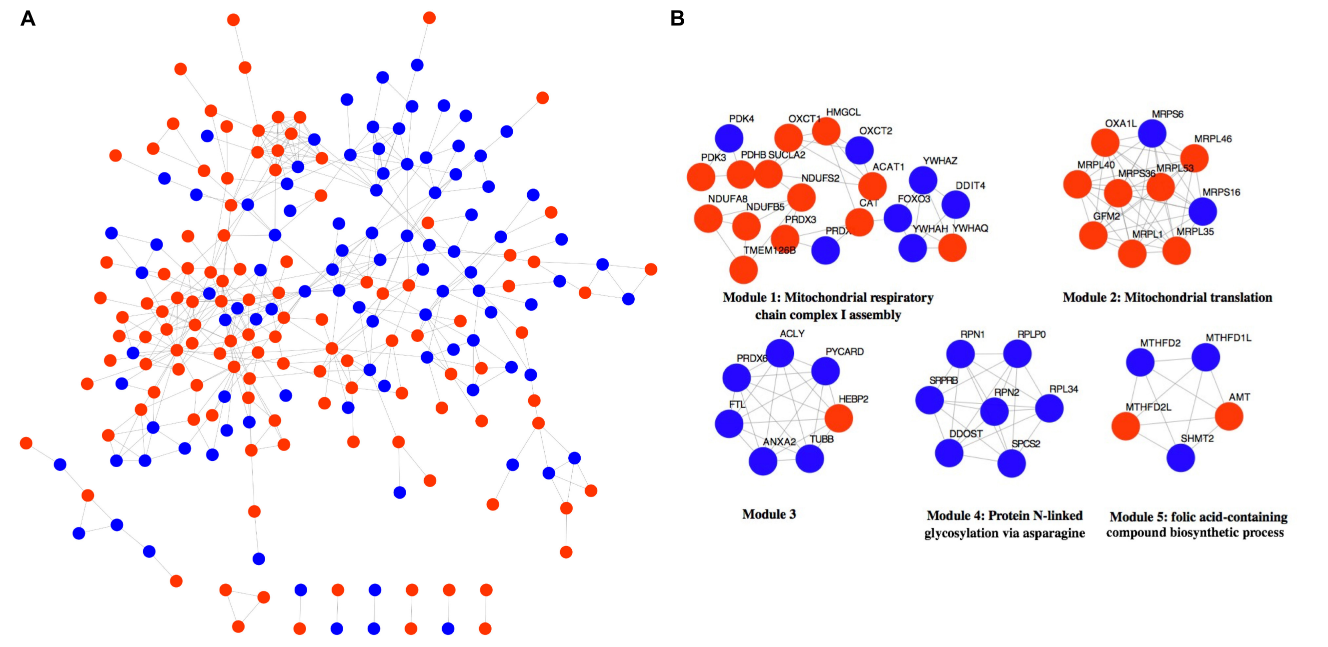


FIGURE 3 | Network analysis of 316 DE-NMGenes. **(A)** The PPI network consists of 440 edges and 215 nodes. Red and blue nodes represent upregulated and downregulated DE-NMGenes, respectively. **(B)** The network representation of modules detected from the PPI network. Red and blue nodes represent upregulated and downregulated NMGenes, respectively.

Binding Protein 1), *KLF4* (*Kruppel Like Factor 4*), *MEF2C* (*Myocyte Enhancer Factor 2C*), *JUND* (*JunD Proto-Oncogene*), *MYC* (*MYC Proto-Oncogene*) and *YY1* (*YY1 Transcription Factor*), have already been proven to play important roles in pathological hypertrophy. The remaining nine genes with unknown roles in HCM were good candidates for further experimental verification.

TABLE 2 | Hubs in the PPI network.

Gene symbol	Full name	Degree
DLD	Dihydrolipoamide Dehydrogenase	19
ACLY	ATP Citrate Lyase	13
CAT	Catalase	13
ACADM	Acyl-CoA Dehydrogenase Medium Chain	12
HADH	Hydroxyacyl-CoA Dehydrogenase	12
MRPL46	Mitochondrial Ribosomal Protein L46	12
MRPL53	Mitochondrial Ribosomal Protein L53	11
MRPL1	Mitochondrial Ribosomal Protein L1	11
MRPL40	Mitochondrial Ribosomal Protein L40	11
MRPS16	Mitochondrial Ribosomal Protein S16	11
ACAT1	Acetyl-CoA Acetyltransferase 1	11
RPLP0	Ribosomal Protein Lateral Stalk Subunit P0	11
OXA1L	*OXA1L* Mitochondrial Inner Membrane Protein	11

DLD, patients with point mutations (p.D479V and p.R482G) at the DLD homodimer interface were affected with HCM (Shany et al., 1999; Odievre et al., 2005). CAT, JMJD1A (Jumonji domain containing 1A) represses the development of cardiomyocyte hypertrophy by upregulating the expression of CAT (Zang et al., 2020). ACLY, ACLY was associated with TAC (thoracic aortic constriction) induced cardiac hypertrophy by regulating autophagy (Marino et al., 2014).

Validation of the Differentially Expressed NMGenes *in vivo*

To validate the identified genes *in vivo*, the samples were extracted from control and transverse aortic constriction (TAC) mice to identify whether the mRNA levels of the top five genes that have not been proven to play important roles in cardiac hypertrophy were consistent with the bioinformatic analysis. In the TAC group, *MYH7*, *ANP*, and *BNP* expression levels were increased (**Figures 4A–C**), indicating that pressure overload successfully induced cardiac hypertrophy in the mouse TAC model. Interestingly, the expression of *DDIT4, TKT, CLIC1, DDOST,* and *SNCA* in the mouse TAC model were all increased compared with the sham operation group (**Figures 4D–H**).

DISCUSSION

Although many studies have been conducted to explore the pathogenesis of HCM, the role of mitochondria in HCM development and progression remains largely unknown. More than 99% of mitochondrial proteins are encoded by nDNA, so NMGenes are more responsible for mitochondrial function (Ferramosca and Zara, 2013). Exploring the expression pattern of NMGenes in HCM will help us better understand the molecular mechanism of mitochondria in HCM. Therefore, we performed a comprehensive comparative analysis of NMGenes in HCM by comparing transcriptional data in HCM patients and normal healthy controls.

Differential expression analysis showed that the proportion of differentially expressed genes in NMGenes was significantly

TABLE 3 | TFs potentially regulating the expression of 316 DE-NMGenes.

#TF	NES	#Targets	Function in pathological hypertrophy
ATF4 (Activating Transcription Factor 4)	6.73	21	–
BACH1	5.27	101	Deletion of *BACH1* caused significant reductions in left ventricular hypertrophy (Mito et al., 2008).
NFYA (Nuclear Transcription Factor Y Subunit Alpha)	4.77	75	–
PBX3 (PBX Homeobox 3)	4.45	145	–
NFYC (Nuclear Transcription Factor Y Subunit Gamma)	4.34	117	–
ATF3	4.32	46	Ectopic expression of *ATF3* was sufficient to promote cardiac hypertrophy (Koren et al., 2013).
XBP1	3.98	43	Myocardial XBP1s protein was significantly increased in hypertrophic and failing heart (Duan et al., 2016).
KLF4	3.63	45	Overexpression of *KLF4* in neonatal rat ventricular myocytes inhibits cardiomyocyte hypertrophy (Liao et al., 2010).
MEF2C	3.54	91	*MEF2C* silencing attenuated load-induced left ventricular hypertrophy by modulating mTOR/S6K pathway in mice (Pereira et al., 2009).
GATA1 (GATA Binding Protein 1)	3.37	19	–
JUND	3.27	15	*JUND* could attenuate phenylephrine-mediated cardiomyocyte hypertrophy by negatively regulating AP-1 transcriptional activity (Hilfiker-Kleiner et al., 2006).
IRF2 (Interferon Regulatory Factor 2*)*	3.27	34	–
MYBL2 (MYB Proto-Oncogene Like 2)	3.27	90	–
MYC	3.17	17	*MYC* overexpression could induce cardiac hypertrophy (Olson et al., 2013).
GATA5 (GATA Binding Protein 5)	3.14	13	–
RARA (Retinoic Acid Receptor Alpha)	3.12	9	–
YY1	3.02	39	*YY1* could prevent cardiac hypertrophy (Sucharov et al., 2008)and suppresses dilated cardiomyopathy and cardiac fibrosis (Tan et al., 2019).

#TFs regulating DE-NMGenes are showed. TFs already proved to play important roles in pathological hypertrophy are marked in bold.

higher than the proportion of overall genes detected on the microarray, highlighting the importance of NMGenes in HCM. For the top 10 NMGenes with the highest fold change, four upregulated genes (i.e., *PDK4*, *STAT3*, *HCLS1,* and *FKBP11*) and four downregulated genes (i.e., *GATM*, *ATPIF1*, *CPT1B,* and *GJA1*) have already been shown to play important roles in pathological hypertrophy (**Table 1**). Importantly, the other genes with undetermined roles in HCM are good candidates for further experimental verification. Consistent with the bioinformatic analysis, *DDIT4, TKT, CLIC1* and *DDOST* mRNA expression increased in TAC mice compared with the sham operation group, suggesting that these genes may play an important role in promoting pathological hypertrophy. However, in contrast with the bioinformatic analysis, its expression at the mRNA level increased at 4 weeks after TAC. We speculate that it would decrease in the earlier or later time of TAC, as the duration of the pressure overload can affect the expression of associated genes (Souders et al., 2012). These DE-NMGenes provide a new perspective on the mechanisms in HCM. For example, *CLIC1,* as a metamorphic protein, is abundantly expressed in the heart (Ponnalagu et al., 2016); however, its function in the heart is far from fully understood. Direct evidence has shown that *CLIC1* plays a significant role in ischemia-reperfusion (IR) injury by regulating reactive oxygen species (ROS) generation (Gururaja Rao et al., 2020). Previous studies have demonstrated that abnormal production of ROS in cardiomyocytes is closely related to the occurrence and development of HCM (Hafstad et al., 2013; Brown and Griendling, 2015). We speculate that *CLIC1* is involved in the progression of HCM by regulating the generation of ROS and might be a potential therapeutic target for cardiac hypertrophy. *DDIT4* is an inhibitor of mTOR signaling, which plays a key regulatory role in cardiovascular pathology (Sciarretta et al., 2014). It is possible that *DDIT4* is involved in the progression of HCM by regulating mTOR signaling.

KEGG pathway analysis showed that abnormal expression of metabolically related pathways such as pyruvate metabolism and fatty acid metabolism (**Figure 2B**) may contribute to the pathogenesis of HCM. In the normal heart, mitochondrial fatty acid oxidation is the main (70–80%) source of energy, and the remaining 20–30% of ATP production derives largely from glucose oxidation (Sacchetto et al., 2019). Fatty acid metabolism disturbances are common in HCM patients, and mutations in the fatty acid oxidation pathway can result in HCM (Marin-Garcia and Goldenthal, 2002a). Fatty acid oxidation involves two key steps: fatty acid transfer and ββ-oxidation. Our results showed that genes involved in fatty acid transport (*CPT1B and CPT2)* or β-oxidation (*ACADSB, ACADM, ACADL,* and *HADH)* were all significantly downregulated (**Supplementary Table 1**). *CPT1B* was one of the top 10 downregulated DE-NMGenes and its deficiency could cause heart lipotoxicity, leading to exacerbated

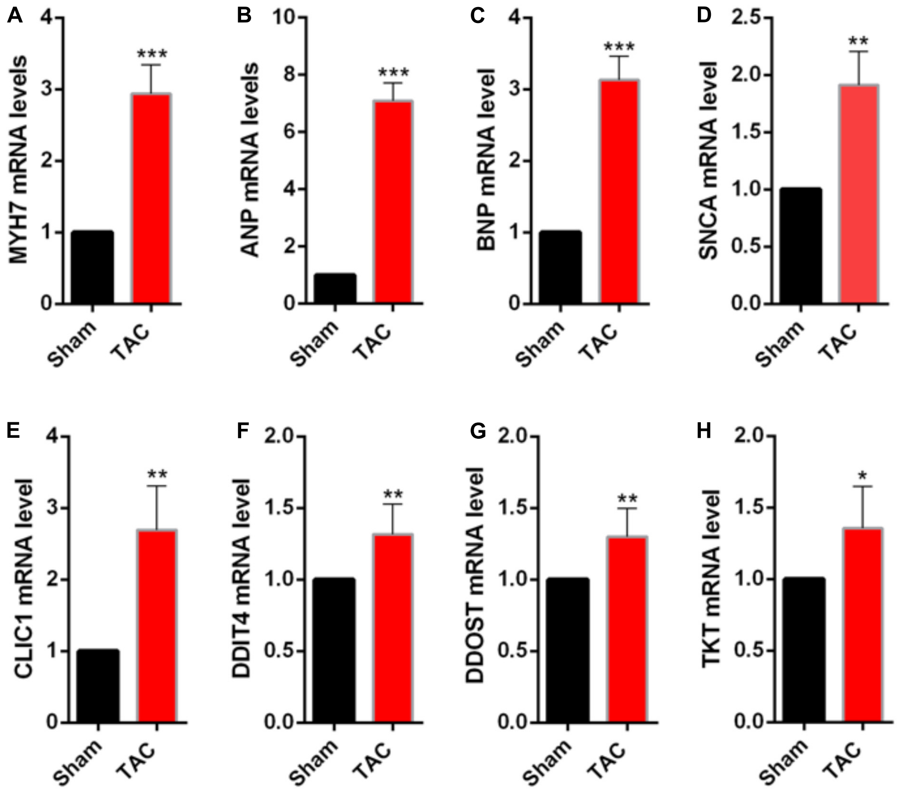


FIGURE 4 | Validation of the differentially expressed NMGenes in TAC mice. **(A–C)** mRNA of the genes indicating cardiac hypertrophy increased. **(D–H)** mRNA of the differentially expressed NMGenes also increased. Data are presented as the mean ± SD. *P < 0.05, **P < 0.01, ***P < 0.001 (Student's *t*-test); *n* = 5 samples per group. TAC, transverse aortic constriction.

cardiac pathology (He et al., 2012). *ACADM* and *HADH* were two hubs in the PPI network. Rats with hypertrophic myocardium showed impaired fatty acid oxidation and decreased expression of *ACADM* and *ACADL* (Doenst et al., 2010). In rats with cardiac hypertrophy caused by left ventricular volume overload, HADH activity was significantly reduced (Lachance et al., 2014). Pyruvate metabolism is a key step in glucose oxidation. Compared with normal hearts, glucose oxidation was actually lower in hypertrophied hearts (Allard et al., 1997). Glucose oxidation and fatty acid oxidation are under fine regulation during disease progression, although there is still controversy, allowing us to consider the treatment of HCM from the perspective of energy metabolism. New treatments include inhibiting enzymes related to fatty acid oxidation and directly increasing the oxidation of glucose and pyruvate, which may bring light to patients with cardiomyopathy in the future.

Genes rarely act alone and usually perform their functions in connection with other genes. Moreover, genes with relatively small but significant changes in expression can also contribute to the phenotypes of interest. However, differential expression analysis focuses only on individual gene expression without considering its close connection with other genes. PPI network-based analysis might largely overcome these limitations by combining gene expression and connections. In the present study, we performed PPI network analysis and obtained 440 interactions among 215 DE-NMGenes. We found that compared with random gene sets from the genome, DE-NMGenes formed significantly more interactions, which indicated that DE-NMGenes were biologically connected to form a group. Our results identified five closely connected modules that might contribute to the development of HCM. In addition, we highlighted 13 hub genes with a high level of network connectivity but relatively modest changes in expression. Hubs *DLD, CAT, ACADM,* and *HADH* have already been proven to be involved in the progression of HCM or pathological hypertrophy, and the role of the remaining hubs in HCM deserves further investigation. The top 2 hub gene, *ACLY,* is an essential cytosolic enzyme for generating acetyl-CoA, a key metabolite for glucose, fatty acid, and amino acid catabolism. A mendelian randomization study found *ACLY* to be a promising target for cardiovascular protection (Ference et al., 2019). In addition, we also noticed that five of the 13 hubs were genes encoding mitochondrial ribosomal proteins (MRPs), which assist protein synthesis within mitochondria. These MRPs were grouped into module 2 in the network module analysis (**Figure 3B**). Mutations in *MRPL3*, *MRPS14*, *MRPS22*, and *MRPL44* could cause HCM accompanied by multiorgan diseases (Smits et al., 2011; El-Hattab and Scaglia, 2016). Therefore, these MRPs are functionally connected, and the consistent downregulation of *MRPL46*, *MRPL53*, *MRPL1,* and *MRPL40* may cause mitochondrial translation deficiency, which would result in a severe phenotype in HCM.

Generally, gene expression is under the fine turn regulation of TFs. Among the 17 TFs predicted in this work, more than half have been shown to be associated with pathological hypertrophy (**Table 3**). The remaining 9 TFs (i.e., *ATF4*, *NFYA*, *PBX3*, *NFYC*, *GATA1*, *MYBL2*, *GATA5,* and *RARA* were good candidates for further experimental verification. *NFY* (*nuclear transcription factor Y*) is a heterotrimeric TF complex consisting of three subunits, *NFYA*, *NFYB* and *NFYC*. In this work, *NFYA* and *NFYC* were predicted to regulate 75 and 117 DE-NMGenes, respectively. By analyzing the targets of *NFYA* and *NFYC* in DE-NMGenes, we found that they were both enriched in the GO term "negative regulation of apoptotic process" with *p*-values of $4.1{*}10^{-7}$ and $2.5{*}10^{-4}$, respectively. Although the role of *NFY* in cardiovascular disease has not been reported, *NFY* is involved in cancer by regulating apoptosis (Gurtner et al., 2010). Moreover, *NFYA* and *NFYC* were both significantly differentially expressed in HCM. We speculate that *NFYA* and *NFYC* may be involved in the pathogenesis of HCM by regulating apoptosis, which provides us with a new perspective to understand the relationship between *NFY* and HCM.

The GATA TF family comprises six members (named GATA1-6) that are involved in the regulation of growth, differentiation, survival and maintenance of body function. Previous studies have underscored the pivotal roles of the GATA family in cardiac hypertrophy (Pikkarainen et al., 2004). Mutations in *GATA2*, *GATA4,* and *GATA6* were identified in patients with HCM (Alonso-Montes et al., 2017). Overexpression of either *GATA4* or *GATA6* could induce cardiac hypertrophy both *in vitro* and *in vivo* (Liang et al., 2001). *GATA5* and *GATA1* are closely related to cardiomyopathy diseases such as dilated cardiomyopathy, although their role in HCM has not yet been reported (Zhang et al., 2015). In this work, *GATA1* and *GATA5* were predicted to regulate 19 and 13 DE-NMGenes, respectively (**Table 3**). Given that the functional characteristics of *GATA5* and *GATA1* overlap at least partly with those of other *GATA* TFs and that *GATA1* and *GATA5* regulate DE-NMGenes, it is reasonable to speculate that *GATA1* and *GATA5* may contribute to HCM.

CONCLUSION

The present study was the first effort to evaluate the global expression pattern of NMGenes in HCM. Based on differential expression analysis, we found that NMGenes were significantly changed and identified 316 DE-NMGenes. Further GO enrichment analysis showed that downregulated DE-NMGenes were more functionally diverse. These DE-NMGenes participated in 10 significant pathways, and nine of these pathways were metabolically related. PPI network analysis showed that 13 DE-NMGenes with high node connectivity were selected as hubs. Finally, a total of 17 TFs were predicted to potentially regulate the expression of the 316 DE-NMGenes, and TFs (such as *ATF4, NFYA*, *NFYC, GATA1,* and *GATA5*) might play roles in HCM. This analysis will provide valuable information for future research on the molecular mechanisms of HCM and offer clues for the discovery of novel therapeutic strategies.

MATERIALS AND METHODS

Data Collection

Normalized gene expression data (GSE36961) were collected from the GEO database (Clough and Barrett, 2016). NMGenes were collected from the MitoCarta (Version 2.0) (Calvo et al., 2016), MitoMiner (Version 4.0) (Smith and Robinson, 2019), IMPI[2] and UniProt databases (UniProt, 2019).

Differential Expression Analysis

To identify DEGs between HCM and normal healthy hearts, limma (Version 3.40.6), an R package in Bioconductor, was utilized (Ritchie et al., 2015). Genes with BH-corrected *p*-values less than 0.01 and fold changes (FCs) larger than 1.2 were selected as significantly differentially expressed. We have deposited the analysis code to a public repository[3].

Functional Enrichment Analysis

GO BP and KEGG pathway enrichment analyses of DE-NMGenes were performed using DAVID, an online functional annotation tool, to understand the biological significance of a list of genes (da Huang et al., 2009). In this work, GO BP and KEGG pathways with BH-corrected *p*-values less than 0.05 were selected as significant.

PPI Network Construction

The PPI network of DE-NMGenes was constructed using the STRING database, and an online database provides information regarding the predicted and experimental protein interactions (Szklarczyk et al., 2019). In this work, PPIs between DE-NMGenes with interaction scores larger than 0.7 were retained.

Network Module Analysis

A network module is defined as a group of genes participating in the same biological function. In this work, we detected network modules from the constructed PPI network using MCODE (Bader and Hogue, 2003), a plugin in Cytoscape[4]. Given the following parameters: a degree cutoff = 2, node score cutoff = 0.2, k-score = 2 and max. depth = 100, modules with scores > 3 and a number of nodes > 5 were selected. GO BP enrichment analysis of modules was performed using DAVID, and BH-corrected *p* values < 0.05 were selected as significant.

Prediction of TFs Regulating DE-NMGenes

To predict TFs regulating DE-NMGenes, iRegulon (Version: 1.3), a plugin in Cytoscape, was employed (Janky et al., 2014). The iRegulon plugin uses motif and track discovery

[2] http://impi.mrc-mbu.cam.ac.uk/

[3] https://github.com/ZhenhongJiang/Nuclear-encoded-mitochondrial-genes-in-HCM.git

[4] https://cytoscape.org/

in a set of coregulated genes to identify regulons. Given the following parameters: motif collection (10 kb, 9,713 PWMs), track collection (1120 ChIP-seq tracks of ENCODE raw signals), putative regulatory region (20 kb centered around TSS), motif rankings database (20 kb region centered around TSS, 7 species), track of rankings database (20 kb centered around TSS, ChIP-seq-derived), minimum identity between orthologous genes = 0 and maximum false discovery rate on motif similarity = 0.001, TFs with the NES (normalized enrichment score) larger than 3 were selected. The higher the NES was, the more reliable the TFs were.

Animals and Surgical Procedures

All experiments involving animals were approved by the Animal Ethics and Experimentation Committee of Nanchang University and carried out according to the "Guide for the Care and Use of Laboratory Animals." Male C57BL/6 mice, aged 8 weeks and weighing 20–25g, were purchased from the SlacJingda Experimental Animals Company [Changsha, Hunan Province, China]. A total of 20 mice were divided into two groups (ten mice per group): the sham operation group and the TAC group. TAC was performed as previously described (Zhang et al., 2020). Briefly, mice were induced with 5% isoflurane and intubated orally and then maintained at 2% isoflurane during surgery with mechanical ventilation. After a midline sternotomy, the aortic arch was exposed. Constriction was performed by tying a 5-0 silk suture around a 27-gauge needle overlying the arch between the origin of the brachiocephalic trunk and left common carotid artery. For the sham operation group, 10 mice underwent the same procedure, but the suture was withdrawn without tying. Then, the thorax and skin were closed by using 6-0 polypropylene sutures. Four weeks after surgery, the mice were euthanized, and their hearts were quickly excised for further evaluation.

Quantitative Real-Time PCR Analysis

Total RNA was extracted from mouse cardiac tissues using TRIzol reagent (Invitrogen, New York, United States), and then the quality and concentration of RNA were determined using an Agilent Bioanalyzer 2100 according to the manufacturer's instructions. The cDNAs were generated by MMLV transcriptase (BioRAD, United States), and quantitative real-time PCR assays were performed as previously described (Yu et al., 2018). Triplicate PCR amplifications were performed for each sample, and the mRNA levels were normalized to GAPDH. The comparative threshold cycle method (2-ΔΔCT) was applied to estimate the relative gene expression of cardiac tissues between the TAC and sham operation groups. The primer sequences for *CLIC1, DDIT4, TKT, DDOST, SNCA, MYH7, ANP*, and *BNP* are listed in **Supplementary Table 4**. The differences in mRNA levels between the two groups were evaluated by using Student's *t*-tests. A *P*-value < 0.05 was considered significant.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GEO (Gene Expression Omnibus) database, Accession number GSE36961: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36961.

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Ethics and Experimentation Committee of Nanchang University.

AUTHOR CONTRIBUTIONS

KH and ZJ were responsible for the entire project and revised the draft of the manuscript. ZT, LW, and ZJ collected the data, performed the analyses, and drafted the first version of the manuscript. All authors took part in the interpretation of the results and preparation of the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (NSFC, 32060154).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.670787/full#supplementary-material

Supplementary Table 1 | 2927 DEGs between HCM and control samples

Supplementary Table 2 | Annotation results for 141 upregulated DE-NMGenes, 175 DE-NMGenes, and 4 network modules

Supplementary Table 3 | Modules detected from PPI network

Supplementary Table 4 | Real-time PCR Primer sequences.

REFERENCES

Allard, M. F., Henning, S. L., Wambolt, R. B., Granleese, S. R., English, D. R., and Lopaschuk, G. D. (1997). Glycogen metabolism in the aerobic hypertrophied rat heart. *Circulation* 96, 676–682. doi: 10.1161/01.cir.96.2.676

Alonso-Montes, C., Rodriguez-Reguero, J., Martin, M., Gomez, J., Coto, E., Naves-Diaz, M., et al. (2017). Rare genetic variants in GATA transcription factors in patients with hypertrophic cardiomyopathy. *J. Investig. Med.* 65, 926–934. doi: 10.1136/jim-2016-000364

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Barabasi, A. L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Brown, D. I., and Griendling, K. K. (2015). Regulation of signal transduction by reactive oxygen species in the cardiovascular system. *Circ. Res.* 116, 531–549. doi: 10.1161/CIRCRESAHA.116.303584

Calvo, S. E., Clauser, K. R., and Mootha, V. K. (2016). MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* 44, D1251–D1257. doi: 10.1093/nar/gkv1003

Chen, L., Zhao, L., Samanta, A., Mahmoudi, S. M., Buehler, T., Cantilena, A., et al. (2017). STAT3 balances myocyte hypertrophy vis-a-vis autophagy in response to Angiotensin II by modulating the AMPKalpha/mTOR axis. *PLoS One* 12:e0179835. doi: 10.1371/journal.pone.0179835

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

da Huang, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Doenst, T., Pytel, G., Schrepper, A., Amorim, P., Farber, G., Shingu, Y., et al. (2010). Decreased rates of substrate oxidation ex vivo predict the onset of heart failure and contractile dysfunction in rats with pressure overload. *Cardiovasc. Res.* 86, 461–470. doi: 10.1093/cvr/cvp414

Duan, Q., Ni, L., Wang, P., Chen, C., Yang, L., Ma, B., et al. (2016). Deregulation of XBP1 expression contributes to myocardial vascular endothelial growth factor-A expression and angiogenesis during cardiac hypertrophy in vivo. *Aging Cell* 15, 625–633. doi: 10.1111/acel.12460

El-Hattab, A. W., and Scaglia, F. (2016). Mitochondrial cardiomyopathies. *Front. Cardiovasc. Med.* 3:25. doi: 10.3389/fcvm.2016.00025

Ference, B. A., Ray, K. K., Catapano, A. L., Ference, T. B., Burgess, S., Neff, D. R., et al. (2019). Mendelian randomization study of ACLY and cardiovascular disease. *N. Engl. J. Med.* 380, 1033–1042. doi: 10.1056/NEJMoa1806747

Ferramosca, A., and Zara, V. (2013). Biogenesis of mitochondrial carrier proteins: molecular mechanisms of import into mitochondria. *Biochim. Biophys. Acta* 1833, 494–502. doi: 10.1016/j.bbamcr.2012.11.014

Fontes, M. S., van Veen, T. A., de Bakker, J. M., and van Rijen, H. V. (2012). Functional consequences of abnormal Cx43 expression in the heart. *Biochim. Biophys. Acta* 1818, 2020–2029. doi: 10.1016/j.bbamem.2011.07.039

Gao, J., Collyer, J., Wang, M., Sun, F., and Xu, F. (2020). Genetic dissection of hypertrophic cardiomyopathy with myocardial RNA-seq. *Int. J. Mol. Sci.* 21:3040. doi: 10.3390/ijms21093040

Gersh, B. J., Maron, B. J., Bonow, R. O., Dearani, J. A., Fifer, M. A., Link, M. S., et al. (2011). 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: executive summary: a report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines. *Circulation* 124, 2761–2796. doi: 10.1161/CIR.0b013e318223e230

Gurtner, A., Fuschi, P., Martelli, F., Manni, I., Artuso, S., Simonte, G., et al. (2010). Transcription factor NF-Y induces apoptosis in cells expressing wild-type p53 through E2F1 upregulation and p53 activation. *Cancer Res.* 70, 9711–9720. doi: 10.1158/0008-5472.CAN-10-0721

Gururaja Rao, S., Patel, N. J., and Singh, H. (2020). Intracellular chloride channels: novel biomarkers in diseases. *Front. Physiol.* 11:96. doi: 10.3389/fphys.2020.00096

Hafstad, A. D., Nabeebaccus, A. A., and Shah, A. M. (2013). Novel aspects of ROS signalling in heart failure. *Basic Res. Cardiol.* 108:359. doi: 10.1007/s00395-013-0359-8

He, L., Kim, T., Long, Q., Liu, J., Wang, P., Zhou, Y., et al. (2012). Carnitine palmitoyltransferase-1b deficiency aggravates pressure overload-induced cardiac hypertrophy caused by lipotoxicity. *Circulation* 126, 1705–1716. doi: 10.1161/CIRCULATIONAHA.111.075978

Hilfiker-Kleiner, D., Hilfiker, A., Castellazzi, M., Wollert, K. C., Trautwein, C., Schunkert, H., et al. (2006). JunD attenuates phenylephrine-mediated cardiomyocyte hypertrophy by negatively regulating AP-1 transcriptional activity. *Cardiovasc. Res.* 71, 108–117. doi: 10.1016/j.cardiores.2006.02.032

Hu, X., Shen, G., Lu, X., Ding, G., and Shen, L. (2019). Identification of key proteins and lncRNAs in hypertrophic cardiomyopathy by integrated network analysis. *Arch. Med. Sci.* 15, 484–497. doi: 10.5114/aoms.2018.75593

Janky, R., Verfaillie, A., Imrichova, H., Van de Sande, B., Standaert, L., Christiaens, V., et al. (2014). iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput. Biol.* 10:e1003731. doi: 10.1371/journal.pcbi.1003731

Jensen, M., Muller, C., Choe, C. U., Schwedhelm, E., and Zeller, T. (2020). Analysis of L-arginine:glycine amidinotransferase-, creatine- and homoarginine-dependent gene regulation in the murine heart. *Sci. Rep.* 10:4821. doi: 10.1038/s41598-020-61638-3

Koren, L., Elhanani, O., Kehat, I., Hai, T., and Aronheim, A. (2013). Adult cardiac expression of the activating transcription factor 3, ATF3, promotes ventricular hypertrophy. *PLoS One* 8:e68396. doi: 10.1371/journal.pone.0068396

Lachance, D., Dhahri, W., Drolet, M. C., Roussel, E., Gascon, S., Sarrhini, O., et al. (2014). Endurance training or beta-blockade can partially block the energy metabolism remodeling taking place in experimental chronic left ventricle volume overload. *BMC Cardiovasc. Disord.* 14:190. doi: 10.1186/1471-2261-14-190

Li, J., Wu, Z., Zheng, D., Sun, Y., Wang, S., and Yan, Y. (2019). Bioinformatics analysis of the regulatory lncRNAmiRNAmRNA network and drug prediction in patients with hypertrophic cardiomyopathy. *Mol. Med. Rep.* 20, 549–558. doi: 10.3892/mmr.2019.10289

Liang, Q., De Windt, L. J., Witt, S. A., Kimball, T. R., Markham, B. E., and Molkentin, J. D. (2001). The transcription factors GATA4 and GATA6 regulate cardiomyocyte hypertrophy in vitro and in vivo. *J. Biol. Chem.* 276, 30245–30253. doi: 10.1074/jbc.M102174200

Liao, X., Haldar, S. M., Lu, Y., Jeyaraj, D., Paruchuri, K., Nahori, M., et al. (2010). Kruppel-like factor 4 regulates pressure-induced cardiac hypertrophy. *J. Mol. Cell Cardiol.* 49, 334–338. doi: 10.1016/j.yjmcc.2010.04.008

Lim, D. S., Roberts, R., and Marian, A. J. (2001). Expression profiling of cardiac genes in human hypertrophic cardiomyopathy: insight into the pathogenesis of phenotypes. *J. Am. Coll. Cardiol.* 38, 1175–1180. doi: 10.1016/s0735-1097(01)01509-1

Lin, C. Y., Lee, T. L., Chiu, Y. Y., Lin, Y. W., Lo, Y. S., Lin, C. T., et al. (2015). Module organization and variance in protein-protein interaction networks. *Sci. Rep.* 5:9386. doi: 10.1038/srep09386

Liu, X., Ma, Y., Yin, K., Li, W., Chen, W., Zhang, Y., et al. (2019). Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy. *Sci. Data* 6:90. doi: 10.1038/s41597-019-0094-6

Marian, A. J., and Braunwald, E. (2017). Hypertrophic cardiomyopathy: genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circ. Res.* 121, 749–770. doi: 10.1161/CIRCRESAHA.117.311059

Marin-Garcia, J., and Goldenthal, M. J. (2002a). Fatty acid metabolism in cardiac failure: biochemical, genetic and cellular analysis. *Cardiovasc. Res.* 54, 516–527. doi: 10.1016/s0008-6363(01)00552-1

Marin-Garcia, J., and Goldenthal, M. J. (2002b). Understanding the impact of mitochondrial defects in cardiovascular disease: a review. *J. Card. Fail.* 8, 347–361. doi: 10.1054/jcaf.2002.127774

Marino, G., Pietrocola, F., Kong, Y., Eisenberg, T., Hill, J. A., Madeo, F., et al. (2014). Dimethyl alpha-ketoglutarate inhibits maladaptive autophagy in pressure overload-induced cardiomyopathy. *Autophagy* 10, 930–932. doi: 10.4161/auto.28235

Mito, S., Ozono, R., Oshima, T., Yano, Y., Watari, Y., Yamamoto, Y., et al. (2008). Myocardial protection against pressure overload in mice lacking Bach1, a transcriptional repressor of heme oxygenase-1. *Hypertension* 51, 1570–1577. doi: 10.1161/HYPERTENSIONAHA.107.102566

Mori, J., Basu, R., McLean, B. A., Das, S. K., Zhang, L., Patel, V. B., et al. (2012). Agonist-induced hypertrophy and diastolic dysfunction are associated with selective reduction in glucose oxidation: a metabolic contribution to heart failure with normal ejection fraction. *Circ. Heart Fail.* 5, 493–503. doi: 10.1161/CIRCHEARTFAILURE.112.966705

Odievre, M. H., Chretien, D., Munnich, A., Robinson, B. H., Dumoulin, R., Masmoudi, S., et al. (2005). A novel mutation in the dihydrolipoamide dehydrogenase E3 subunit gene (DLD) resulting in an atypical form of alpha-ketoglutarate dehydrogenase deficiency. *Hum. Mutat.* 25, 323–324. doi: 10.1002/humu.9319

Olson, A. K., Ledee, D., Iwamoto, K., Kajimoto, M., O'Kelly Priddy, C., Isern, N., et al. (2013). C-Myc induced compensated cardiac hypertrophy increases free fatty acid utilization for the citric acid cycle. *J. Mol. Cell Cardiol.* 55, 156–164. doi: 10.1016/j.yjmcc.2012.07.005

Pereira, A. H., Clemente, C. F., Cardoso, A. C., Theizen, T. H., Rocco, S. A., Judice, C. C., et al. (2009). MEF2C silencing attenuates load-induced left ventricular hypertrophy by modulating mTOR/S6K pathway in mice. *PLoS One* 4:e8472. doi: 10.1371/journal.pone.0008472

Pfanner, N., Warscheid, B., and Wiedemann, N. (2019). Mitochondrial proteins: from biogenesis to functional networks. *Nat. Rev. Mol. Cell Biol.* 20, 267–284. doi: 10.1038/s41580-018-0092-0

Pikkarainen, S., Tokola, H., Kerkela, R., and Ruskoaho, H. (2004). GATA transcription factors in the developing and adult heart. *Cardiovasc. Res.* 63, 196–207. doi: 10.1016/j.cardiores.2004.03.025

Ponnalagu, D., Gururaja Rao, S., Farber, J., Xin, W., Hussain, A. T., Shah, K., et al. (2016). Molecular identity of cardiac mitochondrial chloride intracellular channel proteins. *Mitochondrion* 27, 6–14. doi: 10.1016/j.mito.2016.01.001

Reibel, D. K., O'Rourke, B., Foster, K. A., Hutchinson, H., Uboh, C. E., and Kent, R. L. (1986). Altered phospholipid metabolism in pressure-overload hypertrophied hearts. *Am. J. Physiol.* 250, H1–H6. doi: 10.1152/ajpheart.1986.250

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

Sacchetto, C., Sequeira, V., Bertero, E., Dudek, J., Maack, C., and Calore, M. (2019). Metabolic alterations in inherited cardiomyopathies. *J. Clin. Med.* 8:2195. doi: 10.3390/jcm8122195

Saoura, M., Powell, C. A., Kopajtich, R., Alahmad, A., Al-Balool, H. H., Albash, B., et al. (2019). Mutations in ELAC2 associated with hypertrophic cardiomyopathy impair mitochondrial tRNA 3′-end processing. *Hum. Mutat.* 40, 1731–1748. doi: 10.1002/humu.23777

Sciarretta, S., Volpe, M., and Sadoshima, J. (2014). Mammalian target of rapamycin signaling in cardiac physiology and disease. *Circ. Res.* 114, 549–564. doi: 10.1161/CIRCRESAHA.114.302022

Shany, E., Saada, A., Landau, D., Shaag, A., Hershkovitz, E., and Elpeleg, O. N. (1999). Lipoamide dehydrogenase deficiency due to a novel mutation in the interface domain. *Biochem. Biophys. Res. Commun.* 262, 163–166. doi: 10.1006/bbrc.1999.1133

Smith, A. C., and Robinson, A. J. (2019). MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. *Nucleic Acids Res.* 47, D1225–D1228. doi: 10.1093/nar/gky1072

Smits, P., Saada, A., Wortmann, S. B., Heister, A. J., Brink, M., Pfundt, R., et al. (2011). Mutation in mitochondrial ribosomal protein MRPS22 leads to Cornelia de Lange-like phenotype, brain abnormalities and hypertrophic cardiomyopathy. *Eur. J. Hum. Genet.* 19, 394–399. doi: 10.1038/ejhg.2010.214

Souders, C. A., Borg, T. K., Banerjee, I., and Baudino, T. A. (2012). Pressure overload induces early morphological changes in the heart. *Am. J. Pathol.* 181, 1226–1235. doi: 10.1016/j.ajpath.2012.06.015

Sucharov, C. C., Dockstader, K., and McKinsey, T. A. (2008). YY1 protects cardiac myocytes from pathologic hypertrophy by interacting with HDAC5. *Mol. Biol. Cell* 19, 4141–4153. doi: 10.1091/mbc.E07-12-1217

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

Tan, C. Y., Wong, J. X., Chan, P. S., Tan, H., Liao, D., Chen, W., et al. (2019). Yin Yang 1 suppresses dilated cardiomyopathy and cardiac fibrosis through regulation of Bmp7 and Ctgf. *Circ. Res.* 125, 834–846. doi: 10.1161/CIRCRESAHA.119.314794

Tian, R. (2003). Transcriptional regulation of energy substrate metabolism in normal and hypertrophied heart. *Curr. Hypertens. Rep.* 5, 454–458. doi: 10.1007/s11906-003-0052-7

UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Wang, X., Deng, Y., Zhang, G., Li, C., Ding, G., May, H. I., et al. (2019). Spliced X-box binding protein 1 stimulates adaptive growth through activation of mTOR. *Circulation* 140, 566–579. doi: 10.1161/CIRCULATIONAHA.118.038924

Weissler-Snir, A., Allan, K., Cunningham, K., Connelly, K. A., Lee, D. S., Spears, D. A., et al. (2019). Hypertrophic cardiomyopathy-related sudden cardiac death in young people in Ontario. *Circulation* 140, 1706–1716. doi: 10.1161/CIRCULATIONAHA.119.040271

Yang, K., Long, Q., Saja, K., Huang, F., Pogwizd, S. M., Zhou, L., et al. (2017). Knockout of the ATPase inhibitory factor 1 protects the heart from pressure overload-induced cardiac hypertrophy. *Sci. Rep.* 7:10501. doi: 10.1038/s41598-017-11251-8

Yang, W., Li, Y., He, F., and Wu, H. (2015). Microarray profiling of long non-coding RNA (lncRNA) associated with hypertrophic cardiomyopathy. *BMC Cardiovasc. Disord.* 15:62. doi: 10.1186/s12872-015-0056-7

Yu, P., Hu, L., Xie, J., Chen, S., Huang, L., Xu, Z., et al. (2018). O-GlcNAcylation of cardiac Nav1.5 contributes to the development of arrhythmias in diabetic hearts. *Int. J. Cardiol.* 260, 74–81. doi: 10.1016/j.ijcard.2018.02.099

Zang, R., Tan, Q., Zeng, F., Wang, D., Yu, S., and Wang, Q. (2020). JMJD1A represses the development of cardiomyocyte hypertrophy by regulating the expression of catalase. *Biomed. Res. Int.* 2020:5081323. doi: 10.1155/2020/5081323

Zhang, N., Zhang, Y., Qian, H., Wu, S., Cao, L., and Sun, Y. (2020). Selective targeting of ubiquitination and degradation of PARP1 by E3 ubiquitin ligase WWP2 regulates isoproterenol-induced cardiac remodeling. *Cell Death Differ.* 27, 2605–2619. doi: 10.1038/s41418-020-0523-2

Zhang, X. L., Dai, N., Tang, K., Chen, Y. Q., Chen, W., Wang, J., et al. (2015). GATA5 loss-of-function mutation in familial dilated cardiomyopathy. *Int. J. Mol. Med.* 35, 763–770. doi: 10.3892/ijmm.2014.2050

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.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 24 June 2021
doi: 10.3389/fgene.2021.673192

Immune and Metabolic Dysregulated Coding and Non-coding RNAs Reveal Survival Association in Uterine Corpus Endometrial Carcinoma

Da Liu[1] and Min Qiu[2]*

[1] Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University, Shenyang, China, [2] Department of Orthopedics, Shengjing Hospital of China Medical University, Shenyang, China

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Tingfang Chen,
University of Pennsylvania,
United States
Yu Dong,
Shanghai Jiao Tong University, China

***Correspondence:**
Min Qiu
qiumin8735@163.com

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

Received: 27 February 2021
Accepted: 14 April 2021
Published: 24 June 2021

Citation:
Liu D and Qiu M (2021) Immune
and Metabolic Dysregulated Coding
and Non-coding RNAs Reveal
Survival Association in Uterine Corpus
Endometrial Carcinoma.
Front. Genet. 12:673192.
doi: 10.3389/fgene.2021.673192

Uterine corpus endometrial carcinoma (UCEC) is one of the most common gynecologic malignancies, but only a few biomarkers have been proven to be effective in clinical practice. Previous studies have demonstrated the important roles of non-coding RNAs (ncRNAs) in diagnosis, prognosis, and therapy selection in UCEC and suggested the significance of integrating molecules at different levels for interpreting the underlying molecular mechanism. In this study, we collected transcriptome data, including long non-coding RNAs (lncRNAs), microRNAs (miRNAs), and messenger RNAs (mRNAs), of 570 samples, which were comprised of 537 UCEC samples and 33 normal samples. First, differentially expressed lncRNAs, miRNAs, and mRNAs, which distinguished invasive carcinoma samples from normal samples, were identified, and further analysis showed that cancer- and metabolism-related functions were enriched by these RNAs. Next, an integrated, dysregulated, and scale-free biological network consisting of differentially expressed lncRNAs, miRNAs, and mRNAs was constructed. Protein-coding and ncRNA genes in this network showed potential immune and metabolic functions. A further analysis revealed two clinic-related modules that showed a close correlation with metabolic and immune functions. RNAs in the two modules were functionally validated to be associated with UCEC. The findings of this study demonstrate an important clinical application for improving outcome prediction for UCEC.

Keywords: dysregulated network, endometrial carcinoma, miRNA, lncRNA, integrative analysis, TCGA, immunity, metabolism

INTRODUCTION

Cancer is one of the major public health problems worldwide and is the second leading cause of death in the United States (Siegel et al., 2021). After the rapid development in healthcare, the total decline in the cancer death rate has reached approximately 31% (Siegel et al., 2021). Nonetheless, uterine corpus endometrial carcinoma (UCEC) is still one of the most common gynecologic malignancies in many countries (Matteson et al., 2018). In the United States alone, there will be approximately 14,000 new UCEC patients and 4,000 deaths in the 2021, as predicted by Siegel et al. (Siegel et al., 2021). Generally, UCEC is prevalent among postmenopausal women due to the unstable level of estrogen (Chen et al., 2015). Different risk factors, such as smoking, high blood pressure, and being overweight, also contribute to the generation and development of UCEC (Zhang et al., 2014). In particular, changes in molecular levels are one factor contributing the development of UCEC (Li et al., 2020). However, effective therapeutic targets are still scarce in clinical practice.

Non-coding RNAs (ncRNAs), including microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), have been regarded as transcriptional noise and useless due to their low effective transcription and expression (Hyashizaki, 2004). Taking advantage of the large-scale, next-generation transcriptomic sequencing, more ncRNAs have been identified. In GENCODE v29, there are 16,066 annotated lncRNA genes, 7,577 annotated small ncRNA genes (e.g., miRNA) and thousands of other ncRNA genes. In total, there are more than 30,000 annotated ncRNA genes, which are more than protein-coding genes whose annotated number is less than 20,000. Many ncRNAs have been functionally associated with human diseases, such as cancers (Gutschner and Diederichs, 2012). HOX antisense intergenic RNA (HOXAIR), one of the most famous lncRNAs, has been reported to be associated with metastases in colorectal, liver, pancreatic, breast, and gastric cancers (Gupta et al., 2010; Kogo et al., 2011; Yang et al., 2011). Furthermore, some ncRNAs have been functionally related with UCEC. Wang found a six-miRNA signature that can predict the survival of UCEC patients (Wang et al., 2019). Many studies have investigated the pathogenesis at genomic levels using the combination of different kinds of molecules and have discovered clinical diagnostic and prognostic biomarkers. It reported that miR-21 and lncRNA AWPPH are associated with the poor prognosis of hepatocellular carcinoma but regulate cancer cell chemosensitivity and proliferation in triple-negative breast cancer (Liu et al., 2019). Dong et al. revealed two patient survival-associated RNA sets, including lncRNAs, miRNAs, and messenger RNAs (mRNAs), in invasive breast carcinoma (Dong et al., 2020). Moreover, Liu et al. identified six triplets of mRNA–lncRNA–miRNA that play a function in UCEC (Liu et al., 2017) based on the expression profiles. However, their underlying molecular mechanisms still need to be uncovered.

In this study, to investigate the underlying molecular mechanisms of the generation and development of UCEC, the expression profiles of 537 UCEC and their 33 counterpart normal samples were downloaded from the Cancer Genome Atlas (TCGA). Three different kinds of RNAs, namely, lncRNAs, miRNAs, and mRNAs, were extracted from the profiles. First, a differential expression analysis was performed, followed by a functional enrichment analysis, including a gene ontology (GO) analysis, KEGG analysis, and gene set enrichment analysis (GSEA). Then, a lncRNA–miRNA–mRNA dysregulated network was constructed, and two modules related with the survival time, metabolic function, and immune function were identified. RNAs from each module have showed a functional role in UCEC.

MATERIALS AND METHODS

Acquisition of RNA Sequencing Datasets

RNA sequencing datasets of 570 samples were downloaded from TCGA[1], including 537 UCEC samples and 33 normal samples (**Supplementary Table 1**). Each sample contained miRNAs, lncRNAs, and mRNAs simultaneously were used for downstream analyses. The annotation from GENCODE database (GENCODE v36) was used to extract lncRNAs from the expression profile. Based on the annotation file, the following biotypes were regarded as known lncRNAs: antisense, lincRNA, lncRNA, processed_transcript, sense_intronic, sense_overlapping, and TEC. The biotype "protein_coding" was used to extract mRNAs from the expression profile. Finally, 19,597 mRNAs, 15,088 lncRNAs, and 188 miRNAs were used for the downstream analysis.

Differential Expression Analysis

To remove biases, RNAs with an expression level in less than 10% of the samples were ignored, followed by a differential expression analysis with p-value < 0.05 and fold change > 2 using a t-test (Ye et al., 2018). In total, 648 differentially expressed lncRNAs, 5,831 differentially expressed mRNAs, and 342 differentially expressed miRNAs were identified (**Supplementary Table 2**). Unsupervised clustering was performed, and heat maps were drawn for differentially expressed lncRNAs, mRNAs, and miRNAs using the R package pheatmap, separately. Moreover, the R package Prcomp was used to conduct the principal component analysis (PCA).

MiRNAs and Their Targets

MiRNA target sites were downloaded from one of the most popular databases in the field, starBase v3.0 (Li et al., 2014), which predicts the miRNA target using five algorithms, i.e., TargetScan (Lewis et al., 2005), miRanda (Enright et al., 2003), Pictar2 (Krek et al., 2005), PITA (Kertesz et al., 2007), and RNA22 (Loher and Rigoutsos, 2012). MiRNAs play a function in RNA-induced silencing complexes (RISCs), or the ribonucleoprotein complexes (Fabian et al., 2010). The components of RISCs, i.e., Argonaute (AGO) family proteins, are the best characterized protein elements and are central to RISC functions (Chekulaeva and Filipowicz, 2009). Ultraviolet (UV) crosslinking and immunoprecipitation (CLIP) is one of the useful techniques in identifying specific protein–RNA interactions, including identifying the AGO–RNA–miRNA complex to illustrate miRNA functions (König et al., 2012). Thus, in this study, AGO CLIP-Seq datasets downloaded from starBase v3.0 were used to identify AGO binding sites. MiRNA-target pairs with at least one AGO binding site were considered. Finally, 40,042 miRNA–lncRNA and 1224,551 miRNA–mRNA regulatory relationships were used, which include 3,228 lncRNAs, 413 miRNAs, and 14,645 mRNAs.

Functional Enrichment Analysis

To explore the functional roles of differentially expressed molecules, GO and KEGG analyses were performed using

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

clusterProfiler (Yu et al., 2012). For ncRNAs, we first calculated the Pearson correlation coefficient between each ncRNA-mRNA pair based on the expression value across the samples, followed by the calculation of the average Pearson correlation coefficient for each mRNA across ncRNAs. Then, the top 500 co-expressed mRNAs were used. Barplots were drawn using ggplot2. To further investigate the functional roles of the key RNAs, GSEA was also performed using clusterProfiler (Yu et al., 2012).

To determine if genes in each immune (or metabolism)-related pathway are enriched in each sample, the Gene Set Variation Analysis (GSVA) (Hänzelmann et al., 2013) was performed. Gene sets annotated in immune (or metabolism)-related pathways were obtained from MSigDB[2]. GSVA scores were calculated using the R package GSVA with the single-sample GSEA method.

Construction of the Dysregulated lncRNA–miRNA–mRNA Network

First, the miRNA–lncRNA and miRNA–mRNA interactions from starBase v3.0 (Li et al., 2014) were obtained. Only differentially expressed miRNAs, lncRNAs, and mRNAs were considered for the downstream analysis. Then, the dysregulated lncRNA–miRNA–mRNA network was constructed based on the interactions. Afterward, a two-step filtering was used: (1) The correlations between each miRNA-target pair should be significant (p-value < 0.01 and $|$ correlation coefficient$| > 0.3$) across all samples using the Pearson correlation coefficient. (2) Only miRNAs shared by mRNAs and lncRNAs were used. Finally, a dysregulated network was constructed containing 1243 interactions, including 323 mRNAs, 52 miRNAs, and 53 lncRNAs (**Supplementary Table 3**). To identify functional modules, CytoCluster (Li et al., 2017), a graphical algorithm, was used with the hierarchical clustering algorithm in protein interaction networks (HC-PIN) and default parameters. CytoCluster is a Cytoscape plugin integrating six clustering algorithms, i.e., identifying overlapping and hierarchical modules in protein interaction networks (OH-PIN), identifying protein complex algorithm (IPCA), clustering with overlapping neighborhood expansion (ClusterONE), detecting complexes based on an uncertain graph model (DCU), identifying protein complexes based on maximal complex extension (IPC-MCE), and the Biological Networks Gene Ontology (BinGO) function. CytoCluster is a very popular algorithm used to identify functional modules, predict protein complexes and network biomarkers, and visualize clustering results.

Survival Analysis

The clinical data of all the UCEC and normal samples were obtained from TCGA, and the survival time was extracted using a customized Perl script. For each module, the samples were clustered into two different groups via k-means clustering based on the expression across the RNAs, followed by the comparison of the survival durations between the two groups using a log-rank test. Finally, an R package survival was used to conduct the statistical test.

RESULTS

Dysregulated RNAs Can be Used to Distinguish UCEC Samples From Normal Ones

The expression profiles of 570 samples for miRNAs, lncRNAs, and mRNAs were downloaded from TCGA, which include 537 UCEC samples and 33 counterpart normal samples (**Supplementary Table 1**). To investigate the underlying molecular mechanism on how UCEC occurs and develops, a differential expression analysis was performed for each expression profile using a t-test with a p-value < 0.05 and fold change > 2 as the cutoff (see section "Materials and Methods"). A total of 5831 differentially expressed mRNAs between the UCEC and normal samples were identified, which include 2810 upregulated and 3021 downregulated genes (**Supplementary Table 2**). Moreover, 648 differentially expressed lncRNAs were identified, including 219 upregulated and 428 downregulated lncRNAs (**Supplementary Table 2**). We also identified 342 differentially expressed miRNAs, in which 280 were upregulated and 62 were downregulated (**Supplementary Table 2**).

To further investigate the differentially expressed mRNAs, lncRNAs, and miRNAs between the UCEC and their counterpart normal samples, an unsupervised hierarchical clustering analysis was performed using the R package pheatmap. Each molecule can clearly distinguish UCEC samples from their counterpart normal samples (**Figures 1A–C**). Furthermore, PCA was conducted for the differentially expressed lncRNAs, mRNAs, and miRNAs using the R function prcomp. Again, the majority of the UCEC samples and their counterpart normal samples were separated into two groups (**Figures 1D–F**).

The known tumor suppressor lncRNA HAND2 Antisense RNA 1 (HAND2-AS1) was identified as one of the differentially expressed lncRNAs in high-grade serous ovarian carcinoma (Yang et al., 2018). The significant downregulation in UCEC indicated the potential role as a tumor suppressor in UCEC (**Figure 2A**). Another lncRNA example is FRMD6 Antisense RNA 2 (FRMD6-AS2), which is also downregulated in UCEC (**Figure 2B**). Wang et al. reported the tumor suppressive effect of this lncRNA in UCEC, whose expression is consistent here (Wang et al., 2020). For the protein-coding gene, Homeobox protein Hox-A11 (HOXA11) was significantly downregulated in UCEC (**Figure 2C**) and was reported to play roles in malignant cancer (Zhang et al., 2018). WT1 was also downregulated in UCEC (**Figure 2D**), which was reported to be a prognostic marker in advanced serous epithelial ovarian carcinoma (Netinatsunthorn et al., 2006). MicroRNA-21 (miR-21), which was upregulated in UCEC (**Figure 2E**), is also a cancer biomarker (Bautista-Sánchez et al., 2020). The suppression role for the proliferation and metastasis of miR-522 in non-small cell lung cancer was reported by Zhang et al. (2016), in which miR-522 was upregulated

[2] http://www.gsea-msigdb.org/gsea/msigdb

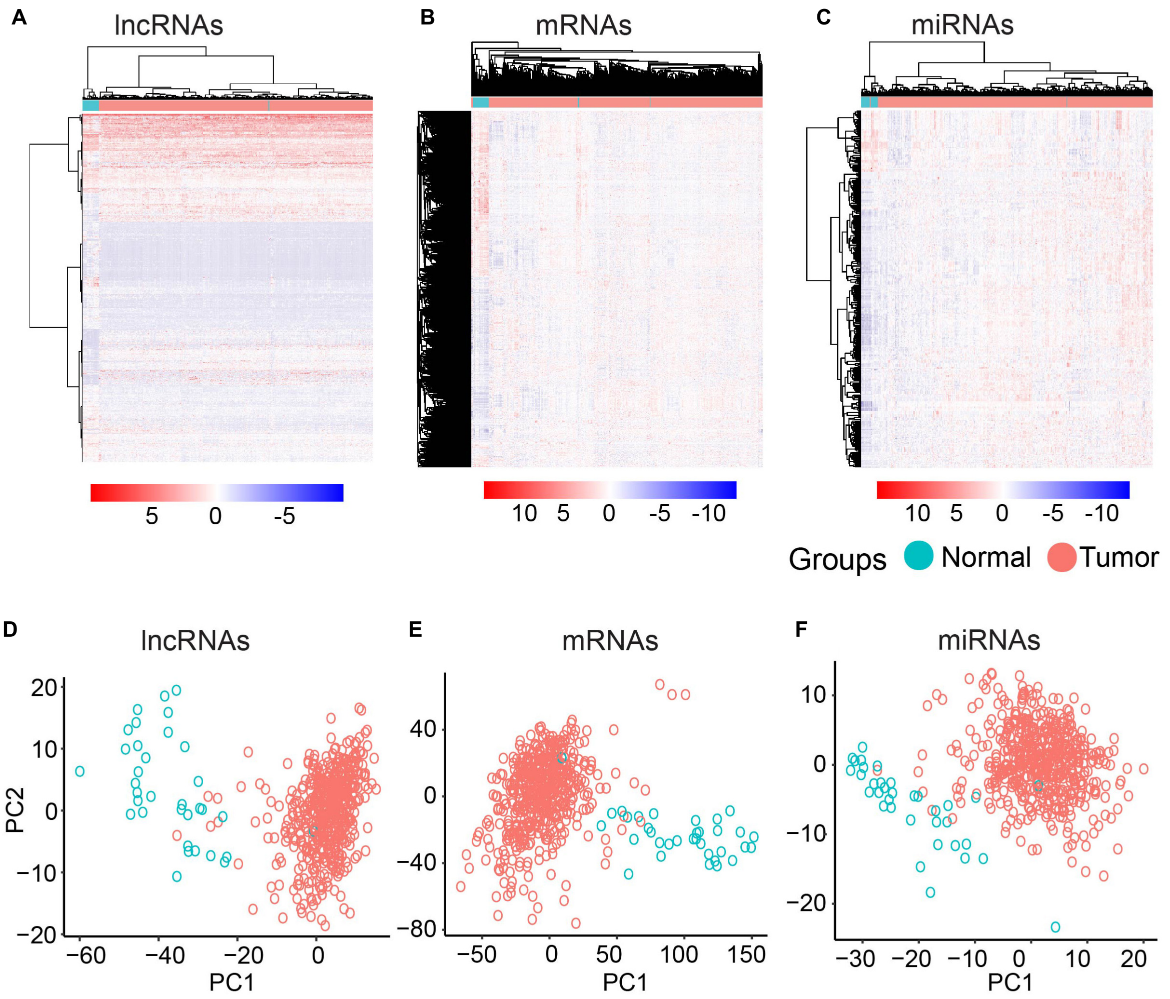


FIGURE 1 | Clustering based on differentially expressed molecules. Heatmap of clustering for UCEC and the normal samples based on differentially expressed lncRNAs **(A)** mRNAs **(B)** and miRNAs **(C)**. PCA analysis for differentially expressed lncRNAs **(D)** mRNAs **(E)** and miRNAs **(F)**.

in UCEC (**Figure 2F**). All these data indicate the potential functional roles of these key RNA molecules.

Dysregulated Genes Are Highly Enriched in Cancer- and Metabolism-Related Pathways

As we mentioned above, genes playing an important function in tumor generation and development were identified to be up- or downregulated in UCEC. To determine the functional roles for all differentially expressed mRNAs, an unbiased functional enrichment analysis for GO using clusterProfiler (Yu et al., 2012) was performed. Cancer hallmark-related terms were enriched (**Figure 3A**). Apoptotic processes, such as "dendritic cell apoptotic process," and cell proliferation-related pathways, such as "mesenchymal cell proliferation" and "regulation of mesenchymal cell proliferation," were enriched. Moreover, immunity-related terms were enriched, such as "establishment of T-cell polarity."

A functional enrichment analysis for KEGG was also performed by the UCEC-related genes (**Figure 3B**). Phosphatidylinositol-4,5-bisphosphate 3-kinase (PI3K)/protein kinase B (Akt) pathway, which is associated with cellular quiescence, proliferation, cancer, and longevity, is an intracellular signaling pathway of great importance in the cell cycle process. It was enriched by UCEC-related genes. The pathway "proteoglycans in cancer" was also enriched, which suggested the functional roles of differentially expressed mRNAs in cancer.

To further investigate the roles of these UCEC-related genes, GSEA was performed using clusterProfiler (Yu et al., 2012; **Figures 3C–F**). The glycolytic pathway, whose high level in tumors, including UCEC, exhibits specific driver genes in most cancer types (Wei et al., 2020), was enriched by upregulated genes in UCEC (**Figure 3D**). Upregulated genes in UCEC were also enriched in a hypoxia-related pathway (Ruan et al., 2009; **Figure 3E**). Moreover, known tumor-related pathways, i.e., G2M checkpoint (**Figure 3C**) and TNFA (**Figure 3F**)

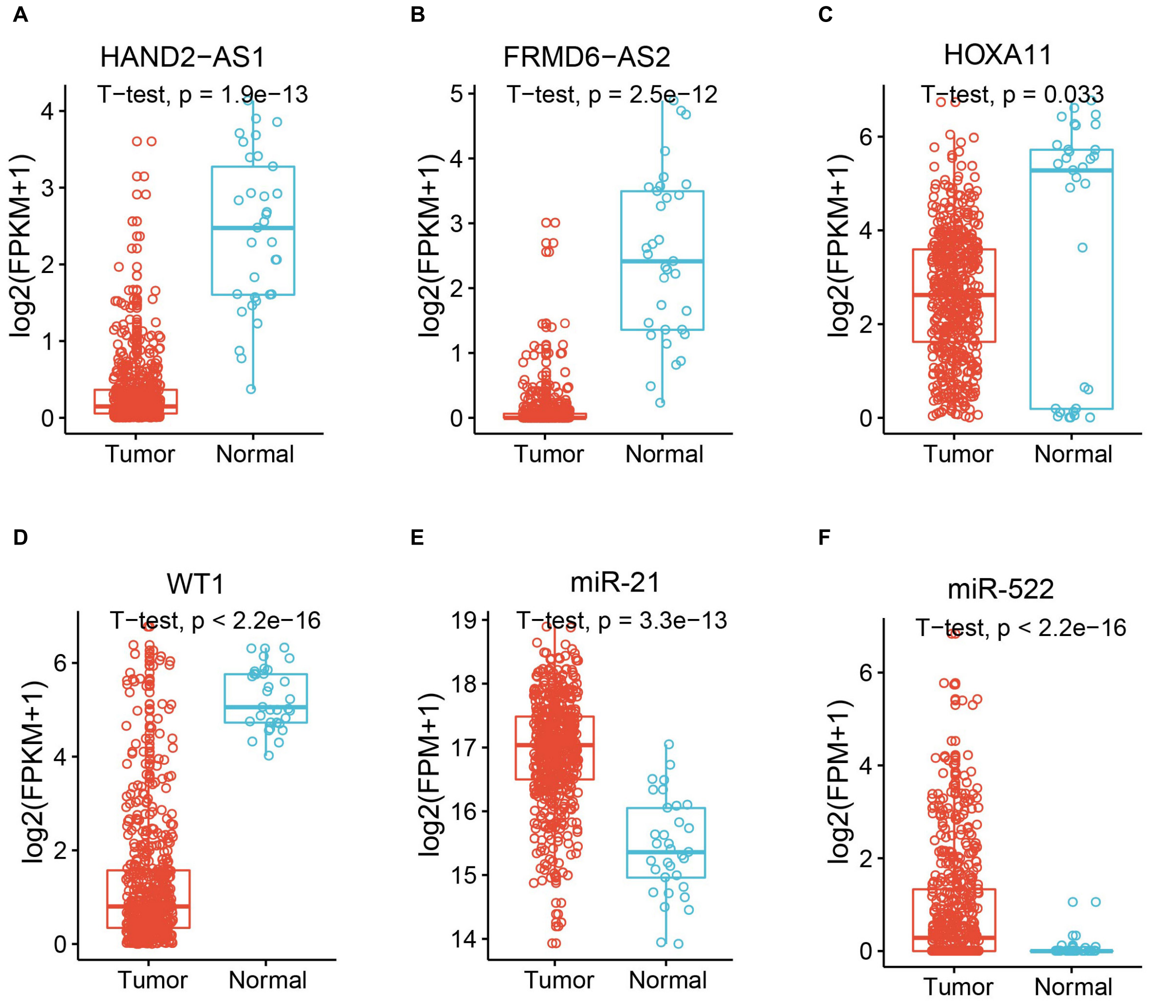


FIGURE 2 | Expression of example molecules in UCEC and the normal samples. The comparison of gene expression between tumor sample and the normal sample for differentially expressed lncRNAs HAND2-AS1 **(A)** and FRMD6-AS2 **(B)**, differentially expressed genes HOXA11 **(C)** and WT1 **(D)**, and differentially expressed miRNAs miR-21 **(E)** and miR-522 **(F)**.

related terms, were enriched by up- and downregulated genes, respectively.

Dysregulated ncRNAs Reveal Immune and Metabolic Functions

NcRNAs have previously been regarded as useless for a long time. However, recently, more studies have attempted to explore the function of ncRNAs (Jiang et al., 2019) and showed functional ncRNAs in tumors (Dong et al., 2020). To determine the functional roles of differentially expressed lncRNAs in UCEC, GO and KEGG analyses were performed (**Figures 4A,B**). For the GO analysis, immunity-related terms, such as "neutrophil-mediated immunity," "neutrophil degranulation," "myeloid leukocyte-mediated immunity," "leukocyte degranulation," "myeloid leukocyte activation" and "interleukin-1-mediated signaling pathway" were enriched (**Figure 4A**). For the KEGG analysis, metabolic pathways, such as "central carbon metabolism in cancer," "glycolysis/gluconeogenesis," "glucagon signaling pathway," "oxidative phosphorylation," and "thermogenesis" were enriched by these lncRNAs (**Figure 4B**).

In addition, to further identify the roles of these lncRNAs, GSEA was also performed (**Figures 4C–F**). Metabolic features, such as "TCA cycle," "Hallmark reactive oxygen species pathway," and "myeloid-derived suppressor cell" were enriched (**Figures 4C–E**). The immunity-related feature "T-cell memory (Tcm) CD8" was also enriched (**Figure 4F**). Interestingly, all these features were enriched by downregulated lncRNAs in UCEC, suggesting the immune and metabolic functional roles of these downregulated lncRNAs.

Besides lncRNAs, miRNAs were also reported to play essential roles in tumor development (Qiu et al., 2020). Thus, to determine the functional role of differentially expressed miRNAs, the same analyses performed on lncRNAs were performed for miRNAs. Again, metabolism and immunity-related GO terms and KEGG pathways were enriched (**Figures 5A,B**). Metabolic GO terms,

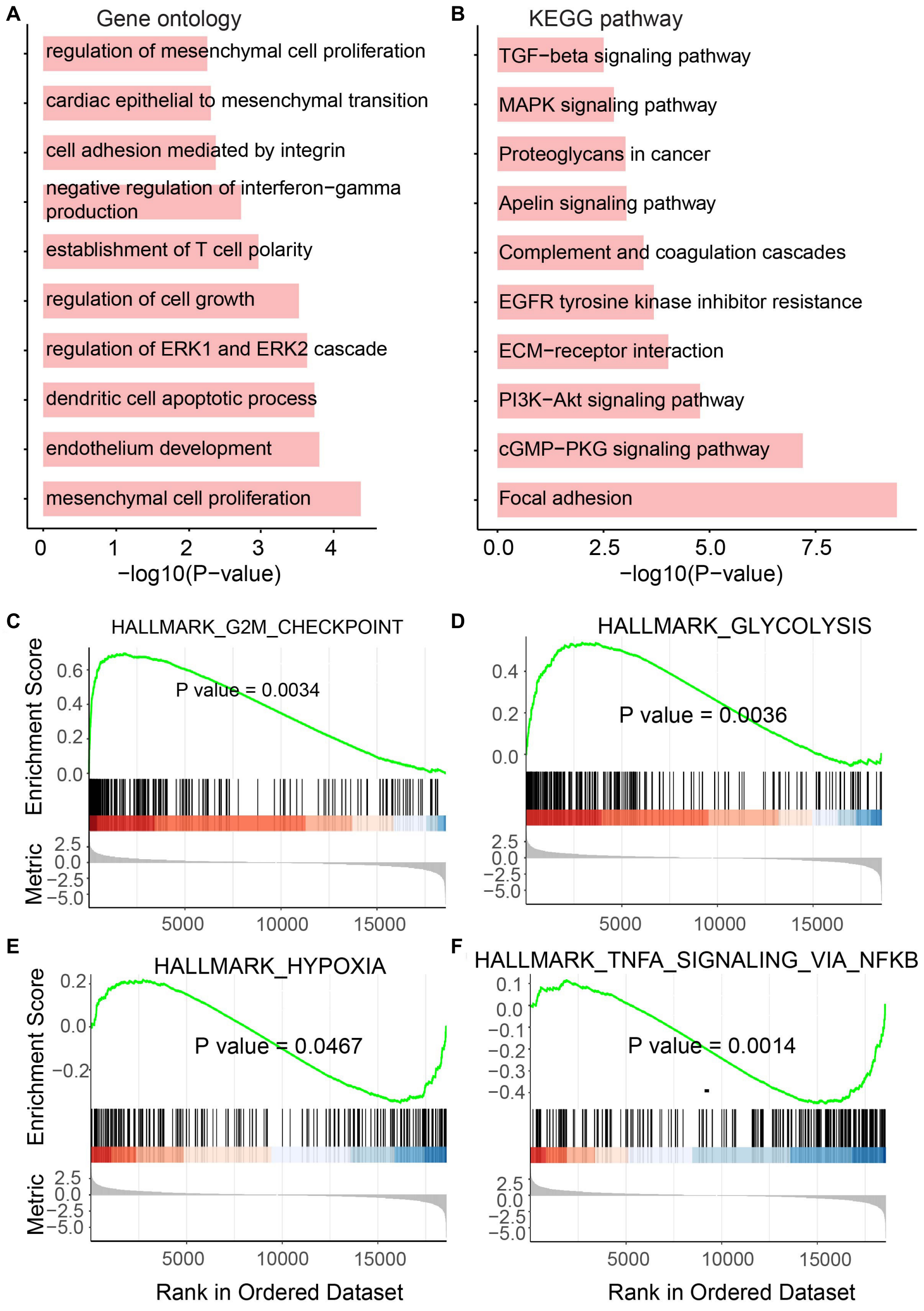


FIGURE 3 | Functional enrichment analysis for differentially expressed mRNAs. **(A)** Enriched GO terms. **(B)** Enriched KEGG pathways. **(C–F)** Results of GSEA analysis.

such as "positive regulation of MAPK cascade" and "regulation of ERK1 and ERK2 cascade," and immunity-related terms, such as "leukocyte activation involved in immune response," "myeloid cell activation involved in immune response" and "neutrophil-mediated immunity" were enriched. Similarly, GSEA also showed the enrichment of pathways involving in cancer and metabolic diseases (**Figures 5C–F**). The DNA repair pathway, which has been reported to be the target for cancer therapies (Helleday et al., 2008) and plays roles in metabolic diseases (Hoeijmakers, 2009), was enriched by upregulated miRNAs in UCEC (**Figure 5C**). The E2F pathway was also enriched by upregulated miRNAs in UCEC (**Figure 5D**). E2F plays a key role in tumor suppression through a specific regulation of tumor suppressor genes (Kurayoshi et al., 2018). Furthermore, estrogen-related and G2M pathways were enriched by downregulated and upregulated miRNAs in UCEC, respectively (**Figures 5E,F**).

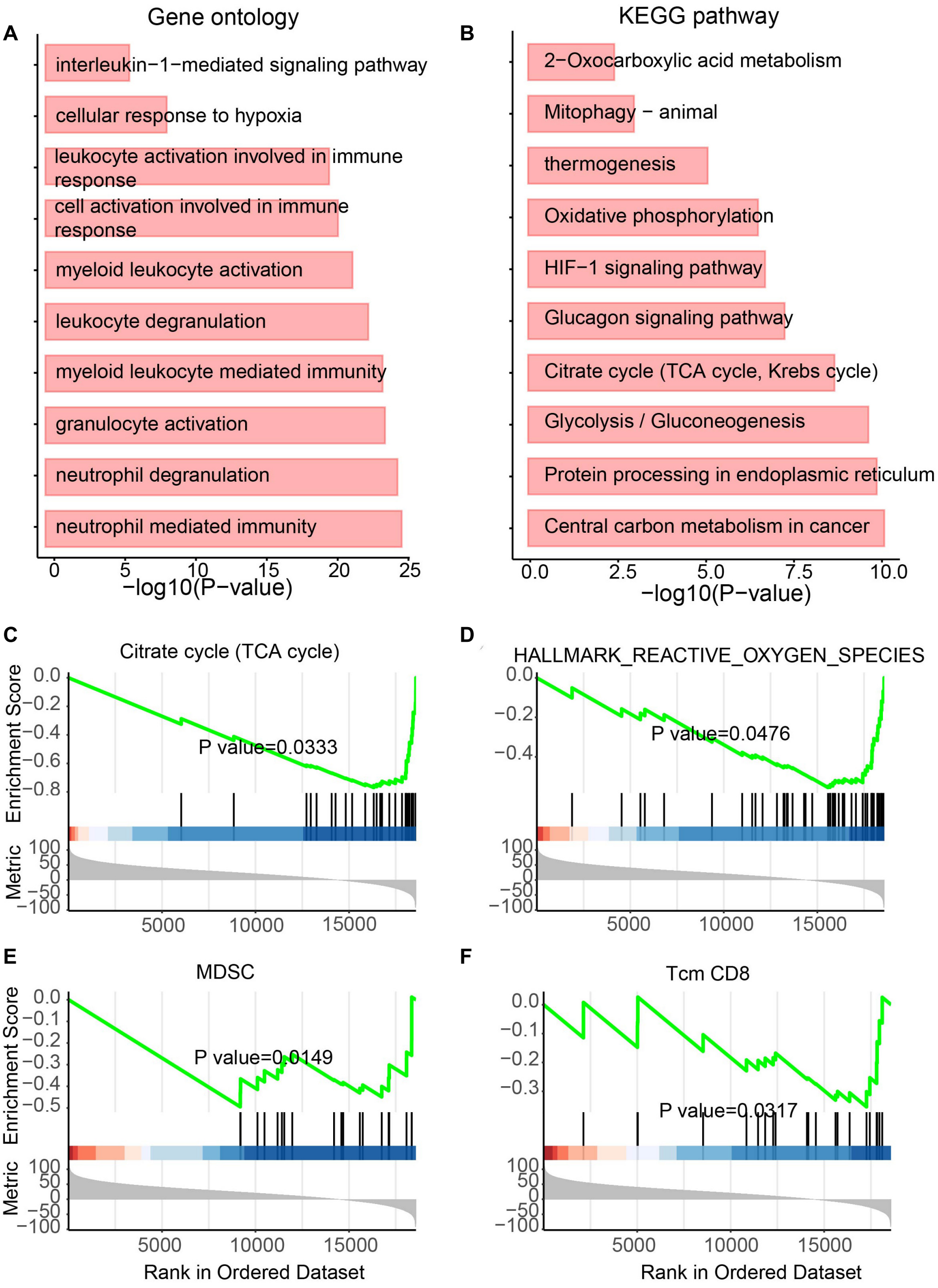


FIGURE 4 | Functional enrichment analysis for differentially expressed lncRNAs. **(A)** Enriched GO terms. **(B)** Enriched KEGG pathways. **(C–F)** Results of GSEA analysis.

Estrogens show function in controlling the energy balance and glucose homeostasis (Mauvais-Jarvis et al., 2013) and play roles in different cancer types (Whiteside, 2008).

Construction of the Dysregulated lncRNA–miRNA–mRNA Network

Based on the interactions between miRNA and its targets downloaded from starBase v3.0 (Li et al., 2014), a dysregulated network containing differentially expressed lncRNAs, miRNAs, and mRNAs was constructed. To provide more confident interactions between miRNA and its targets, AGO CLIP-Seq was used, followed by several filtering steps (see section "Materials and Methods"). A final dysregulated lncRNA–miRNA–mRNA network was constructed with 1243 interactions and 428 differentially expressed molecules, including 323 mRNAs, 53 miRNAs, and 53 lncRNAs (**Figure 6A** and **Supplementary Table 3**).

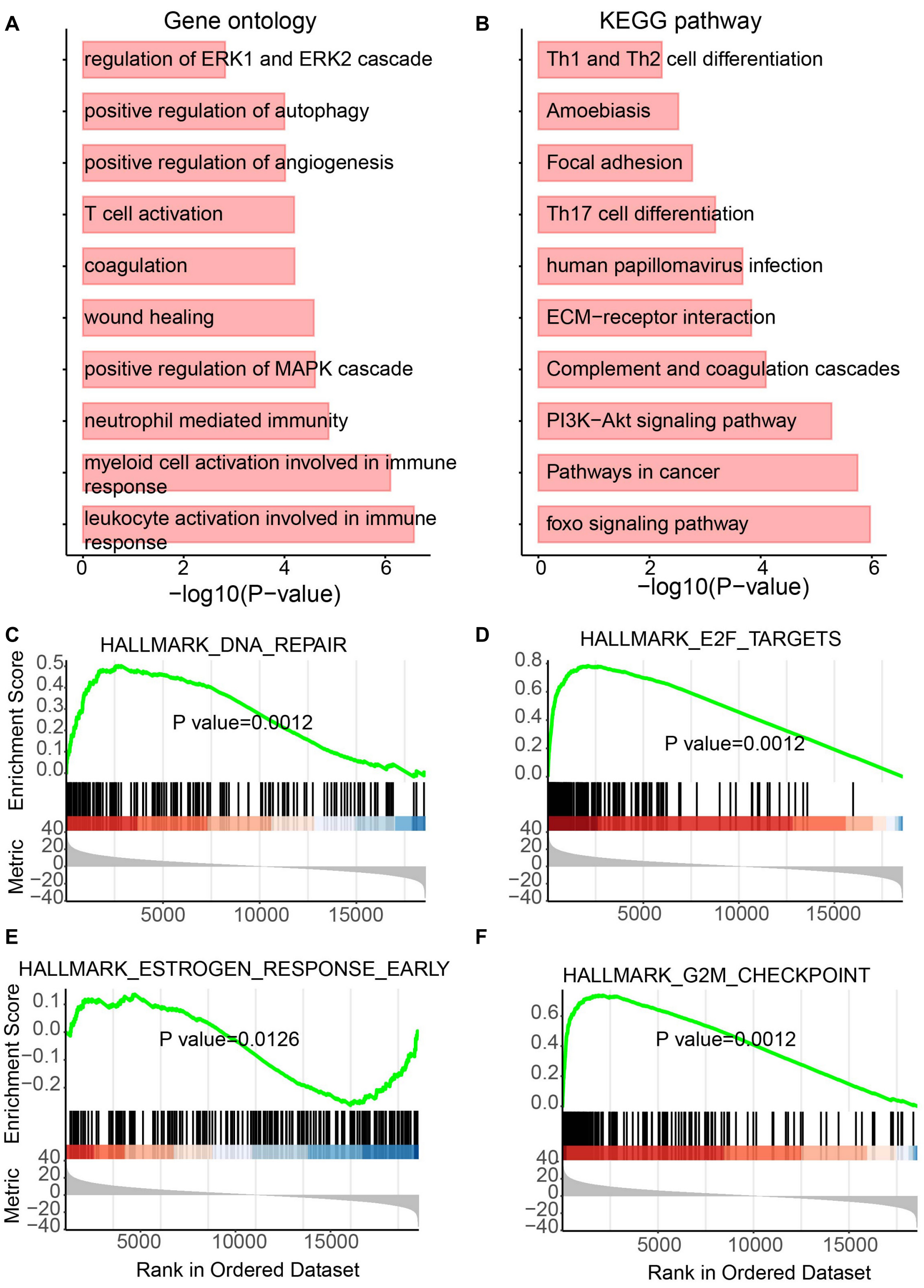


FIGURE 5 | Functional enrichment analysis for differentially expressed miRNAs. **(A)** Enriched GO terms. **(B)** Enriched KEGG pathways. **(C–F)** Results of GSEA analysis.

A biological network is a small-world network (Latora and Marchiori, 2001) or scale-free network (Latora and Marchiori, 2001). To test whether our dysregulated network is a scale-free network, the degree distribution was analyzed (**Supplementary Figure 1**). Approximately 90% of RNAs have less than five edges, whereas only approximately 5% of RNAs have more than 10 interactions. The data support that our dysregulated network is a scale-free network and a meaningful biological network. To further investigate the network, a GO analysis was performed. Cancer hallmark-related functions were enriched, such as the migration-related term "epithelial cell migration" and proliferation-related term "regulation of epithelial cell proliferation" (**Figure 6B**). Moreover, pathways involved in the metabolism were enriched (**Figure 6B**). The Wnt signaling pathway has been shown to direct glycolysis and angiogenesis in colon cancer (Pate et al., 2014). In

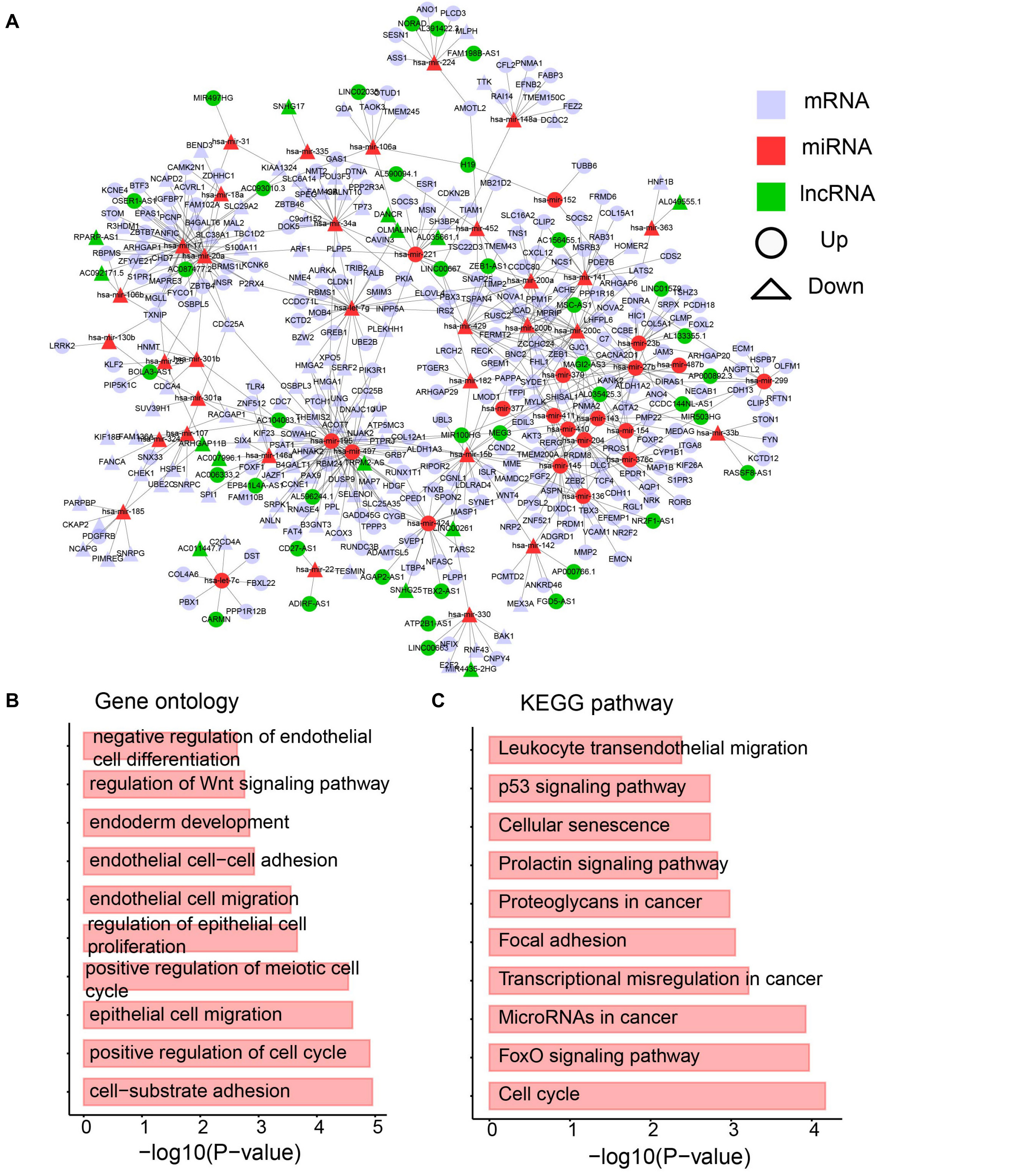


FIGURE 6 | The dysregulated lncRNA-miRNA-mRNA network. **(A)** The network containing differentially expression mRNA, lncRNA and miRNA. **(B)** Enriched GO terms. **(C)** Enriched KEGG pathways.

addition, the KEGG pathway analysis was performed. Pathways playing function in cancers, such as "proteoglycans in cancer," "microRNAs in cancer" and "transcriptional misregulation in cancer" were enriched by the differentially expressed RNAs in the dysregulated network (**Figure 6C**). The FoxO pathway was also enriched (**Figure 6C**), which was reported to be a therapeutic target in cancers (Farhan et al., 2017) and regulate glucose and lipid metabolism (Lee and Dong, 2017). All these data imply the immune and metabolic functions of our dysregulated network.

The Dysregulated Networks Showed Clinical-Related Modules

To maximize the utility of the dysregulated lncRNA–miRNA–mRNA network, the Cytoscape plugin CytoCluster

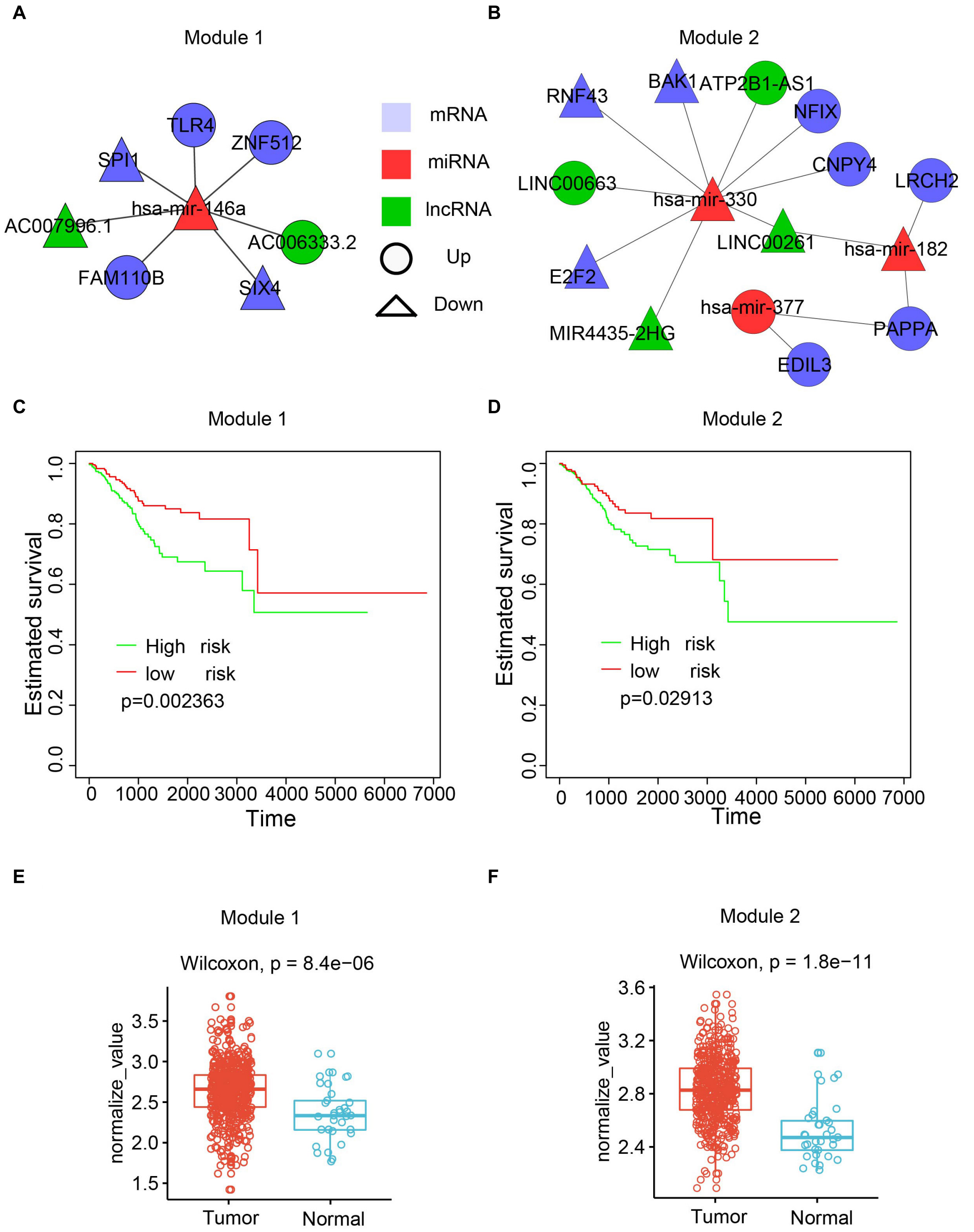


FIGURE 7 | Functional modules identified from the dysregulated network. **(A)** The first module. **(B)** The second module. **(C)** Kaplan-Meier plot of survival for the first module. **(D)** Kaplan-Meier plot of survival for the second module. **(E)** Expression patterns of the first modules in normal and cancer samples. The average expression value of each molecule crossing all normal/cancer samples was used. **(F)** Expression patterns of the second modules in normal and cancer samples.

(Li et al., 2017) was used to identify functional modules from the dysregulated network. CytoCluster is a popular tool used to identify functional modules by integrating seven clustering algorithms, namely, HC-PIN (Wang et al., 2011), OH-PIN (Wang et al., 2012), IPCA (Li et al., 2008), ClusterONE (Nepusz et al., 2012), DCU (Zhao et al., 2014), IPC-MCE (Li et al., 2010), and BinGO function. Accordingly, two modules were identified (**Figures 7A,B**). The first module contained 7 interactions with 5 mRNAs, 2 lncRNAs, and 1 miRNA. The second one consisted of 14 interactions with 8 mRNAs, 4 lncRNAs, and 3 miRNAs.

To explore the biological function of the two modules, the associations of the modules with the patient survival time were evaluated by checking the difference of the survival time between two subpopulations from all UCEC patients divided

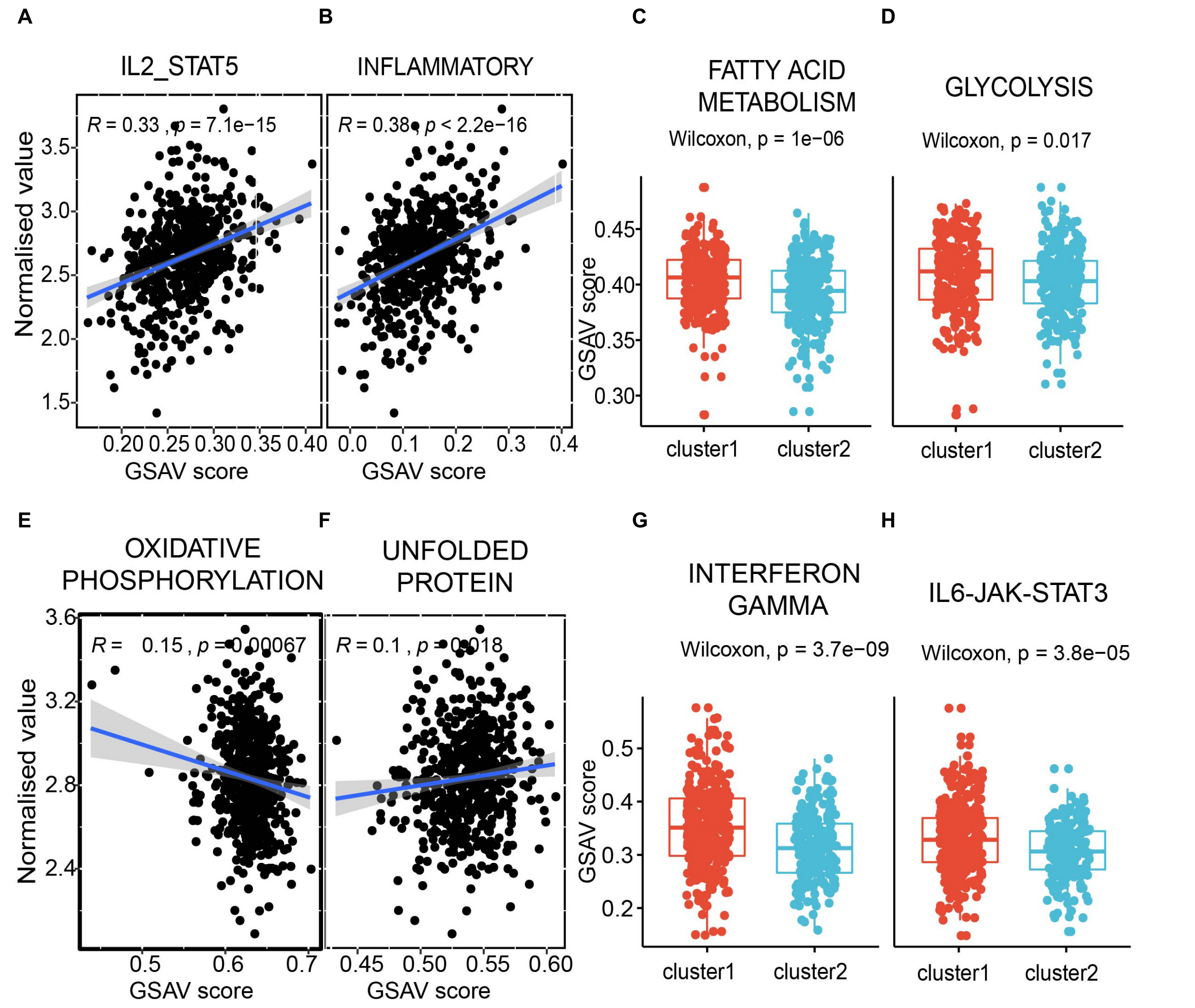


FIGURE 8 | The modules correlated with immunity and metabolism. **(A,B)** the correlation between the first module and pathway IL2-STAT5 **(A)**, and inflammatory **(B)**. Each dot presents a sample. X axis presents the GSVA score, and Y axis presents the normalized expression value. The average expression value of the first module for each sample was used. **(C,D)** the comparison between two subpopulations from **Figure 7C** in fatty acid metabolism pathway **(C)** and glycolysis pathway **(D)**. Y axis presents the GSVA score. **(E,F)** the correlation between the second module and pathway oxidative phosphorylation **(E)**, and unfolded protein **(F)**. **(G,H)** the comparison between two sub-populations from **Figure 7D** in interferon gamma pathway **(G)** and IL6-JAK-STAT3 pathway **(D)**.

by the k-means clustering. Both modules showed a significant correlation with the survival time (**Figures 7C,D**). Next, the Wilcoxon rank-sum test was performed based on the expression values of RNAs between the tumor and normal samples. The results showed that both modules had higher expression in the UCEC samples compared with their counterpart normal samples (**Figures 7E,F**).

The Clinical-Related Modules Are Correlated With Metabolism and Immunology

As immunity- and metabolism-related functions were connected to the dysregulated RNAs in the network, we focused on these related pathways. To determine if the dysregulated RNAs in the two modules are correlated with the immune and metabolic functions, GSVA (Hänzelmann et al., 2013) was performed for each sample. GSVA provides increased power to detect subtle pathway activity changes over a sample population in comparison to corresponding methods.

The first module is positively correlated with interleukin-2 and STAT5 pathway (**Figure 8A**), which was reported to regulate T-cell development and function (Mahmud et al., 2013). A known immune inflammatory pathway was also positively correlated in the first module (**Figure 8B**). Furthermore, two classical metabolic pathways, i.e., fatty acid metabolism pathway and glycolysis pathway, showed significantly different GSVA scores between the two subpopulations with different survival times in the module shown in **Figures 7C**, **8C,D**. The same analyses were also performed to the second module. Oxidative

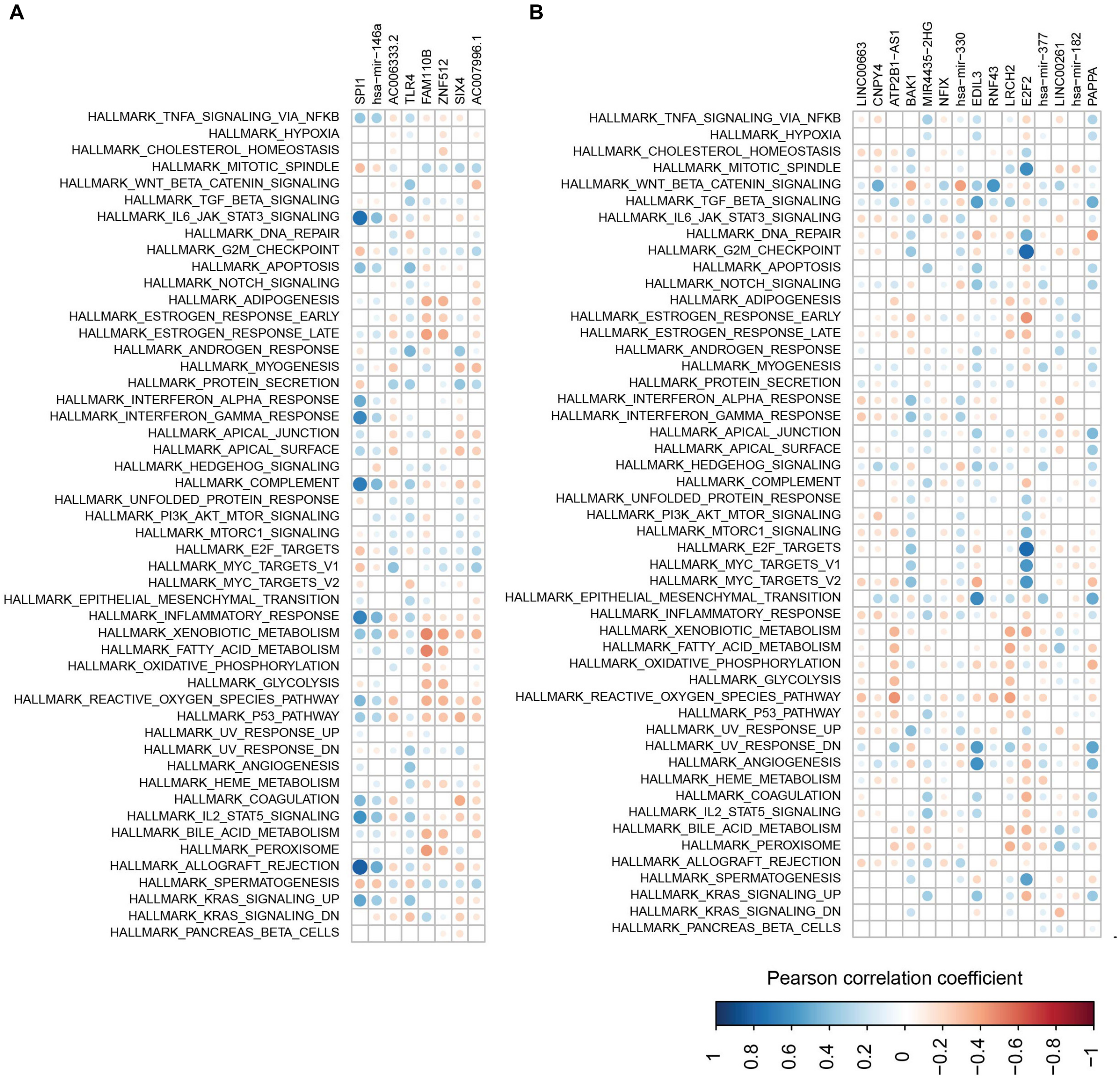


FIGURE 9 | The correlations between RNAs and pathways. **(A,B)** Correlations between pathway and RNAs from the first module **(A)** and the second module **(B)**.

phosphorylation, a classic metabolic pathway, showed a negative correlation with the second module (**Figure 8E**). The unfolded protein pathway, which showed functional roles in different cancer types (McGrath et al., 2018) and metabolic pathways (Lee and Ozcan, 2014), was positively correlated with the second module (**Figure 8F**). The interferon gamma pathway, which affects tumor progression and regression in different cancers (Jorgovanovic et al., 2020) and also metabolic signalings (Siska and Rathmell, 2016), showed significantly different GSVA scores between the two subpopulations with different survival times in the second module shown in **Figures 7D**, **8G**. A similar scenario occurred in the IL6/JAK/STAT3 pathway, a well-known pathway playing a significant role in cancers (Johnson et al., 2018; **Figure 8H**).

To further check the function of the two modules, the correlation between each RNA in the modules and the pathways involved in the immune and metabolic functions was examined (**Figures 9A,B**). Overall, SP11, miR-146a, AC006333.2, and TLR4 from the first module showed a negative correlation with the metabolic and immune functions (**Figure 9A**). Conversely, the other four RNAs in the first module more likely have a positive correlation with the metabolic and immune functions. In the second module, several RNAs, especially for E2F2, showed a negative correlation with the metabolic and immune functions (**Figure 9B**). E2F2 was highly negatively correlated with pathways involved in G2M checkpoints, E2F targets, and mitotic spindles.

DISCUSSION

In this study, a dysregulated lncRNA–miRNA–mRNA network was constructed, in which all RNAs were differentially expressed in UCEC and enriched in cancer and metabolic functions. An integrative analysis on transcriptome data from 570 samples was performed at three different RNA levels, i.e., lncRNAs, miRNAs, and mRNAs. Further analysis identified two clinical-related modules, which showed correlation with metabolic and immune functions. Importantly, some elements from the two modules have been functionally related with UCEC. This framework will help reveal the underlying mechanism for the generation and development of UCEC.

NcRNAs, which constitute more than 90% of RNAs made from the human genome, have attracted increasing attention as more ncRNAs have been functionally validated in different conditions, particularly in human diseases, such as cancers (Anastasiadou et al., 2017; Slack and Chinnaiyan, 2019). In this study, to better determine the potential roles of ncRNAs in UCEC, we focused on dysregulated lncRNAs and miRNAs. By taking advantage of state-of-the-art technologies, we integrated dysregulated lncRNAs, miRNAs, and mRNAs into a single dysregulated network, which is a scale-free and biologically meaningful network. Based on the dysregulated lncRNA–miRNA–mRNA network, a functional enrichment analysis for GO and KEGG was performed, and the results showed that metabolic and immune functions that the network may be involved in were enriched.

Further analysis identified two modules including dysregulated lncRNAs, miRNAs, and mRNAs using a Cytoscape plugin CytoCluster. By integrating the corresponding clinical data, we found that the two modules were survival time related, and both modules were overexpressed in the UCEC samples, indicating the potential carcinogenic roles of some overexpressed elements in the two modules. Through GSVA, we further showed that both modules were immunity and metabolism related. Nevertheless, the biggest limitation is that all the conclusions were drawn without any experiments to support them. Although some elements in the two modules have been functionally validated in UCEC, there are genes (i.e., TLR4, FAM110B, LINC00663, and LINC00261) in the two modules that have not been reported in UCEC, and further experimental and clinical validations are necessary for these RNAs with potential functional roles in UCEC. In the future, we would select one of the genes for further investigation. The counterpart functional experiments such as knockdown and overexpression assays to investigate the mechanism on how the gene paly function in UCEC would be performed. Our study provides new insights into the outcome prediction and will help in the precision medicine for UCEC.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The patient data used in this study was acquired as publicly available datasets that were collected with patients' informed consent.

AUTHOR CONTRIBUTIONS

MQ conceived the project. MQ and DL collected the data and reviewed the manuscript. DL performed analysis and wrote the manuscript. Both authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.673192/full#supplementary-material

REFERENCES

Anastasiadou, E., Jacob, L. S., and Slack, F. J. (2017). Non-coding RNA networks in cancer. *Nat. Rev. Cancer* 18, 5–18. doi: 10.1038/nrc.2017.99

Bautista-Sánchez, D., Arriaga-Canon, C., Pedroza-Torres, A., De La Rosa-Velázquez, I. A., González-Barrios, R., Contreras-Espinosa, L., et al. (2020). The Promising Role of miR-21 as a Cancer Biomarker and Its Importance in RNA-Based Therapeutics. *Mol. Ther. Nucleic Acids* 20, 409–420. doi: 10.1016/j.omtn.2020.03.003

Chekulaeva, M., and Filipowicz, W. (2009). Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. *Curr. Opin. Cell Biol.* 21, 452–460. doi: 10.1016/j.ceb.2009.04.009

Chen, Y., Huang, Q., Chen, Q., Lin, Y., Sun, X., Zhang, H., et al. (2015). The inflammation and estrogen metabolism impacts of polychlorinated biphenyls on endometrial cancer cells. *Toxicol. In Vitro* 29, 308–313. doi: 10.1016/j.tiv.2014.11.008

Dong, Y., Xiao, Y., Shi, Q., and Jiang, C. (2020). Dysregulated lncRNA-miRNA-mRNA Network Reveals Patient Survival-Associated Modules and RNA Binding Proteins in Invasive Breast Carcinoma. *Front. Genet.* 10:1284. doi: 10.3389/fgene.2019.01284

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biol.* 5:R1. doi: 10.1186/gb-2003-5-1-r1

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

Farhan, M., Wang, H., Gaur, U., Little, P. J., Xu, J., and Zheng, W. (2017). FOXO signaling pathways as therapeutic targets in cancer. *Int. J. Biol. Sci.* 13, 815–827. doi: 10.7150/ijbs.20052

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076. doi: 10.1038/nature08975

Gutschner, T., and Diederichs, S. (2012). The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.* 9, 703–719. doi: 10.4161/rna.20481

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

Helleday, T., Petermann, E., Lundin, C., Hodgson, B., and Sharma, R. A. (2008). DNA repair pathways as targets for cancer therapy. *Nat. Rev. Cancer* 8, 193–204. doi: 10.1038/nrc2342

Hoeijmakers, J. H. J. (2009). DNA Damage, Aging, and Cancer. *N. Engl. J. Med.* 361, 1475–1485. doi: 10.1056/nejmra0804615

Hyashizaki, Y. (2004). Neutral evolution of 'non-coding' complementary DNAs (reply). *Nature* 431, 2–3. doi: 10.1038/nature03017

Jiang, C., Ding, N., Li, J., Jin, X., Li, L., Pan, T., et al. (2019). Landscape of the long non-coding RNA transcriptome in human heart. *Brief. Bioinform.* 20, 1812–1825. doi: 10.1093/bib/bby052

Johnson, D. E., O'Keefe, R. A., and Grandis, J. R. (2018). Targeting the IL-6/JAK/STAT3 signalling axis in cancer. *Nat. Rev. Clin. Oncol.* 15, 234–248. doi: 10.1038/nrclinonc.2018.8

Jorgovanovic, D., Song, M., Wang, L., and Zhang, Y. (2020). Roles of IFN-γin tumor progression and regression: a review. *Biomark. Res.* 8:49. doi: 10.1186/s40364-020-00228-x

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat. Genet.* 39, 1278–1284. doi: 10.1038/ng2135

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., et al. (2011). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res.* 71, 6320–6326. doi: 10.1158/0008-5472.CAN-11-1021

König, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2012). Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.* 13, 77–83. doi: 10.1038/nrg3141

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., et al. (2005). Combinatorial microRNA target predictions. *Nat. Genet.* 37, 495–500. doi: 10.1038/ng1536

Kurayoshi, K., Ozono, E., Iwanaga, R., Bradford, A. P., Komori, H., Araki, K., et al. (2018). "The Key Role of E2F in Tumor Suppression through Specific Regulation of Tumor Suppressor Genes in Response to Oncogenic Changes," in *Gene Expression and Regulation in Mammalian Cells - Transcription Toward the Establishment of Novel Therapeutics*, (ed) A. Sebata (London: IntechOpen). doi: 10.5772/intechopen.72125

Latora, V., and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.* 87:198701. doi: 10.1103/PhysRevLett.87.198701

Lee, J., and Ozcan, U. (2014). Unfolded protein response signaling and metabolic diseases. *J. Biol. Chem.* 289, 1203–1211. doi: 10.1074/jbc.R113.534743

Lee, S., and Dong, H. H. (2017). FoxO integration of insulin signaling with glucose and lipid metabolism. *J. Endocrinol.* 233, R67–R79. doi: 10.1530/JOE-17-0002

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, J., Xu, W., and Zhu, Y. (2020). Mammaglobin B may be a prognostic biomarker of uterine corpus endometrial cancer. *Oncol. Lett.* 20:255. doi: 10.3892/ol.2020.12118

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, M., Chen, J. E., Wang, J. X., Hu, B., and Chen, G. (2008). Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics* 9:398. doi: 10.1186/1471-2105-9-398

Li, M., Li, D., Tang, Y., Wu, F., and Wang, J. (2017). Cytocluster: a cytoscape plugin for cluster analysis and visualization of biological networks. *Int. J. Mol. Sci.* 18:1880. doi: 10.3390/ijms18091880

Li, M., Wang, J. X., Liu, B. B., and Chen, J. E. (2010). An algorithm for identifying protein complexes based on maximal clique extension. *J. Cent. South Univ.* 41, 560–565.

Liu, A. N., Qu, H. J., Gong, W. J., Xiang, J. Y., Yang, M. M., and Zhang, W. (2019). LncRNA AWPPH and miRNA-21 regulates cancer cell proliferation and chemosensitivity in triple-negative breast cancer by interacting with each other. *J. Cell. Biochem.* 120, 14860–14866. doi: 10.1002/jcb.28747

Liu, C., Zhang, Y. H., Deng, Q., Li, Y., Huang, T., Zhou, S., et al. (2017). Cancer-Related Triplets of mRNA-lncRNA-miRNA Revealed by Integrative Network in Uterine Corpus Endometrial Carcinoma. *Biomed. Res. Int.* 2017:3859582. doi: 10.1155/2017/3859582

Loher, P., and Rigoutsos, I. (2012). Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics* 28, 3322–3323. doi: 10.1093/bioinformatics/bts615

Mahmud, S. A., Manlove, L. S., and Farrar, M. A. (2013). Interleukin-2 and STAT5 in regulatory T cell development and function. *JAKSTAT* 2:e23154. doi: 10.4161/jkst.23154

Matteson, K. A., Robison, K., and Jacoby, V. L. (2018). Opportunities for early detection of endometrial cancer in women with postmenopausal bleeding. *JAMA Intern. Med.* 178, 1222–1223. doi: 10.1001/jamainternmed.2018.2819

Mauvais-Jarvis, F., Clegg, D. J., and Hevener, A. L. (2013). The role of estrogens in control of energy balance and glucose homeostasis. *Endocr. Rev.* 34, 309–338. doi: 10.1210/er.2012-1055

McGrath, E. P., Logue, S. E., Mnich, K., Deegan, S., Jäger, R., Gorman, A. M., et al. (2018). The unfolded protein response in breast cancer. *Cancers* 10:344. doi: 10.3390/cancers10100344

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* 9, 471–472. doi: 10.1038/nmeth.1938

Netinatsunthorn, W., Hanprasertpong, J., Dechsukhum, C., Leetanaporn, R., and Geater, A. (2006). WT1 gene expression as a prognostic marker in advanced serous epithelial ovarian carcinoma: an immunohistochemical study. *BMC Cancer* 6:90. doi: 10.1186/1471-2407-6-90

Pate, K. T., Stringari, C., Sprowl-Tanio, S., Wang, K., TeSlaa, T., Hoverter, N. P., et al. (2014). Wnt signaling directs a metabolic program of glycolysis and angiogenesis in colon cancer. *EMBO J.* 33, 1454–1473. doi: 10.15252/embj.201488598

Qiu, M., Fu, Q., Jiang, C., and Liu, D. (2020). Machine Learning Based Network Analysis Determined Clinically Relevant miRNAs in Breast Cancer. *Front. Genet.* 11:615864. doi: 10.3389/fgene.2020.615864

Ruan, K., Song, G., and Ouyang, G. (2009). Role of hypoxia in the hallmarks of human cancer. *J. Cell. Biochem.* 107, 1053–1062. doi: 10.1002/jcb.22214

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer Statistics, 2021. *CA. Cancer J. Clin.* 71, 7–33. doi: 10.3322/caac.21654

Siska, P. J., and Rathmell, J. C. (2016). Metabolic Signaling Drives IFN-γ. *Cell Metab.* 24, 651–652. doi: 10.1016/j.cmet.2016.10.018

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

Wang, J., Li, M., Chen, J., and Pan, Y. (2011). A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 607–620. doi: 10.1109/TCBB.2010.75

Wang, J., Li, Z., Wang, X., Ding, Y., and Li, N. (2020). The tumor suppressive effect of long non-coding RNA FRMD6-AS2 in uteri corpus endometrial carcinoma. *Life Sci.* 243:117254. doi: 10.1016/j.lfs.2020.117254

Wang, J., Ren, J., Li, M., and Wu, F. X. (2012). Identification of hierarchical and overlapping functional modules in PPI networks. *IEEE Trans. Nanobioscience* 11, 386–393. doi: 10.1109/TNB.2012.2210907

Wang, Y., Xu, M., and Yang, Q. (2019). A six-microRNA signature predicts survival of patients with uterine corpus endometrial carcinoma. *Curr. Probl. Cancer* 43, 167–176. doi: 10.1016/j.currproblcancer.2018.02.002

Wei, J., Huang, K., Chen, Z., Hu, M., Bai, Y., Lin, S., et al. (2020). Characterization of glycolysis-associated molecules in the tumor microenvironment revealed by pan-cancer tissues and lung cancer single cell data. *Cancers* 12:1788. doi: 10.3390/cancers12071788

Whiteside, T. L. (2008). The tumor microenvironment and its role in promoting tumor growth. *Oncogene* 27, 5904–5912. doi: 10.1038/onc.2008.271

Yang, X., Wang, C. C., Lee, W. Y. W., Trovik, J., Chung, T. K. H., and Kwong, J. (2018). Long non-coding RNA HAND2-AS1 inhibits invasion and metastasis in endometrioid endometrial carcinoma through inactivating neuromedin U. *Cancer Lett.* 413, 23–34. doi: 10.1016/j.canlet.2017.10.028

Yang, Z., Zhou, L., Wu, L. M., Lai, M. C., Xie, H. Y., Zhang, F., et al. (2011). Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Ann. Surg. Oncol.* 18, 1243–1250. doi: 10.1245/s10434-011-1581-y

Ye, Y., Xiang, Y., Ozguc, F. M., Kim, Y., Liu, C. J., Park, P. K., et al. (2018). The Genomic Landscape and Pharmacogenomic Interactions of Clock Genes in Cancer Chronotherapy. *Cell Syst.* 6, 314–328.e2. doi: 10.1016/j.cels.2018.01.013

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, R., Zhang, T. T., Zhai, G. Q., Guo, X. Y., Qin, Y., Gan, T. Q., et al. (2018). Evaluation of the HOXA11 level in patients with lung squamous cancer and insights into potential molecular pathways via bioinformatics analysis. *World J. Surg. Oncol* 16:109. doi: 10.1186/s12957-018-1375-9

Zhang, T., Hu, Y., Ju, J., Hou, L., Li, Z., Xiao, D., et al. (2016). Downregulation of miR-522 suppresses proliferation and metastasis of non-small cell lung cancer cells by directly targeting DENN/MADD domain containing 2D. *Sci. Rep.* 6:19346. doi: 10.1038/srep19346

Zhang, Y., Liu, H., Yang, S., Zhang, J., Qian, L., and Chen, X. (2014). Overweight, obesity and endometrial cancer risk: results from a systematic review and meta-analysis. *Int. J. Biol. Markers* 29, e21–e29. doi: 10.5301/jbm.5000047

Zhao, B., Wang, J., Li, M., Wu, F. X., and Pan, Y. (2014). Detecting protein complexes basedon uncertain graph model. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 486–497. doi: 10.1109/TCBB.2013.2297915

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

The handling editor declared a past co-authorship with the authors MQ and DL.

ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fgene.2021.684349

Identification of a Novel Immune-Related CpG Methylation Signature to Predict Prognosis in Stage II/III Colorectal Cancer

Feng Chen[1†], Lijuan Pei[1†], Siyao Liu[2†], Yan Lin[3], Xinyin Han[4,5], Erhong Meng[2], Xintong Wang[2], Shuai Hong[2], Dongliang Wang[2], Feide Liu[1], Yang Fei[1]* and Guangda Wang[6]**

[1] Department of General Surgery, The Fourth Medical Center of PLA General Hospital, Beijing, China, [2] ChosenMed Technology Co., Ltd., Beijing, China, [3] Library, The Fourth Hospital of Hebei Medical University, Shijiazhuang, China, [4] Computer Network Information Center, Chinese Academy of Sciences, Beijing, China, [5] University of the Chinese Academy of Sciences, Beijing, China, [6] Department of Radiology, The Fourth Hospital of Hebei Medical University, Shijiazhuang, China

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Qian Chen,
Guangxi Medical University Cancer Hospital, China
Jinhui Liu,
Nanjing Medical University, China

***Correspondence:**
Yang Fei
feiyang304py@sina.com
Feide Liu
Feide301@163.com
Guangda Wang
today_5@sina.com

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

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

Received: 23 March 2021
Accepted: 07 June 2021
Published: 28 June 2021

Citation:
Chen F, Pei L, Liu S, Lin Y, Han X, Meng E, Wang X, Hong S, Wang D, Liu F, Fei Y and Wang G (2021) Identification of a Novel Immune-Related CpG Methylation Signature to Predict Prognosis in Stage II/III Colorectal Cancer. Front. Genet. 12:684349. doi: 10.3389/fgene.2021.684349

With the increasing incidence of colorectal cancer (CRC) and continued difficulty in treating it using immunotherapy, there is an urgent need to identify an effective immune-related biomarker associated with the survival and prognosis of patients with this disease. DNA methylation plays an essential role in maintaining cellular function, and changes in methylation patterns may contribute to the development of autoimmunity, aging, and cancer. In this study, we aimed to identify a novel immune-related methylated signature to aid in predicting the prognosis of patients with CRC. We investigated DNA methylation patterns in patients with stage II/III CRC using datasets from The cancer genome atlas (TCGA). Overall, 182 patients were randomly divided into training (n = 127) and test groups (n = 55). In the training group, five immune-related methylated CG sites (cg11621464, cg13565656, cg18976437, cg20505223, and cg20528583) were identified, and CG site-based risk scores were calculated using univariate Cox proportional hazards regression in patients with stage II/III CRC. Multivariate Cox regression analysis indicated that methylated signature was independent of other clinical parameters. The Kaplan–Meier analysis results showed that CG site-based risk scores could significantly help distinguish between high- and low-risk patients in both the training (P = 0.000296) and test groups (P = 0.022). The area under the receiver operating characteristic curve in the training and test groups were estimated to be 0.771 and 0.724, respectively, for prognosis prediction. Finally, stratified analysis results suggested the remarkable prognostic value of CG site-based risk scores in CRC subtypes. We identified five methylated CG sites that could be used as an efficient overall survival (OS)-related biomarker for stage II/III CRC patients.

Keywords: colorectal cancer, CpG methylated sites, biomarker, prognosis, immunotherapy

Abbreviations: CRC, colorectal cancer; MSS, microsatellite stable; TMB, tumor mutational burden; TMB-H, high tumor mutational burden; OS, overall survival; CG score, CG site-based risk score; AUC, area under the curve; MSI, microsatellite instability; DCA, decision curve analysis.

INTRODUCTION

In China, colorectal cancer (CRC) is the fifth most common malignancy, and CRC-related deaths have increased in recent years (Chen, 2015; Fang et al., 2015). Approximately 70% of patients with CRC have stage II/III tumors. At present, the tumor-node-metastasis classification criteria are insufficient to predict prognosis and make clinical decisions, especially in patients with stage II/III CRC (Edge and Compton, 2010). Considerable progress has been made in tumor immunotherapy (immuno-oncology) owing to the enhanced understanding of immune mechanisms. However, the benefit of immunotherapy in patients with CRC is limited, and the advancement in clinical research is relatively lagging (Sun et al., 2016). Programmed cell death protein 1/programmed death-1 ligand 1 antibody inhibitors have been reported to be ineffective in immunotherapy for 85% of patients with microsatellite stable (MSS) CRC (Sillo et al., 2019). In addition, existing biomarkers, including programmed death-1 ligand 1 protein expression, tumor mutational burden (TMB), immune scores, and gamma-interferon signatures, do not effectively predict the prognosis of patients with MSS CRC. Consequently, there is an urgent need to identify immune-related biomarkers for predicting cancer prognosis, which will improve the treatment of CRC.

Aberrant DNA methylation results in the downregulation of various genes and can potentially initiate the pathogenesis of cancer. It is a promising candidate for the development of robust diagnostic, predictive, and prognostic biomarkers for cancer. For instance, hypomethylation of long interspersed nuclear element-1 is correlated with poor survival in CRC patients (Antelo et al., 2012; Rhee et al., 2012). Additionally, long interspersed nuclear element-1 hypomethylation of cell-free DNA is associated with disease progression in CRC (Nagai et al., 2017). Moreover, the hypermethylation level of cyclin-dependent kinase inhibitor 2A predicts recurrence, distant metastasis, and prognosis in patients with CRC (Shen et al., 2007; Kim et al., 2010). Interestingly, cyclin-dependent kinase inhibitor 2A hypermethylation is associated with the poor survival of patients with rectal cancer after surgery and adjuvant 5-fluorouracil chemotherapy (Simpson et al., 1999; Kim et al., 2010). The methylation states of helicase-like transcription factor and hyperplastic polyposis 1 are correlated with tumor aggressiveness, recurrence, and prognosis (Huang et al., 2007). However, only a few studies have focused on identifying immune-related methylated signatures for predicting the prognosis of patients with stage II/III CRC. Therefore, it is necessary to identify prognosis-related methylated biomarkers for this deadly disease.

In this study, we aimed to identify and validate a novel immune-related methylated site-based signature using CRC datasets from the cancer genome atlas (TCGA). Based on our results, we proposed a prognosis-related biomarker that is also effective for patients with CRC subtypes.

MATERIALS AND METHODS

Patients

We downloaded the epigenome-wide DNA CpG site methylation scored as a β-value between 0 and 1 (Illumina 450 K Methylation Beadchip) of stage II/III CRC samples from the Genomic Data Commons data portal[1] (Sanford et al., 2018). Overall, 182 stage II/III CRC samples and 36 paired normal samples were included. The summary of patients is shown in **Table 1**, and the patients were randomly divided into training (*n* = 127) and testing groups (*n* = 55) (**Figure 1**). We obtained the fragments per kilobase of exon per million mapped fragment formats of 182 stage II/III CRC samples in the "HTSeq-FPKM" category, which were further processed, followed by normalized values for gene expression levels (Robinson et al., 2010; Kruppa and Jung, 2017). The "Masked Somatic Mutation" category included four types of mutation data based on diverse processing software, and we selected "MuTect2 Variant" process with 182 stage II/III CRC samples for further mutation analysis. TMB was determined by analyzing the number of somatic mutations per megabase. The cut-off value for high TMB (TMB-H) was determined to be the top 25% of all CRC patients. We obtained the clinical data of the 182 stage II/III CRC samples from TCGA-COAD dataset. The CRC samples gene expression profiles of GSE14333 and GSE103479 were downloaded from GEO databases[2]. The expression data of GSE14333 was based on GPL570 Platforms included 290 primary CRC samples (Submission date: January 08, 2010). The expression data of GSE103479 was based on GPL23985 Platforms included 363 stage II/III CRC samples (Submission date: December 31, 2017).

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

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

TABLE 1 | Summary of patient demographics and characteristics.

Characteristic	Training (*N* = 127)	Test (*N* = 55)
Gender		
Female	58 (45.7%)	26 (47.3%)
Male	69 (54.3%)	29 (52.7%)
Age		
<65 years	50 (39.4%)	28 (50.9%)
≥65 years	77 (60.6%)	27 (49.1%)
Stage		
II	73 (57.5%)	32 (58.2%)
III	54 (42.5%)	23 (41.8%)
Adjuvant chemotherapy		
Adjuvant chemotherapy	49 (38.6%)	25 (45.5%)
None	78 (61.4%)	30 (54.5%)
Vital status		
Living	97 (76.4 %)	44 (80.0%)
Dead	30 (23.6%)	11 (20.0%)

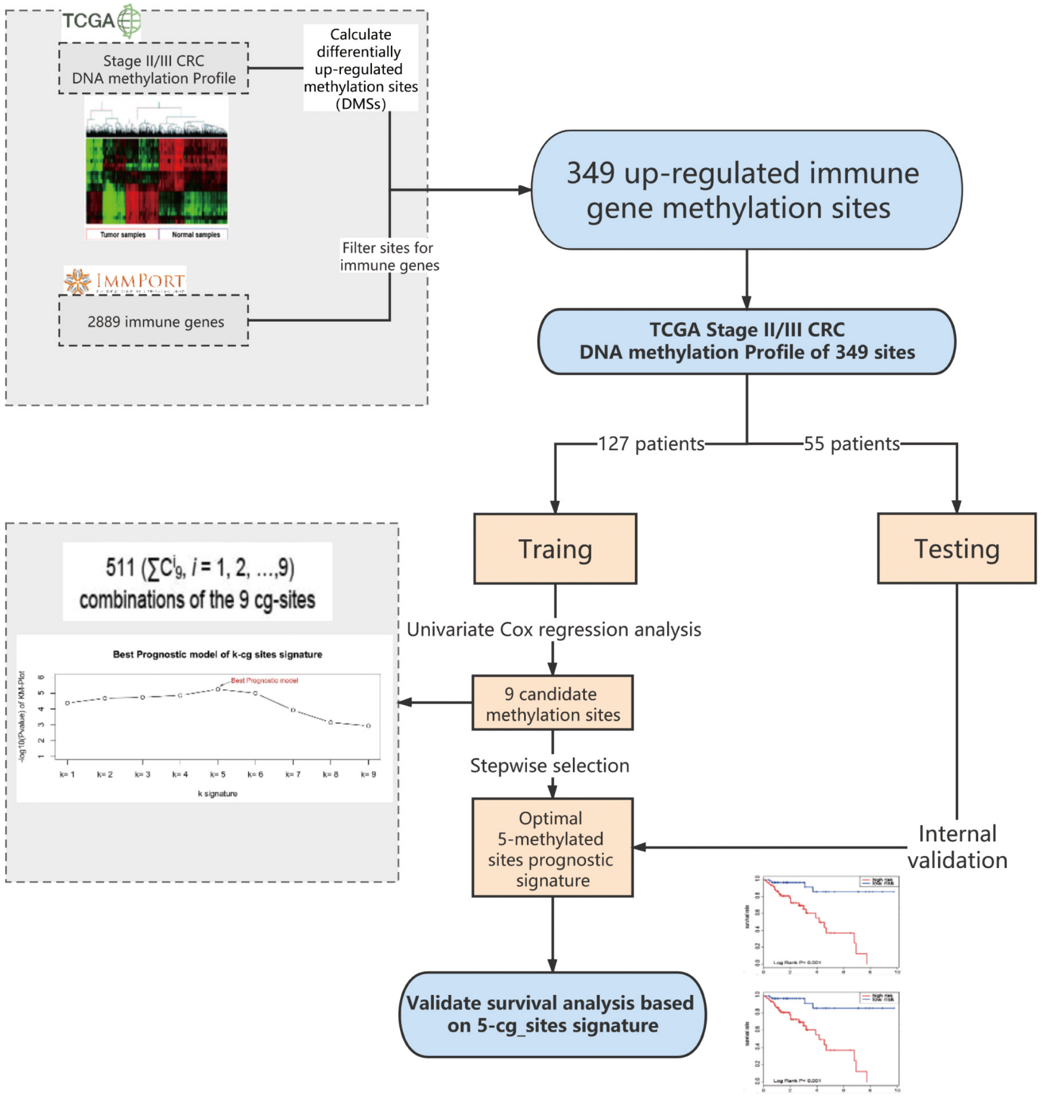

TCGA
Stage II/III CRC
DNA methylation Profile
Calculate differentially up-regulated methylation sites (DMSs)
Tumor samples
Normal samples
IMMPORT
2889 immune genes
Filter sites for immune genes
349 up-regulated immune gene methylation sites
TCGA Stage II/III CRC
DNA methylation Profile of 349 sites
127 patients
55 patients
Traing
Testing
Univariate Cox regression analysis
9 candidate methylation sites
Stepwise selection
Optimal 5-methylated sites prognostic signature
Internal validation
511 (∑Cⁱ₉, i = 1, 2, ...,9)
combinations of the 9 cg-sites
Best Prognostic model of k-cg sites signature
Best Prognostic model
-log10(Pvalue) of KM-Plot
k signature
k= 1
k= 2
k= 3
k= 4
k= 5
k= 6
k= 7
k= 8
k= 9
Validate survival analysis based on 5-cg_sites signature

FIGURE 1 | I Dentification of the methylated signature in the training set. Methylated site profiling in tumor and normal tissue samples. Overall, 349 methylated sites were overlapped between 6450 differentially methylated sites (DMSs) and 2483 immune genes. Correlation between nine methylated sites and the survival of patients with stage II/III CRC in the training group was observed upon performing Univariate Cox regression analysis. Development of a prognostic classifier for all combinations of the nine CG sites using the CG Score. For each combination, patients were classified into high- and low-risk groups based on their median CG Score, and the five-methylated site signature with the largest value of −log(p) was selected as the final signature. DMSs, differentially upregulated methylation sites; CRC, colorectal cancer; TCGA, the cancer genome atlas; CG score, CG site-based risk score.

Identification of Immune-Related Differential Methylation Sites

First, 6450 differentially methylated sites (DMSs) between stage II/III CRC and adjacent normal tissues were identified using the edgeR package, with | log2 FC| > 1.0 and adjusted $P < 0.05$ as thresholds. Thereafter, we focused on the upregulated methylated sites between CRC and adjacent normal tissues and mapped them to immune genes. Overall, 2483 immune genes were downloaded from the immunology database and analysis portal (ImmPort)[3]. Finally, through this analysis, we identified 349 upregulated immune-related methylated sites in stage II/III CRC samples (**Figure 1**).

Statistical Analysis

Machine learning algorithms for predictive models have been described previously (Hu et al., 2019). First, we used univariate and multivariate Cox proportional hazards regression to evaluate the association between overall survival (OS) and the methylation value of each gene site in the training group (Guo et al., 2018). Nine candidate CG sites associated with OS were screened ($P < 0.1$). We used a stepwise selection algorithm for selecting signatures to construct a reliable and an efficient predictive

[3] https://www.immport.org/shared/genelists

prognostic model (**Figure 1**). There are 511 (ΣC^{i}_{9}, $i = 1, 2, \ldots, 9$) combinations of the nine CG sites. For each combination, the CG site-based risk score (CG Score) was calculated based on the following equation, where N is the number of methylated sites of the signature, Meth_{i} is the methylation value of the candidate sites, and Coef_{i} is the univariate Cox regression coefficient:

$$\text{CG Score} = \sum_{i=1}^{\text{N}} \text{Meth}_{\text{i}} * \text{Coef}_{\text{i}} \quad (1)$$

We calculated the CG Score for each sample and the median CG Score in the training group was used as the cut-off value (cut-off = 0.67). Next, we divided all samples into high- and low-risk groups. The Kaplan-Meier survival method, as well as the log-rank test, was applied to compare the prognosis between two groups. In this study, we used area under the curve (AUC) as the performance measurement method for predictive models, which was plotted using the "survivalROC" R package, and all statistical tests were performed using R-3.6.3.

Immune Cell Infiltration in CRC

We calculated relative percent of 22 immune cells in each sample by CIBERSORT which included gene expression of 22 leukocyte subtypes (Newman et al., 2015). Then we compared 22 immune cells infiltrates level between high- and low- risk group samples by Wilcoxon ranked-sum test.

RESULTS

A Five-Methylated Site Signature Predicts the Survival of Patients in the Training Group

The training group, comprising the complete clinical data, was used to further explore the association of 349 methylated sites with prognosis. Survival times were included as dependent variables in univariate Cox proportional hazard regression analysis of the 349 methylated sites. Nine methylated sites were found to be markedly associated with OS ($P < 0.1$) (**Figure 1**). Next, stepwise regression analysis was employed to provide the most effective predictive prognostic model, we developed a five-methylated site signature by selecting the best classification results to construct the final prognostic model (**Supplementary Figure 1**). The CG Score combining the five CG sites (cg11621464, cg13565656, cg18976437, cg20505223, and cg20528583) was determined as follows:

$$\text{CG Score} = \left(1.87 \times \text{meth}_{\text{cg11621464}}\right) + \left(1.11 \times \text{meth}_{\text{cg13565656}}\right)$$
$$+ \left(-1.74 \times \text{meth}_{\text{cg18976437}}\right) + \left(2.40 \times \text{meth}_{\text{cg20505223}}\right) +$$
$$\left(-1.97 \times \text{meth}_{\text{cg20528583}}\right) \quad (2)$$

Confirmation of OS Based on the Methylated Signature in the Training and Test Groups

All patients in the training group were further divided into high- ($n = 64$) and low-risk groups ($n = 63$), and the OS in the low-risk group was higher than that in the high-risk group in the training group (HR: 3.18, 95% CI: 1.82–5.56; $P = 0.000296$, **Figure 2A**). Similarly, using the established prognostic model, patients in the test group were divided into high- ($n = 35$) and low-risk ($n = 20$) groups, and the OS in the low-risk group was higher than that in the high-risk group in the test group (HR: 1.75, 95% CI: 1.03–4.165; $P = 0.022$, **Figure 2B**). We calculated percent of 22 leukocyte cells of high- and low-risk groups by CIBERSORT and then compared immune cell fractions. As a result, we found high-risk group with more naive B cell ($p < 0.05$, **Supplementary Figure 2B**). Su et al. showed that after Chemotherapy-Induced Immunity, the B cells of patients with good curative effects were significantly reduced (Lu et al., 2020).

We used AUC to evaluate the accuracy of the prognostic model. In the training group, the predictive precision of the prognostic signature was more reliable than that of other clinical parameters ($\text{AUC}_{CGScore} = 0.771$, **Figure 2C**). Similar outcomes were obtained for the test group ($\text{AUC}_{CGScore} = 0.724$, **Figure 2D**). The decision curve analysis (DCA) curve showed that the diagnostic value of CG Score is due to clinical indicators, such as stage, age, etc., as well as existing immune biomarkers, such as microsatellite instability (MSI), TMB, etc. The combined model composed of these markers and CG Score can obtain a better net return rate ratio (**Supplementary Figure 3**). Therefore, our results suggest that CG Score may be an efficient prognostic biomarker.

Methylated Signature Has Prognostic Value for CRC Subtypes

Overall, 182 CRC samples were classified into subtypes according to stage, TMB, MSI status, and adjuvant therapy. Next, we carried out a stratified analysis in subtypes to evaluate whether the methylated signature could predict the survival of patients within the same subtype. Log-rank tests of stage II ($P = 0.0002$, **Figure 3A**) and stage III patients ($P = 0.0173$, **Figure 3B**) showed that the methylated signature could classify stage II/III patients into high- and low-risk groups. The standard adjuvant therapy for patients with stage II/III CRC is oxaliplatin and fluorouracil chemotherapy for more than 6 months (Iveson et al., 2019). In the non-adjuvant chemotherapy subtypes, low-risk patients had significantly longer OS than high-risk patients (log-rank $P = 1.3\text{E-}05$, **Figure 3C**).

TMB-H and MSI status are emerging biomarkers associated with immunotherapy for CRC (Schrock et al., 2019), but there were no significant differences estimated in OS between TMB and MSI subgroups (**Supplementary Figure 4A,B**). According to CG Score, low-risk MSS patients had significantly longer OS than high-risk patients (log-rank $P = 0.0004$, **Figure 4A**). This phenomenon was also identified in MSI (log-rank $P = 0.0266$, **Figure 4B**), TMB-L (log-rank $P = 0.0002$, **Figure 4C**) and TMB-H

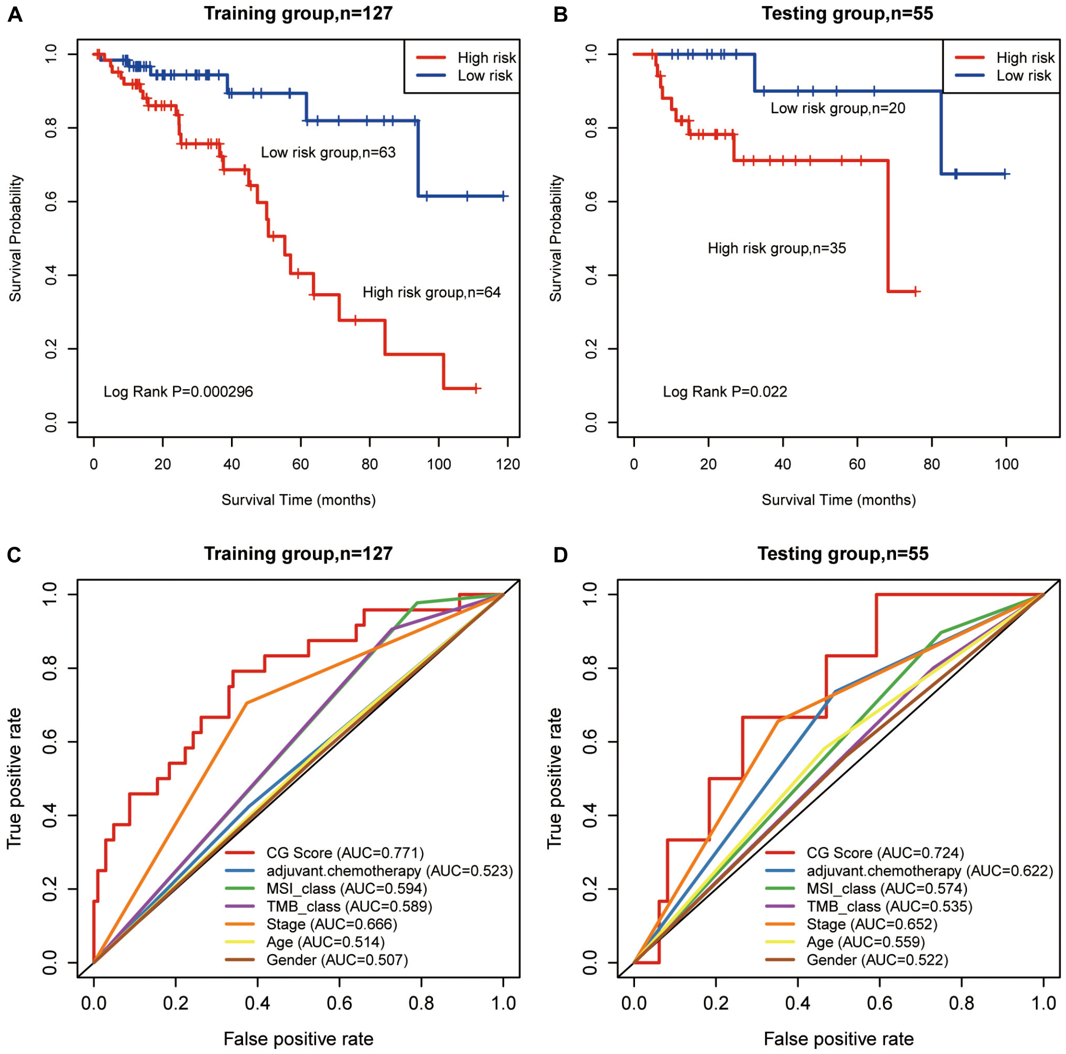


FIGURE 2 | Prognosis of patients with stage II/III CRC was predicted using the methylated signature. **(A,B)** Based on Kaplan–Meier survival curves, patients with stage II/III CRC were classified into high- and low-risk groups using methylation sites as signature in the training and test groups. *P*-values were calculated via log-rank test. **(C,D)** Comparison of the sensitivity and specificity for the prediction of overall survival (OS) based on the CG Score and other clinical parameters. Receiver operating characteristics (ROC) curves for the **(C)** training and **(D)** test groups. CRC, colorectal cancer; MSI, microsatellite instability; TMB, tumor mutational burden; AUC, area under the curve; CG score, CG site-based risk score.

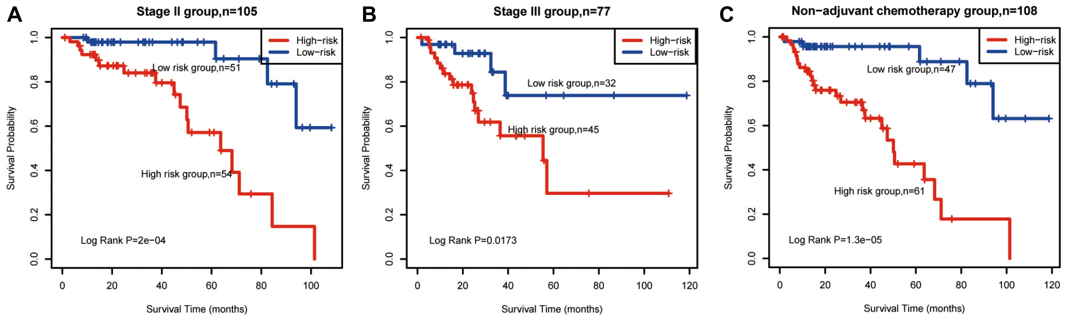


FIGURE 3 | Survival prediction in patients with CRC subtypes. Kaplan–Meier survival curves classified patients into high- and low-risk groups using the methylated signature. **(A)** Stage II group (*n* = 105). **(B)** Stage III group (*n* = 77). **(C)**, non-adjuvant chemotherapy (*n* = 108). Vertical hash marks indicate censored data. CRC: colorectal cancer.

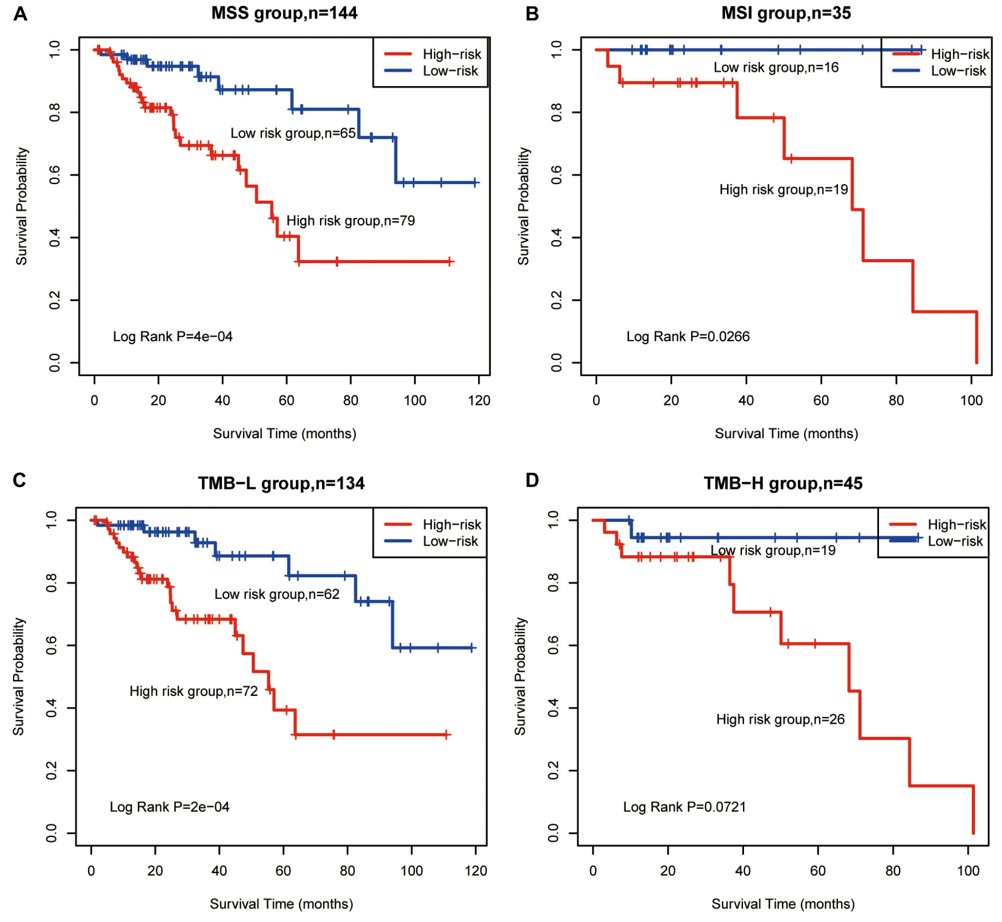


FIGURE 4 | Survival prediction for TMB and MSI subtypes using the methylated signature. Based on Kaplan–Meier survival curves, patients with **(A)** MSS, **(B)** MSI, **(C)** TMB-L, and **(D)** TMB-H were classified into high- and low-risk groups using the methylated signature. Vertical hash marks indicate censored data. MSS, microsatellite stable; MSI, microsatellite instability; TMB-L, low tumor mutational burden; TMB-H, high tumor mutational burden.

groups (log-rank $P = 0.0721$, **Figure 4D**). Thus, the results suggest that CG Score is an efficient prognostic tool for CRC subgroups.

Methylated Signature Is an Independent Prognostic Factor

A multivariate Cox regression analysis using CG Score and clinical parameters (e.g., age, sex, tumor stage, TMB, MSI status, and adjuvant chemotherapy) demonstrated that CG Score was independent of other clinical characteristics both in the training and test groups (**Figure 5A** and **Supplementary Figure 5**). In addition, the CG Score (HR: 6.17, 95% CI: 2.37–16.0, $P < 0.001$, $n = 124$, **Figure 5A**) could be a significant prognostic factor for patients in the high-risk group. According to the multivariate model contained clinicopathological information and the CG site-based risk score, we built a dynamic nomogram (**Figure 5B**).

Correlations Between the Methylated Signature and Immune Biomarkers

Previous research showed immune checkpoint genes including PDL1, interferon-gamma (IFN-γ), PDL2, CTLA4, etc. We found negative correlations between CG Score and other markers (**Supplementary Figure 6**), which indicate the potential of CG Score to be a novel immune-related prognosis biomarker.

Tumor immune dysfunction and exclusion (TIDE) is a gene expression biomarker developed for predicting the clinical response to immune checkpoint blockade (Jiang et al., 2018). We obtained the TIDE score for 182 TCGA-CRC dataset by the online webserver[4]. There is different for TIDE score between high and low CG Score (t-test $p = 0.067$) and the AUC for CG Score under 5 and 3 years are 0.771 and 0.699. In addition, the AUC for TIDE under 5 and 3 years are 0.599 and 0.551 (**Figure 6**). The result indicated that CG Score might be a potential biomarker for immunotherapy especially for Immune checkpoint inhibitors, which show the better performance than existing signatures.

RNA Expression Profile of the Methylated Signature

The five methylated sites were identified on the following genes: *SCTR*, *PIK3CD*, *FGF5*, *PLXNC1*, and *LTBP4*. We observed that

[4] http://tide.dfci.harvard.edu/

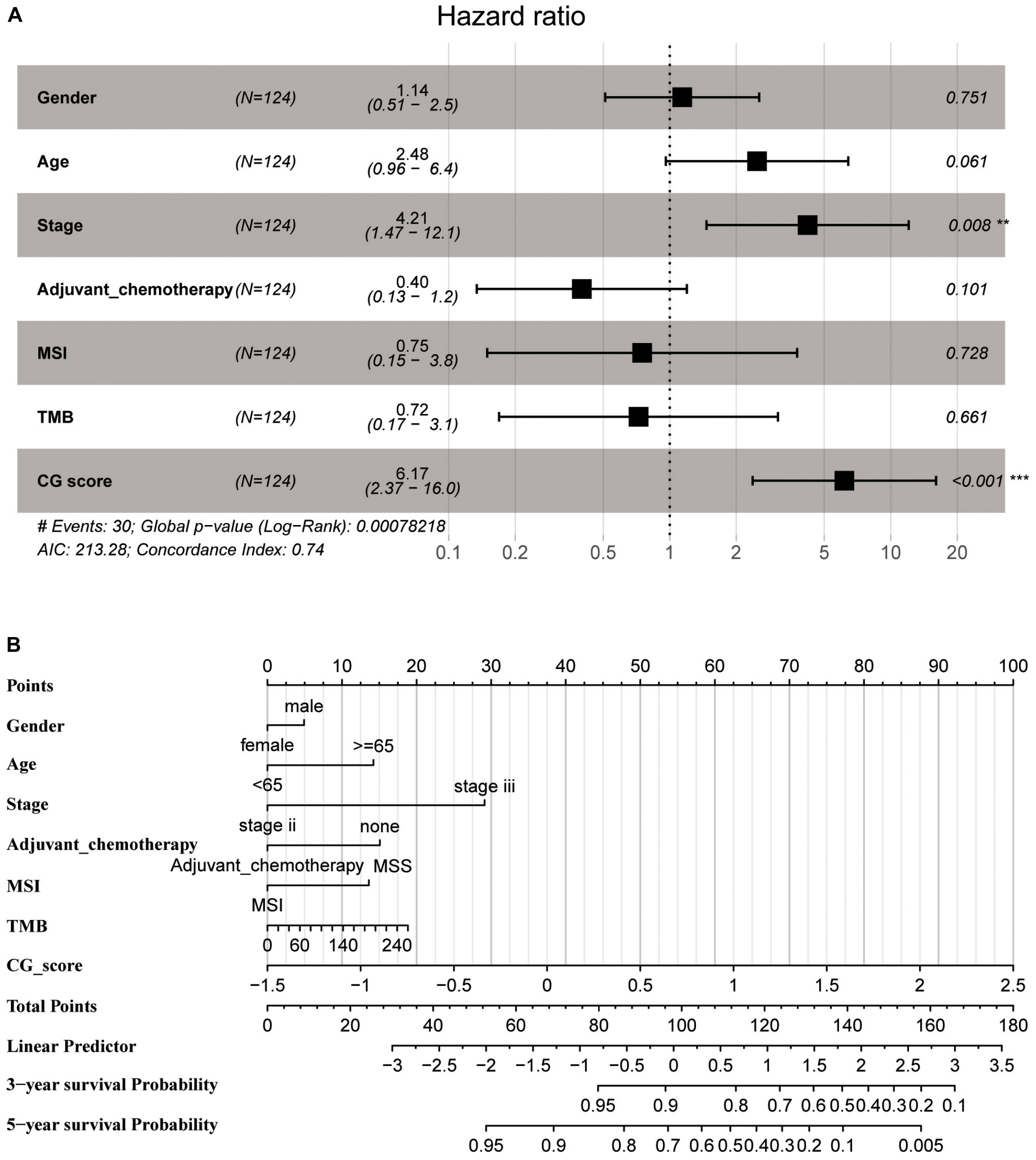


FIGURE 5 | (A) Multivariate Cox regression analysis depicting the association of the methylated signature with the survival of stage II/III CRC patients in the training group. **(B)** The nomogram prediction model was developed by integrating CG Score with the clinical features in the training group. CRC, colorectal cancer; MSI, microsatellite instability; TMB, tumor mutational burden; CG score, CG site-based risk score.

the high expression levels of *PIK3CD*, *PLXNC1*, and *LTBP4* were correlated with the MSI and TMB-H groups (**Figure 7**). Additionally, Chen J.-S. et al. (2019) found that *PIK3CD* was overexpressed in CRC. Li et al. (2019) confirmed that the overexpression of *PLXNC1* could promote cell proliferation and migration. According to a previous study, *LTBP4* acts as a local regulator of transforming growth factor-β expression during tissue deposition and signaling in CRC, and the increase in *LTBP4* expression might cause CRC (Berg et al., 2010). Our RNA expression profile analysis revealed that the above-mentioned genes could be related to MSI or TMB, and therefore, the signature has the potential to replace MSI or TMB as a new prognostic marker.

External Validation of Signature in CRC Datasets

Due to the incompleteness of the methylation profile with survival data or receiving ICB treatment of CRC patients, a large number of studies have confirmed that DNA methylation can cause changes in chromatin structure and DNA stability, thereby inhibiting gene expression (Huang et al., 2021)

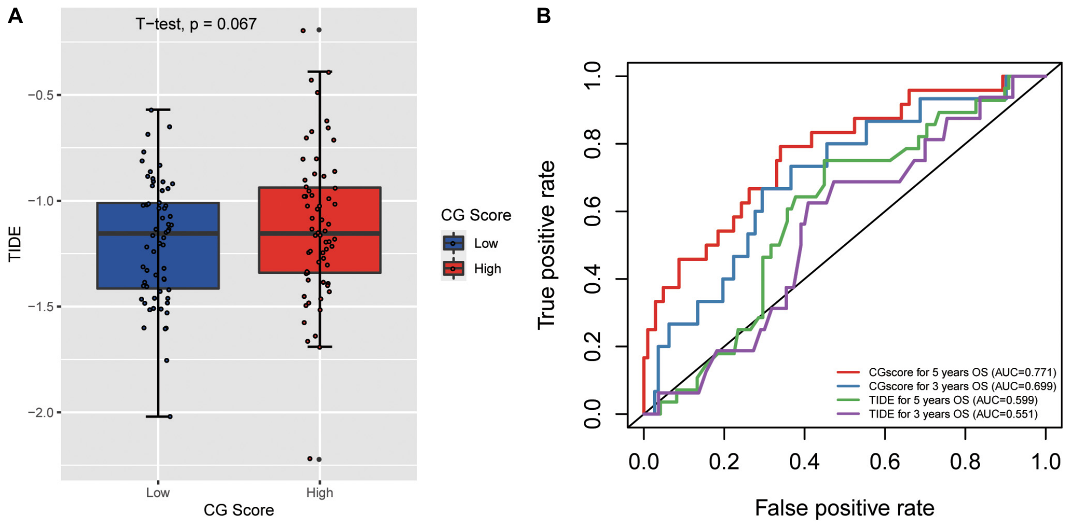


FIGURE 6 | (A) Correlation between tumor mmune dysfunction and exclusion (TIDE) and CG Score. **(B)** The performance of the CG Score and TIDE for overall survival in CRC.

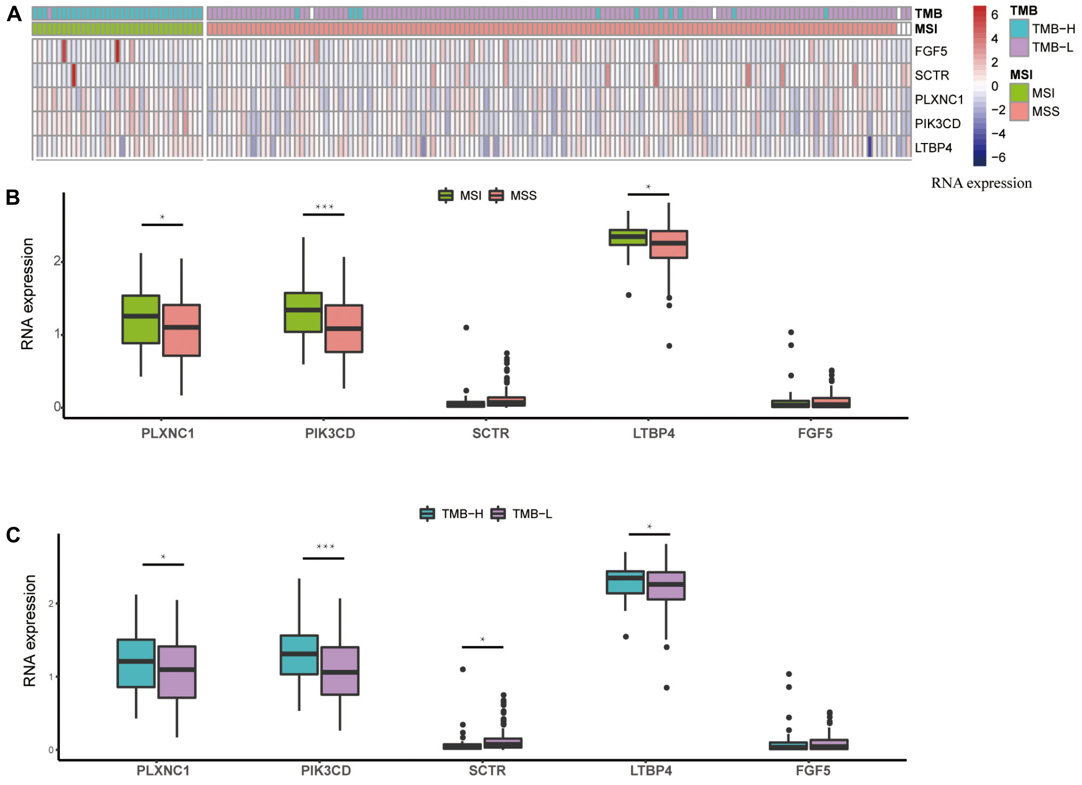


FIGURE 7 | RNA expression profile based on the methylated signature. **(A)** Heatmap of expression levels, after z-score transformation, for the genes involved in the methylated signature. **(B)** The boxplot summarizes the mRNA expression levels in MSS and MSI samples. **(C)** The boxplot summarizes the mRNA expression levels in TMB-H and TMB-L samples. Asterisks indicate genes with significantly ($p < 0.05$) different expression as calculated by t-test. MSS: microsatellite stable; MSI: microsatellite instability; TMB-L, low tumor mutational burden; TMB-H, high tumor mutational burden.

(**Supplementary Figure 7**). We built a gene model based on CG Score (New CG Score = CG Score*correlation between gene expression and methylated sites) to assist in verifying the prognostic value of CG Score, and found that gene model can show good prognostic ability in independent verification datasets (log-rank P-value = 0.043, 0.00021, 0.048) (**Figure 8**).

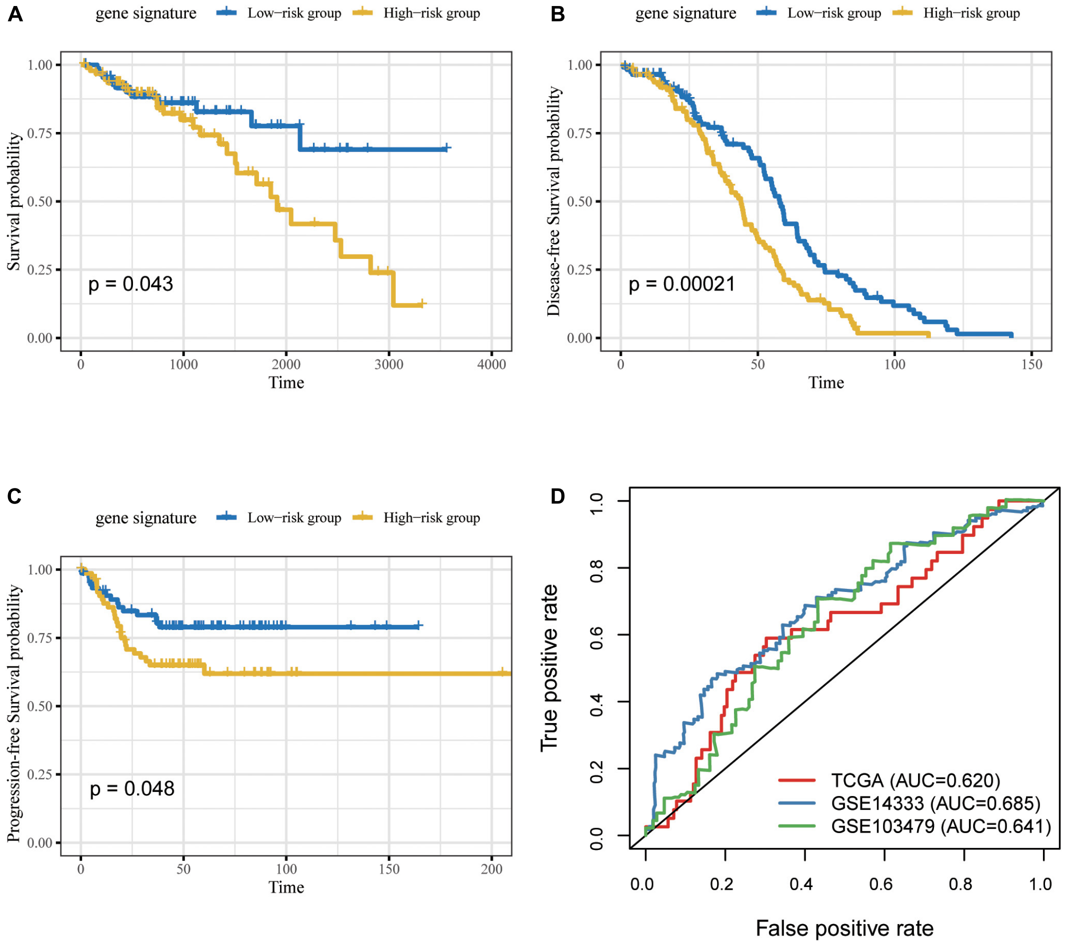


FIGURE 8 | Prognosis of patients with CRC was predicted using the gene signature. Based on Kaplan–Meier survival curves, patients with CRC were classified into high- and low-risk groups using gene expression as signature in the **(A)** TCGA, **(B)** GSE14333 and **(C)** GSE103479 datasets. *P*-values were calculated via log-rank test. **(D)** Comparison of the sensitivity and specificity for the prediction of overall survival based on the gene-based CG Score. CRC, colorectal cancer; AUC, area under the curve.

DISCUSSION

A recent study reported that approximately 30% of CRC patients experience tumor recurrence in the first 3 years after surgery and adjuvant chemotherapy (Sargent et al., 2009). There is a close association between cancer recurrence and clinical or pathological characteristics, such as adjuvant chemotherapy and tumor-node-metastasis classification. However, due to tumor heterogeneity, patients harboring identical clinicopathological features or those undergoing therapeutic interventions present distinct relapse-free survival (Bathe and Farshidfar, 2014). In addition, MSI has become a highly effective immunotherapy biomarker for immune checkpoint inhibitors. About only 15% of stage II and III CRCs present a MSI or deficiency of DNA mismatch repair system (dMMR) phenotype, suggesting that associated with better prognosis than pMMR/MSS tumors (Sinicrope and Sargent, 2012). Moreover, patients with stage II/III dMMR/MSI CRC do not benefit from adjuvant fluoropyrimidine chemotherapy (Tougeron et al., 2016). Thus, it is necessary to propose a new molecular biomarker for predicting the prognosis of patients with stage II/III CRC, especially those with the MSS phenotype. Previous studies have reported that epigenetic modifications play a critical role in carcinogenesis. Promising outcomes have been observed with epigenetic drugs for the treatment of colon cancers (Raynal et al., 2016; Tan et al., 2019). However, it remains unclear whether epigenetic signatures can act as prognostic factors for CRC.

We employed various statistical methods to explore the relationship between methylated signature and prognosis in stage II/III CRC patients, and high-risk patients showed shorter OS than low-risk patients. The AUC for CG Score was estimated to be 0.771 and 0.724 in the training and test groups, respectively. GSEA analysis found that the most significant enrichment pathway in the low- risk group is the cell adhesion pathway, which the results further suggest that process of tumor invasion. Maurer's study found that compared with normal tissues adjacent to cancer, the expression of ICAM-1 (intercellular adhesion molecular-1) in CRC tissues was significantly increased and positively correlated with the infiltration of inflammatory cells in the tumor microenvironment. The results of *in vitro* culture experiments show that the high expression of ICAM-1 depends on the increased dose of IFN-γ and IL-1β(Maurer et al., 1998; Xiang et al., 2001), For the high-risk cohort: KEGG analysis shows that the neuroactive ligand receptor interaction is mainly signaling pathway,that consistent with recent research (Yu et al., 2021) (**Supplementary Figure 8**). Moreover, CG Score was identified as an independent prognosis predictor for patients with CRC. We further discovered that CG Score could distinguish the prognosis of patients in the MSS, TMB-L, and TMB-H subgroups.

The five genes which methylated signature corresponded to after annotation, included *SCTR*, *PIK3CD*, *FGF5*, *PLXNC1*, and *LTBP4*. The hypermethylation of SCTR is a biomarker for precursor lesions in CRC detection (Chen J. et al., 2019; Li et al., 2020). Chen J.-S. et al. (2019) showed that PIK3CD induces CRC cell growth, migration, and invasion by activating AKT/GSK-3β/β-catenin signaling, suggesting that *PIK3CD* could be a novel prognostic biomarker and potential therapeutic target for CRC. Recent studies have shown that the methylated *FGF5* gene could potentially be used as a blood-based biomarker for detecting CRC (Mitchell et al., 2014). *PLXNC1* is involved in intracellular transport, cell migration, and activation of epidermal growth factor receptor and SMAD pathways (Ram et al., 2017). LTBP4 acts as a structural component of the extracellular matrix and local regulator of transforming growth factor-β during tissue deposition and signaling in CRC (Sterner-Kock et al., 2002). Based on the above-mentioned findings, all genes of the methylated signature play critical roles in the tumorigenesis and drug therapy of CRC. To date, only a few studies have investigated the prediction of prognosis in stage II/III CRC patients at the epigenomic level. Thus, in the present study, we proposed specific methylated sites for predicting the prognosis of patients with stage II/III CRC.

Although our CG Score can effectively aid in predicting the prognosis of CRC patients, there is a lack of clinical trials. In addition, there is no definitive evidence to show whether the five methylation sites we identified affect the usage of immune drugs. It will be more convincing if there are data to verify the efficacy of methylated signature and immunotherapy. Moreover, there is a need to verify the ability of the CG Score to distinguish the prognosis of stage I/IV CRC patients as well as the possibility of the CG Score in guiding the immune medication of these patients. In Genomics of Drug Sensitivity in Cancer (GDSC), there are no chemotherapeutics Drugs for the methylation signature. The most sensitive drugs targeting PI3K-Akt signaling pathway (PIK3CA-D) are Alpelisib and Taselisib. If future studies find a drug suitable for the CpG site, experiments can be conducted to verify the sensitivity of the drug.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

GW, YF, and FL conceived and designed the study. FC, LP, EM, and XW acquired the data and drafted the manuscript. SH, DW, YL, and XH analyzed and interpreted the data. XH and SL critically revised the manuscript for important intellectual content. SL, FL, GW, and YF approved the version of the manuscript to be published. All authors contributed to the manuscript and approved the submitted version.

ACKNOWLEDGMENTS

We would like to appreciate editage (www.editage.com) for english language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.684349/full#supplementary-material

REFERENCES

Antelo, M., Balaguer, F., Shia, J., Shen, Y., Hur, K., Moreira, L., et al. (2012). A high degree of LINE-1 hypomethylation is a unique feature of early onset colorectal cancer. *PLoS One* 7:e45357. doi: 10.1371/journal.pone.0045357

Bathe, O. F., and Farshidfar, F. (2014). From genotype to functional phenotype: unraveling the metabolomic features of colorectal cancer. *Genes (Basel)* 5, 536–560. doi: 10.3390/genes5030536

Berg, M., Agesen, T. H., Thiis-Evensen, E., INFAC-study group, Merok, M. A., Teixeira, M. R., et al. (2010). Distinct high resolution genome profiles of early onset and late onset colorectal cancer integrated with gene expression data identify candidate susceptibility loci. *Mol. Cancer* 9:100. doi: 10.1186/1476-4598-9-100

Chen, J., Gingold, J. A., and Su, X. (2019). Immunomodulatory TGF-β signaling in hepatocellular carcinoma. *Trends Mol. Med.* 25, 1010–1023. doi: 10.1016/j.molmed.2019.06.007

Chen, J.-S., Huang, J.-Q., Luo, B., Dong, S.-H., Wang, R.-C., Jiang, Z.-K., et al. (2019). PIK3CD induces cell growth and invasion by activating AKT/GSK-3β/β-catenin signaling in colorectal cancer. *Cancer Sci.* 110, 997–1011. doi: 10.1111/cas.13931

Chen, W. (2015). Cancer statistics: updated cancer burden in China. *Chin. J. Cancer Res.* 27:1. doi: 10.3978/j.issn.1000-9604.2015.02.07

Edge, S. B., and Compton, C. C. (2010). The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann. Surg. Oncol.* 17, 1471–1474. doi: 10.1245/s10434-010-0985-4

Fang, J.-Y., Dong, H.-L., Sang, X.-J., Xie, B., Wu, K.-S., Du, P.-L., et al. (2015). Colorectal cancer mortality characteristics and predictions in China, 1991–2011. *Asian Pac. J. Cancer Prev.* 16, 7991–7995. doi: 10.7314/apjcp.2015.16.17.7991

Guo, J.-C., Wu, Y., Chen, Y., Pan, F., Wu, Z.-Y., Zhang, J.-S., et al. (2018). Protein-coding genes combined with long noncoding RNA as a novel transcriptome molecular staging model to predict the survival of patients with esophageal squamous cell carcinoma. *Cancer Commun (Lond).* 38: 4. doi: 10.1186/s40880-018-0277-0

Hu, S., Yin, X., Zhang, G., and Meng, F. (2019). Identification of DNA methylation signature to predict prognosis in gastric adenocarcinoma. *J. Cell Biochem.* 120, 11708–11715. doi: 10.1002/jcb.28450

Huang, H. Y., Li, J., Tang, Y., Huang, Y. X., Chen, Y. G., Xie, Y. Y., et al. (2021). MethHC 2.0: information repository of DNA methylation and gene expression in human cancer. *Nucleic Acids Res.* 49, D1268–D1275. doi: 10.1093/nar/gkaa1104

Huang, Z.-H., Li, L.-H., Yang, F., and Wang, J.-F. (2007). Detection of aberrant methylation in fecal DNA as a molecular screening tool for colorectal cancer and precancerous lesions. *World J. Gastroenterol.* 13, 950–954. doi: 10.3748/wjg.v13.i6.950

Iveson, T., Boyd, K. A., Kerr, R. S., Robles-Zurita, J., Saunders, M. P., Briggs, A. H., et al. (2019). 3-month versus 6-month adjuvant chemotherapy for patients with high-risk stage II and III colorectal cancer: 3-year follow-up of the SCOT non-inferiority RCT. *Health Technol. Assess.* 23, 1–88. doi: 10.3310/hta23640

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24, 1550–1558. doi: 10.1038/s41591-018-0136-1

Kim, J. C., Choi, J. S., Roh, S. A., Cho, D. H., Kim, T. W., and Kim, Y. S. (2010). Promoter methylation of specific genes is associated with the phenotype and progression of colorectal adenocarcinomas. *Ann. Surg. Oncol.* 17, 1767–1776. doi: 10.1245/s10434-009-0901-y

Kruppa, J., and Jung, K. (2017). Automated multigroup outlier identification in molecular high-throughput data using bagplots and gemplots. *BMC Bioinform.* 18:232. doi: 10.1186/s12859-017-1645-5

Li, D., Zhang, L., Fu, J., Huang, H., Sun, S., Zhang, D., et al. (2020). *SCTR* hypermethylation is a diagnostic biomarker in colorectal cancer. *Cancer Sci.* 111, 4558–4566. doi: 10.1111/cas.14661

Li, R., Teng, X., Zhu, H., Han, T., and Liu, Q. (2019). MiR-4500 regulates PLXNC1 and inhibits papillary thyroid cancer progression. *Horm. Cancer.* 10, 150–160. doi: 10.1007/s12672-019-00366-1

Lu, Y., Zhao, Q., Liao, J. Y., Song, E., Xia, Q., Pan, J., et al. (2020). Complement signals determine opposite effects of B cells in chemotherapy-induced immunity. *Cell* 180, 1081–1097.e1024. doi: 10.1016/j.cell.2020.02.015

Maurer, C. A., Friess, H., Kretschmann, B., Wildi, S., Muller, C., Graber, H., et al. (1998). Over-expression of ICAM-1, VCAM-1 and ELAM-1 might influence tumor progression in colorectal cancer. *Int. J. Cancer* 79, 76–81. doi: 10.1002/(sici)1097-0215(19980220)79:1<76::aid-ijc15<3.0.co;2-f

Mitchell, S. M., Ross, J. P., Drew, H. R., Ho, T., Brown, G. S., Saunders, N. F., et al. (2014). A panel of genes methylated with high frequency in colorectal cancer. *BMC Cancer* 14:54. doi: 10.1186/1471-2407-14-54

Nagai, Y., Sunami, E., Yamamoto, Y., Hata, K., Okada, S., Murono, K., et al. (2017). LINE-1 hypomethylation status of circulating cell-free DNA in plasma as a biomarker for colorectal cancer. *Oncotarget* 8, 11906–11916. doi: 10.18632/oncotarget.14439

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

Ram, M., Najafi, A., and Shakeri, M. T. (2017). Classification and biomarker genes selection for cancer gene expression data using random forest. *Iran J. Pathol.* 12, 339–347.

Raynal, N. J.-M., Lee, J. T., Wang, Y., Beaudry, A., Madireddi, P., Garriga, J., et al. (2016). Targeting calcium signaling induces epigenetic reactivation of tumor suppressor genes in cancer cells. *Cancer Res.* 76, 1494–1505. doi: 10.1158/0008-5472.CAN-14-2391

Rhee, Y.-Y., Kim, M. J., Bae, J. M., Koh, J. M., Cho, N.-Y., Juhnn, Y. S., et al. (2012). Clinical outcomes of patients with microsatellite-unstable colorectal carcinomas depend on L1 methylation level. *Ann. Surg. Oncol.* 19, 3441–3448. doi: 10.1245/s10434-012-2410-7

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

Sanford, T., Meng, M. V., Railkar, R., Agarwal, P. K., and Porten, S. P. (2018). Integrative analysis of the epigenetic basis of muscle-invasive urothelial carcinoma. *Clin. Epigenet.* 10:19. doi: 10.1186/s13148-018-0451-x

Sargent, D., Sobrero, A., Grothey, A., O'Connell, M. J., Buyse, M., Andre, T., et al. (2009). Evidence for cure by adjuvant therapy in colon cancer: observations based on individual patient data from 20,898 patients on 18 randomized trials. *J. Clin. Oncol.* 27, 872–877. doi: 10.1200/JCO.2008.19.5362

Schrock, A. B., Ouyang, C., Sandhu, J., Sokol, E., Jin, D., Ross, J. S., et al. (2019). Tumor mutational burden is predictive of response to immune checkpoint inhibitors in MSI-high metastatic colorectal cancer. *Ann. Oncol.* 30, 1096–1103. doi: 10.1093/annonc/mdz134

Shen, L., Catalano, P. J., Benson, A. B. III, O'Dwyer, P., Hamilton, S. R., and Issa, J.-P. J. (2007). Association between DNA methylation and shortened survival in patients with advanced colorectal cancer treated with 5-fluorouracil based chemotherapy. *Clin. Cancer Res.* 13, 6093–6098. doi: 10.1158/1078-0432.CCR-07-1011

Sillo, T. O., Beggs, A. D., Morton, D. G., and Middleton, G. (2019). Mechanisms of immunogenicity in colorectal cancer. *Br J Surg.* 106, 1283–1297. doi: 10.1002/bjs.11204

Simpson, D. J., Bicknell, J. E., McNicol, A. M., Clayton, R. N., and Farrell, W. E. (1999). Hypermethylation of the *p16/CDKN2A/MTSI* gene and loss of protein expression is associated with nonfunctional pituitary adenomas but not somatotrophinomas. *Genes Chromos. Cancer* 24, 328–336. doi: 10.1002/(SICI)1098-2264(199904)24:4<328::AID-GCC6<3.0.CO;2-P

Sinicrope, F. A., and Sargent, D. J. (2012). Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. *Clin. Cancer Res.* 18, 1506–1512. doi: 10.1158/1078-0432.CCR-11-1469

Sterner-Kock, A., Thorey, I. S., Koli, K., Wempe, F., Otte, J., Bangsow, T., et al. (2002). Disruption of the gene encoding the latent transforming growth factor-beta binding protein 4 (LTBP-4) causes abnormal lung development, cardiomyopathy, and colorectal cancer. *Genes Dev.* 16, 2264–2273. doi: 10.1101/gad.229102

Sun, X., Suo, J., and Yan, J. (2016). Immunotherapy in human colorectal cancer: challenges and prospective. *World J. Gastroenterol.* 22, 6362–6372. doi: 10.3748/wjg.v22.i28.6362

Tan, X., Tong, J., Wang, Y.-J., Fletcher, R., Schoen, R. E., Yu, J., et al. (2019). BET inhibitors potentiate chemotherapy and killing of *SPOP*-mutant colon cancer cells via induction of DR5. *Cancer Res.* 79, 1191–1203. doi: 10.1158/0008-5472.CAN-18-3223

Tougeron, D., Mouillet, G., Trouilloud, I., Lecomte, T., Coriat, R., Aparicio, T., et al. (2016). Efficacy of adjuvant chemotherapy in colon cancer with microsatellite instability: a large multicenter AGEO study. *J. Natl. Cancer Inst.* 108:438. doi: 10.1093/jnci/djv438

Xiang, R., Primus, F. J., Ruehlmann, J. M., Niethammer, A. G., Silletti, S., Lode, H. N., et al. (2001). A dual-function DNA vaccine encoding carcinoembryonic antigen and CD40 ligand trimer induces T cell-mediated protective immunity against colon cancer in carcinoembryonic antigen-transgenic mice. *J. Immunol.* 167, 4560–4565. doi: 10.4049/jimmunol.167.8.4560

Yu, J., Zhang, Q., Wang, M., Liang, S., Huang, H., Xie, L., et al. (2021). Comprehensive analysis of tumor mutation burden and immune microenvironment in gastric cancer. *Biosci. Rep.* 41:BSR20203336. doi: 10.1042/BSR20203336

Conflict of Interest: SL, XW, EM, SH, and DW were employed by ChosenMed Technology (Beijing) Co., Ltd.

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

Prognosis of Non-small-cell Lung Cancer Patients With Lipid Metabolism Pathway Alternations to Immunotherapy

*Tianli Cheng[1,2], Jing Zhang[3], Danni Liu[3], Guorong Lai[3] and Xiaoping Wen[1,2]**

[1] Thoracic Medicine Department I, Hunan Cancer Hospital, Changsha, China, [2] Thoracic Medicine Department I, Affiliated Tumor Hospital of Xiangya Medical School of Central South University, Changsha, China, [3] HaploX Biotechnology, Shenzhen, China

Immune checkpoint inhibitors (ICIs) significantly improve the survival of patients with non-small-cell lung cancer (NSCLC), but only some patients obtain clinical benefits. Predictive biomarkers for ICIs can accurately identify people who will benefit from immunotherapy. Lipid metabolism signaling plays a key role in the tumor microenvironment (TME) and immunotherapy. Hence, we aimed to explore the association between the mutation status of the lipid metabolism pathway and the prognosis of patients with NSCLC treated with ICIs. We downloaded the mutation data and clinical data of a cohort of patients with NSCLC who received ICIs. Univariate and multivariate Cox regression models were used to analyze the association between the mutation status of the lipid metabolism signaling and the prognosis of NSCLC receiving ICIs. Additionally, The Cancer Genome Atlas (TCGA)–NSCLC cohort was used to explore the relationships between the different mutation statuses of lipid metabolism pathways and the TME. Additionally, we found that patients with high numbers of mutations in the lipid metabolism pathway had significantly enriched macrophages (M0- and M1-type), CD4 + T cells (activated memory), CD8 + T cells, Tfh cells and gamma delta T cells, significantly increased expression of inflammatory genes [interferon-γ (IFNG), CD8A, GZMA, GZMB, CXCL9, and CXCL10] and enhanced immunogenic factors [neoantigen loads (NALs), tumor mutation burden (TMB), and DNA damage repair pathways]. In the local-NSCLC cohort, we found that the group with a high number of mutations had a significantly higher tumor mutation burden (TMB) and PD-L1 expression. High mutation status in the lipid metabolism pathway is associated with significantly prolonged progression-free survival (PFS) in NSCLC, indicating that this marker can be used as a predictive indicator for patients with NSCLC receiving ICIs.

Keywords: immune checkpoint inhibitors, non-small-cell lung cancer, predictive marker, lipid metabolism pathway, immune microenvironment

OPEN ACCESS

Edited by:
Shibiao Wan,
St. Jude Children's Research Hospital, United States

Reviewed by:
Tingfang Chen,
University of Pennsylvania, United States
Jingting Yu,
Salk Institute for Biological Studies, United States

****Correspondence:***
Xiaoping Wen
drxpwen@163.com

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

Received: *26 December 2020*
Accepted: *08 June 2021*
Published: *14 July 2021*

Citation:
Cheng T, Zhang J, Liu D, Lai G and Wen X (2021) Prognosis of Non-small-cell Lung Cancer Patients With Lipid Metabolism Pathway Alternations to Immunotherapy. Front. Genet. 12:646362. doi: 10.3389/fgene.2021.646362

INTRODUCTION

Lung cancer is a malignant tumor with the highest morbidity and mortality worldwide (Bray et al., 2018). Non-small-cell lung cancer (NSCLC) is the most common pathological type of lung cancer, and the 5-year survival rate is less than 15% (Herbst et al., 2018; Siegel et al., 2018). Immune checkpoint inhibitors (ICIs) have an antitumor effect by restoring T cell-mediated antitumor immune function and have become a novel clinical treatment tool for NSCLC; however, growing evidence suggests that not all NSCLC patients benefit from ICIs. In the unscreened NSCLC populations, the objective response rate (ORR) to ICIs is commonly less than 20% (Garon et al., 2015). Thus, predicting the effectiveness of ICIs, identifying patients who can benefit from ICIs (Gibney et al., 2016), and maximizing the efficacy of immunotherapy are of great significance for the precise treatment of NSCLC.

PD-L1 expression and tumor mutation burden (TMB) are commonly used markers of immune efficacy. Additionally, high microsatellite instability (MSI-H), deficient mismatch repair (dMMR), tumor-infiltrating lymphocytes (TILs), and the intestinal microbial flora have also show certain predictive value. Although the research conclusions are constantly evolving, some limitations remain (Herbst et al., 2016; Langer et al., 2016; Brody et al., 2017; Chen et al., 2019). For example, a small number of NSCLC patients with low PD-L1 expression seem to be "biomarker negative" but still respond to ICI-based treatment. In contrast, not all patients with high PD-L1 expression can obtain clinical benefit from ICIs (Langer et al., 2016). Additionally, there are many challenges for detecting TMB in clinical practice, including determining the ideal approach for detecting TMB, determining the appropriate cutoff for high or low TMB, and reaching a consensus regarding the different numbers of genes detected by different platforms (Chowell et al., 2018). Moreover, the incidence of MSI-H in NSCLC is very low, so the values of MSI-H and dMMR for predicting the efficacy of ICIs in NSCLC remain to be verified (Warth et al., 2016; Vanderwalde et al., 2018). Hence, how to identify which patients with NSCLC should be treated with ICIs has become an urgent problem in clinical practice.

Metabolic reprogramming processes, such as lipid metabolism, play an important role in the tumor microenvironment (TME) and immunotherapy (DeBerardinis et al., 2008; Yoshida, 2015; Sun, 2016; Baek et al., 2017; Ma et al., 2019). Tumor cells produce large amounts of fatty acids through *de novo* synthesis, and a fatty acid-enriched TME affects the function of effector T cells and M1-type macrophages and is conducive to the production of Tregs and M2 macrophages (Gaber et al., 2017), causing an immunosuppressive TME. Jiang et al. (2018) found that the overexpression of fatty acid synthase (FASN) in ovarian cancer contributed to lipid accumulation in tumor-infiltrating dendritic cells (DCs), causing T cell dysfunction, which in turn induced an impaired antitumor immune response and thus inhibited the ability of fatty acid synthesis to enhance antitumor immunity. Lin et al. (Lin R. et al., 2020) found that tissue-resident memory T (Trm) cells in gastric adenocarcinoma do not use glucose but rather rely on fatty acid oxidation for energy. Cancer cells and Trm cells compete for lipid metabolism, leading to Trm cell death. Blocking PD-L1 can regulate Trm cell metabolism, promote lipid uptake, and further enhance antitumor immune ability. Moreover, several studies have suggested that alterations in specific signaling pathways are associated with the prognosis of patients receiving ICIs and can be used as novel markers to identify patients who will gain benefit from immunotherapy (Teo et al., 2018; Wang et al., 2018). Hence, based on the above results, we aimed to illustrate the association between the mutation status of the lipid metabolism pathway and the prognosis of NSCLC patients treated with ICIs to identify a means to further predict which population of patients with NSCLC will respond to ICIs.

MATERIALS AND METHODS

Immunotherapy Cohort, The Cancer Genome Atlas Cohort, and Local Cohort

One cohort of NSCLC patients who received ICIs [anti-PD-(L)1 monotherapy or anti-PD-(L)1 in combination with anti-CTLA-4 therapy] was derived from a published study reported by Rizvi et al. (2018). This immunotherapy cohort included a total of 240 NSCLC patients with clinical data and mutation data. Additionally, we used the TCGAbiolinks R package to download mutation data, expression data and clinical data from the LUAD and LUSC cohorts in The Cancer Genome Atlas (TCGA) database (Colaprico et al., 2016). The TCGA-LUAD and TCGA-LUSC cohorts were combined into one cohort in the subsequent analysis, called the TCGA cohort (Tomczak et al., 2015). We collected 115 formalin-fixed paraffin-embedded (FFPE) NSCLC samples from the Thoracic Medicine Department I, Hunan Cancer Hospital and Thoracic Medicine Department I, Affiliated Tumor Hospital of Xiangya Medical School of Central South University and performed panel sequencing. The human NSCLC tumor specimens, panel sequencing, data processing, and pathological diagnosis are detailed in the **Supplementary Methods**.

Mutation Data Preprocessing

To explore the association between the mutation status of the lipid metabolism pathway and the prognosis of NSCLC patients receiving ICIs, we downloaded the lipid metabolism gene set from MSigDB (Liberzon et al., 2011). First, we filtered the mutation data and retained only the non-synonymous mutation data. Next, we counted the non-synonymous mutations in the lipid metabolism pathway in each sample. According to the median number of non-synonymous mutations that occurred in this pathway in each dataset, each sample was divided into a group with a high number of mutations and a group with a low number of mutations in lipid metabolism molecules (**Supplementary Table 1**). In the subsequent analysis, we will refer to these two groups as the high mutation group and the low mutation group for short. Additionally, in the mutation frequency analysis, we only compared the top 20 mutations in each cohort.

Immune Microenvironment Analysis

We used the CIBERSORT algorithm and LM22.txt to estimate the proportions of 22 types of TILs from the expression data of NSCLC patients (Newman et al., 2015). Additionally, immune-related genes, immune checkpoint genes and immune-related scores were obtained from published studies (Rooney et al., 2015; Thorsson et al., 2018). The gene set enrichment analysis

(GSEA) algorithm was used to determine the pathways that were significantly enriched or downregulated in the high mutation and low mutation groups (Subramanian et al., 2007). We analyzed and compared the gene ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Reactome pathways enriched in the high and low mutation groups. The enrichment score (ES) and *P*-value were used to evaluate the activity of the pathway and whether there was a significant difference.

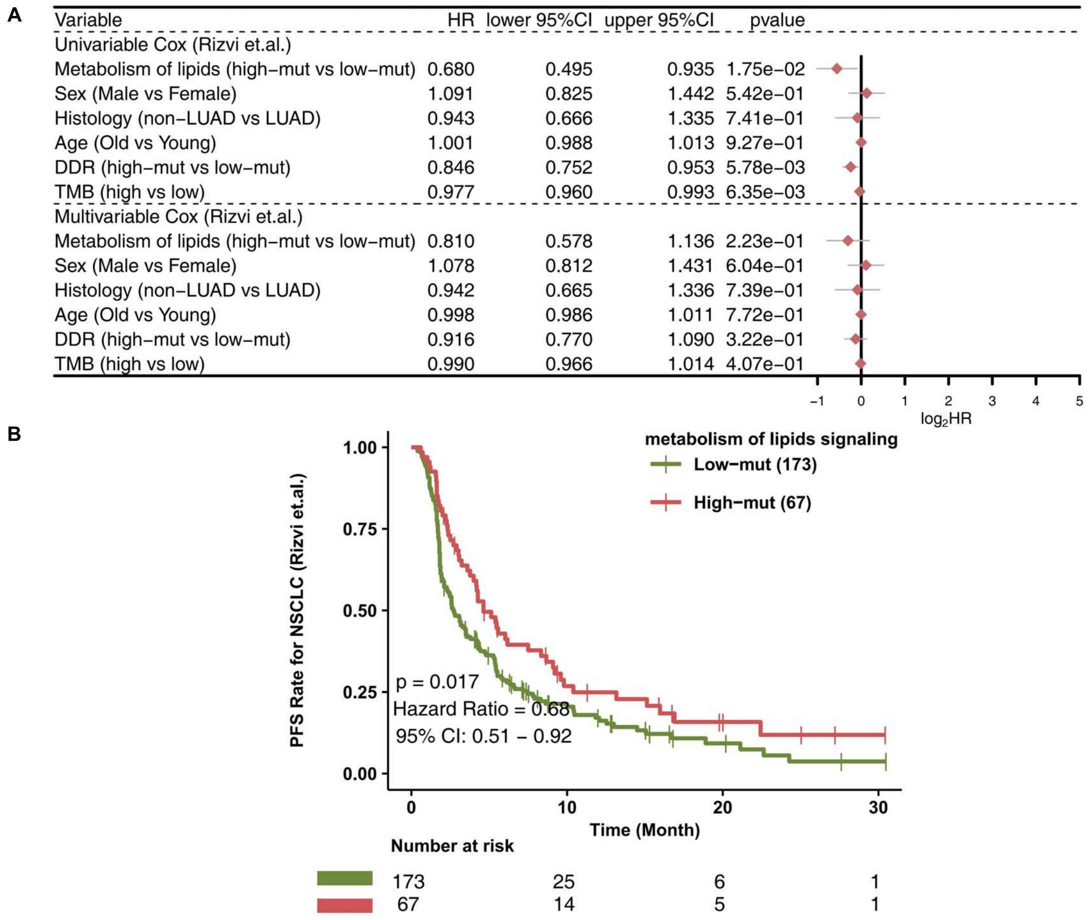
Statistical Analysis

The Mann–Whitney *U* test and Fisher's exact test were applied to the comparison of the difference between the continuous and categorical variables between high-mut and low-mut groups, respectively. We used the Kaplan–Meier (KM) curve, univariate and multivariate Cox model, and the log-rank test to evaluate the effect of the mutation status of lipid metabolism on the prognosis of NSCLC receiving ICIs. Also, the "ggpubr" R package was used to visualize boxplots (Kassambara, 2018). *P* less than 0.05 was regarded as statistically significant.

RESULTS

A Higher Number of Mutations in the Lipid Metabolism Pathway Was Associated With Favorable Prognosis in Patients Treated With ICIs

In the ICI-treated cohort, we used a univariate-Cox model to analyze the effects of a high number of mutations in the lipid metabolism pathway, sex, histological type, and age at the time of prognosis of NSCLC patients receiving immunotherapy (**Figure 1A**). We found that a high number of mutations in the lipid metabolism pathway, a high TMB and a high number of alterations in DNA damage repair (DDR) signaling were associated with prolonged progression-free survival (PFS) in the ICI-treated cohort; however, the results of the multivariate Cox analysis showed that a high number of mutations in the lipid metabolism pathway, a high TMB, or a high number of mutations in DDR signaling could not be used as an independent predictor of the prognosis of patients with NSCLC receiving ICIs. Similarly,



Variable	HR	lower 95%CI	upper 95%CI	pvalue
Univariable Cox (Rizvi et.al.)				
Metabolism of lipids (high−mut vs low−mut)	0.680	0.495	0.935	1.75e−02
Sex (Male vs Female)	1.091	0.825	1.442	5.42e−01
Histology (non−LUAD vs LUAD)	0.943	0.666	1.335	7.41e−01
Age (Old vs Young)	1.001	0.988	1.013	9.27e−01
DDR (high−mut vs low−mut)	0.846	0.752	0.953	5.78e−03
TMB (high vs low)	0.977	0.960	0.993	6.35e−03
Multivariable Cox (Rizvi et.al.)				
Metabolism of lipids (high−mut vs low−mut)	0.810	0.578	1.136	2.23e−01
Sex (Male vs Female)	1.078	0.812	1.431	6.04e−01
Histology (non−LUAD vs LUAD)	0.942	0.665	1.336	7.39e−01
Age (Old vs Young)	0.998	0.986	1.011	7.72e−01
DDR (high−mut vs low−mut)	0.916	0.770	1.090	3.22e−01
TMB (high vs low)	0.990	0.966	1.014	4.07e−01

FIGURE 1 | The value of clinical characteristics and the number of mutations in the lipid metabolism pathway for predicting ICI efficacy. **(A)** Forest plot displaying the results of the univariate and multivariate Cox regression analyses in the ICI-treated cohort (Rizvi et al., 2018). The main portion of the forest plot presents the hazard ratio (HR) and 95% confidence interval (95% CI), and red dots indicate $P < 0.05$. Predictors of favorable outcomes have an HR < 1, and predictors of poor outcome have an HR > 1. **(B)** KM survival curves for PFS in 240 NSCLC patients from the ICI-treated cohort (Rizvi et al., 2018).

NSCLC patients with a high number of mutations in the lipid metabolism pathway had significantly prolonged PFS than those with a low number of mutations [P = 0.017; HR = 0.68; 95% confidence interval (95% CI): 0.51–0.92; **Figure 1B**]. Moreover, we found that the PFS time of the high number of mutations in the lipid metabolism pathway combined with the high TMB group was significantly prolonged than that of the low number of mutations in the lipid metabolism pathway combined with the low TMB group (**Supplementary Figure 1**; P = 0.008; HR = 0.548). Also, we found that the high-TMB group had significantly prolonged PFS time compared with the low-TMB group (**Supplementary Figure 2**; P = 0.024; HR = 0.73).

Comparison of Mutated Genes Between the High and low Mutation Groups

To compare the differences in known cancer driver genes between the high and low mutation groups, we visualized the top 20 mutated driver genes in each group and used Fisher's exact test to calculate the statistical differences. In the ICI-treated cohort, the high mutation group had more gene mutations than the low mutation group. Compared with the low mutation group, the high mutation group had significantly increased TP53 mutations (79.1% vs. 50.9%; $P < 0.05$), PTPRD mutations (20.9% vs. 9.2%; $P < 0.05$), NF1 mutations (17.9% vs. 7.5%; $P < 0.05$), and PTPRT mutations (17.9% vs. 7.5%; $P < 0.05$; **Figure 2A**). Among the above-mentioned genetic mutations with significant differences, most of the mutations were missense mutations, followed by frameshift mutations. In the TCGA cohort, the high mutation group had a higher frequency of driver genes than the low mutation group ($P < 0.05$; **Figure 2B**), while three genes (KRAS, KEAP1, and NFE2L2) showed no significant difference between high and low mutation groups. The results of the mutual exclusivity analysis of the lipid metabolism genes in the high mutation group compared to the low mutation group showed no significant difference (**Supplementary Figure 3**). We also compared lipid metabolism mutation frequency differences between the high mutation group and the low mutation group (**Supplementary Figure 4**). Compared with the low mutation group, the high mutation group had significantly increased mutations in PIK3CG (15.0% vs. 5.20%), PIK3CA (15.0% vs. 2.31%), PIK3C2G (13.4% vs. 1.73%), PIK3C3 (11.9% vs. 1.16%), INPP4B (10.4% vs. 1.16%), NCOR1 (10.4% vs. 1.16%), EP300 (10.4% vs. 1.58%), PTEN (8.96% vs. 1.16%), INPP4A (7.46% vs. 0%), and PIK3R2 (5.97% vs. 0.98%).

Comparison of the Immune Microenvironment Between the High- and Low-Mutation Groups

To explore differences in the TME between the high-mutation group and the low-mutation group, the CIBERSORT algorithm was applied to evaluate the proportions of twenty-two different immune cells in the TME. Compared with the low-mutation group, the high-mutation group had significantly enriched macrophages (M0- and M1-type), CD4 + T cells (activated memory), CD8 + T cells, Tfh cells, and gamma delta T cells (all $P < 0.05$; **Figure 3A**). Additionally, as shown in **Figure 3B**, the number of mutations in the lipid metabolism pathway had a significantly positive correlation with the proportion of macrophages (M1-type), CD4 + T cells (activated memory), CD8 + T cells, Tfh cells, and gamma delta T cells ($R > 0$, $P < 0.05$). A high proportion of CD8 + T cells was significantly correlated with a high proportion of Tfh cells, macrophages (M1-type) and CD4 + T cells (activated memory) ($R > 0$, $P < 0.05$; **Figure 3B**). In contrast, some activated immune cells had a significantly negative correlation with the ratio of resting/suppressive immune cells ($R < 0$, $P < 0.05$; **Figure 3B**). Moreover, we found that the high-mutation group had higher expression levels of immune checkpoint molecules (**Figure 3C**), such as CD274 (PD-L1), LAG3, CD276, and PDCD1 (PD-1), than the low-mutation group. In the local-NSCLC cohort, patients with a high number of mutations in the lipid metabolism pathway had high levels of PD-L1 ($P < 0.05$; **Figure 3D**). **Figure 3E** shows typical cases for each TPS level (lipid metabolism: 3 high-mutation vs. 3 low-mutation cases). Similarly, the expression of inflammatory genes, such as cytotoxicity markers (CD8A, GZMA, and GZMB), antigen processing and presentation markers (MICB and TAP1), and inflammatory cytokines (CXCL9, CXCL10, CCL5, IFNG, IL12A, and TNFRSF18), was significantly higher in the high-mutation group than in the low-mutation group (all $P < 0.05$, **Figure 3F**).

Comparison of Immunogenicity Between the High and Low Mutation Groups

Immunogenicity is a vital factor affecting the prognosis of patients with NSCLC receiving ICIs and the efficacy of ICIs. We determined the differences in immunogenicity between the high and low mutation groups. For TMB, in both the ICI-treated cohort and the TCGA cohort, compared with the low mutation group, the high mutation group had a significantly enhanced TMB (all $P < 0.05$; **Figures 4A,B**). In the Local-NSCLC cohort, we found that patients with a high number of mutations in the lipid metabolism pathway had high levels of TMB ($P < 0.05$; **Figure 4C**). Additionally, the high mutation group had a higher neoantigen load (NAL) than the low mutation group ($P < 0.05$; **Figure 4D**). DDR signaling pathways play a key role in correcting DNA damage. We downloaded eight DDR signaling pathway gene sets from MSigDB and merged these gene sets into one (the merged DDR pathway gene set). In the ICI-treated cohort, in most DDR pathways such as homologous recombination (HR), single-strand break (SSB), double-strand break (DSB), nucleotide excision repair (NER), non-homologous end joining (NHEJ), Fanconi anemia (FA), and merged DDR pathways, the high mutation group had a significantly increased number of mutations ($P < 0.05$; **Figure 4E**). In the TCGA cohort, the high mutation group had a higher number of non-synonymous mutations in all DDR pathways than the low mutation group (all $P < 0.05$; **Figure 4F**).

Differences in Pathway Activity Between the High and Low Mutation Groups

Alterations in functional pathway activity also have impacts on the efficacy of ICIs and the prognosis of NSCLC patients

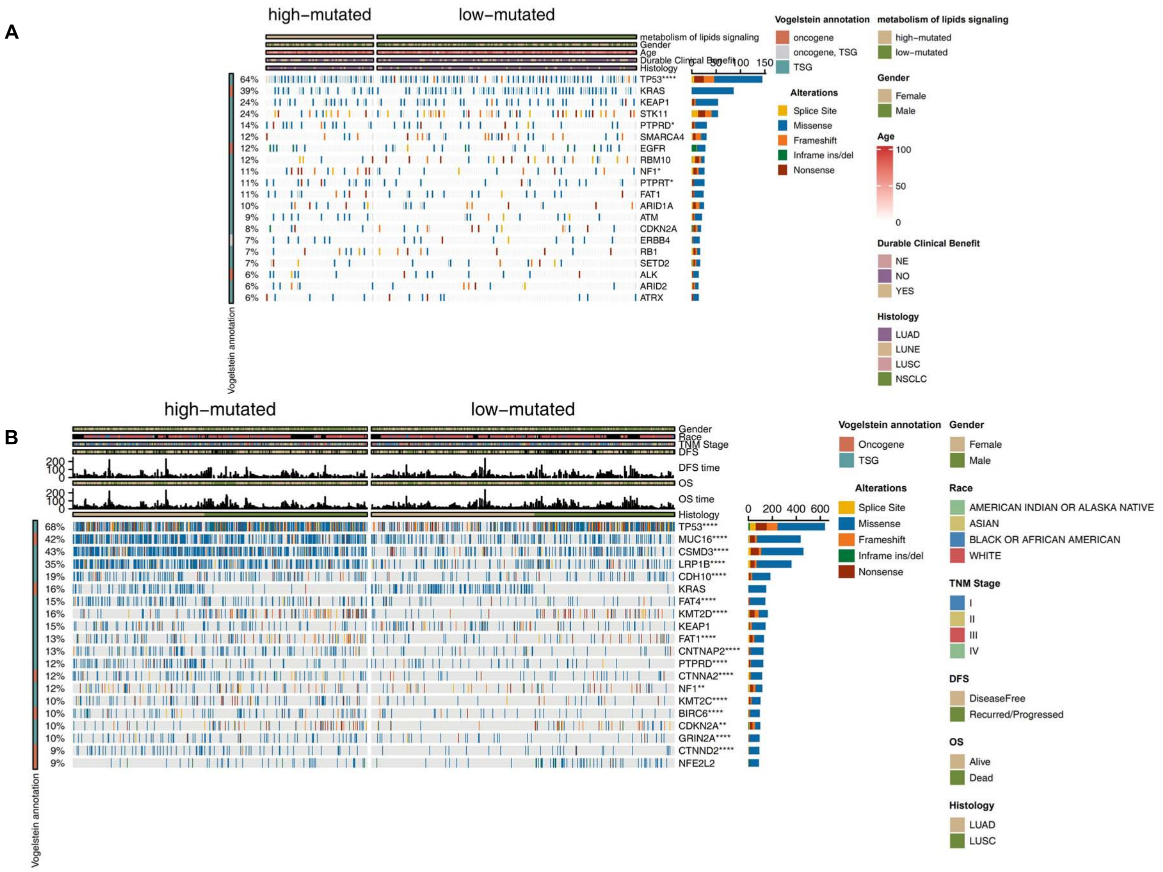


FIGURE 2 | Genomic profiles of NSCLC patients in the ICI-treated cohort (Rizvi et al., 2018) **(A)** and TCGA-NSCLC **(B)** cohorts. The top 20 genes with the highest mutation frequencies and the corresponding clinical information are shown. The top five genes with the highest mutation frequencies in the ICI-treated cohort (Rizvi et al., 2018) were TP53, KRAS, KEAP1, STK11, and PTPRD. The top five genes with the highest mutation frequencies in the TCGA cohort were TP53, TTN, MUC16, CSMD3, and RYR2. The mutation types are indicated as follows: yellow indicates splice site mutations, blue indicates missense mutations, orange indicates frameshift mutations, green indicates in-frame insertions/deletions, and brown indicates nonsense mutations. The clinical characteristics are shown as patient annotations.

receiving ICIs. We used the ClusterProfiler R package to perform GSEA with the NSCLC expression data from the high and low mutation groups. Immune-related pathway terms, such as lymphocyte recruitment and participation in the inflammatory response, lymphocyte aggregation, interleukin 1, and BCR pathway activation, were significantly enriched in the high mutation group (**Figure 5A**). In contrast, some pathway terms related to immune depletion, such as fatty acid synthesis, fatty acid metabolism and regulation of fibroblast proliferation, were significantly downregulated in the high mutation group (**Figure 5B**). Additionally, some carcinogenic pathways, such as the canonical WNT pathway and the NOTCH pathway, were significantly upregulated in the low mutation group compared with the high mutation group (**Figure 5C**).

DISCUSSION

To date, with the gradual increase in in-depth research on immune checkpoints, breakthroughs have been made in the research of ICIs, which have revolutionized the diagnosis and treatment of NSCLC; however, many challenges remain in clinical application, such as the limited population that benefits and the lack of effective biomarkers (Garon et al., 2019; Garassino et al., 2020). In the TME, both tumor cells and immune cells can undergo metabolic reorganization to adapt to a microenvironment with low oxygen, acidity and low nutrition (Wu and Dai, 2017). The activity of the lipid metabolism pathway can affect the recruitment, infiltration and activation of TILs (Yang et al., 2016; Saleh and Elkord, 2019; Jiang et al., 2020). Novel treatments that regulate lipid metabolism may effectively improve the immunotherapy efficacy and patient prognosis. In this study, we found that a high number of mutations in the lipid metabolism pathway was related to a favorable prognosis in patients with NSCLC receiving ICIs. Next, we analyzed the potential relationships between the number of mutations in the lipid metabolism pathway and immunogenicity and the TME. Patients with a high number of mutations in the lipid metabolism pathway had significantly enhanced immunogenic factors (such as

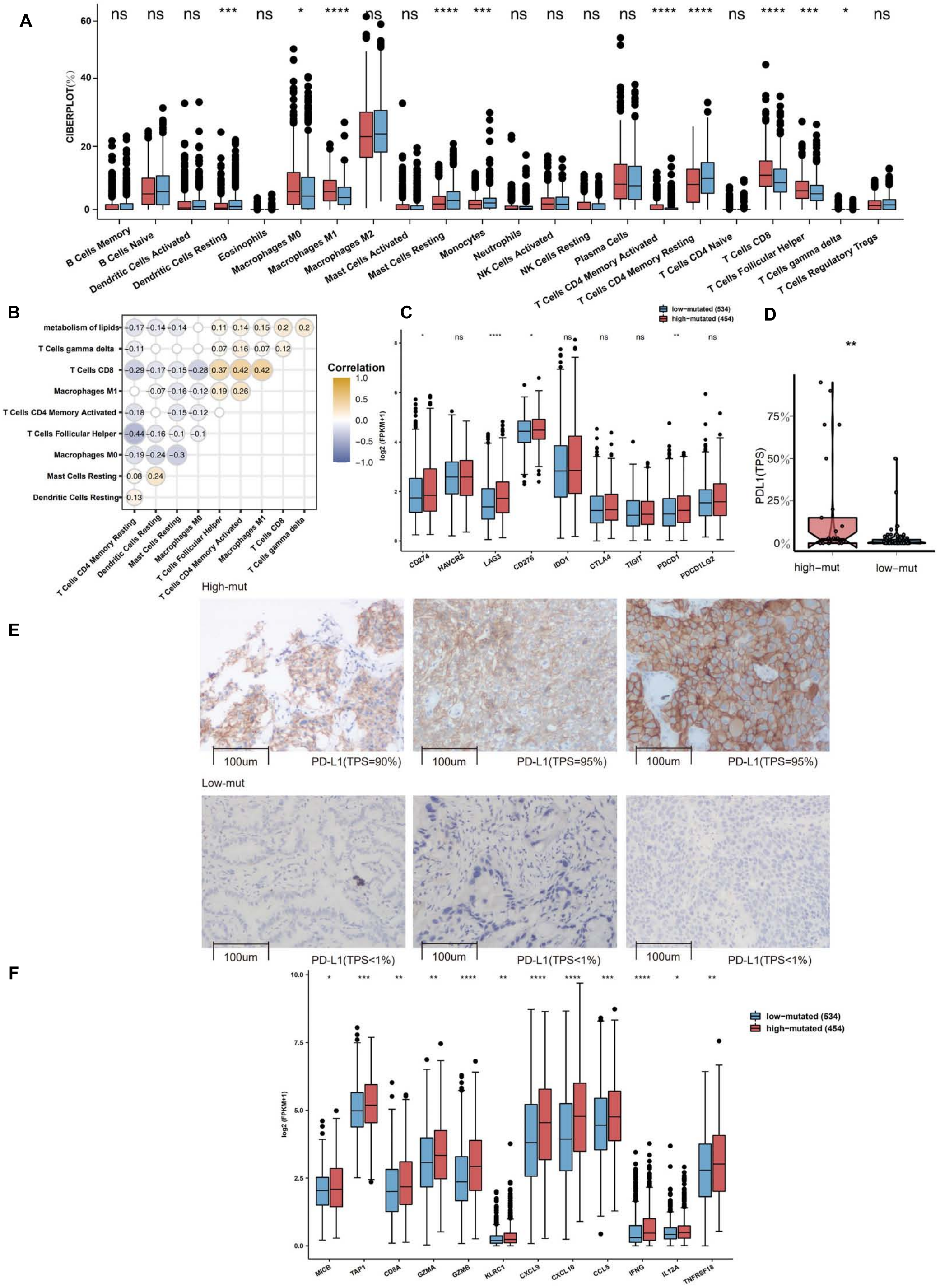


FIGURE 3 | (A) Comparison of the fractions of 22 types of immune cells as estimated by the CIBERSORT algorithm between the high and low mutation groups in the TCGA cohort. **(B)** The correlations between the number of mutations in the lipid metabolism pathway and the proportions of immune cells. **(C)** Comparison of the expression of immune-related genes between the high and low mutation groups in the TCGA cohort. **(D)** Comparison of the expression of PD-L1 (TPS) between the high and low mutation groups in the Local-NSCLC cohort. **(E)** The typical cases for each TPS level between the high (three samples; high PD-L1 TPS) and low mutation (three samples; no PD-L1 TPS) groups in the Local-NSCLC. Using HE and PD-L1 stained slides, we manually assessed the number of tumor cells, the sample size (diameter), the crush rate with a cut-off value of <1% (no PD-L1 TPS), 1–50% (low PD-L1 TPS), 50% < (high PD-L1 TPS), and the TPS for each biopsy sample using the slide that contained the most tumor cells. The TPS level was evaluated by pathologists who completed training courses in TPS estimation.
(F) Heatmap depicting the mean differences in the expression of proinflammatory and antigen presentation genes between the high and low mutation groups in the TCGA cohort. Each square represents the fold change or the mean difference in the expression of these genes between the high and low mutation groups in the TCGA cohort. Red indicates upregulation.

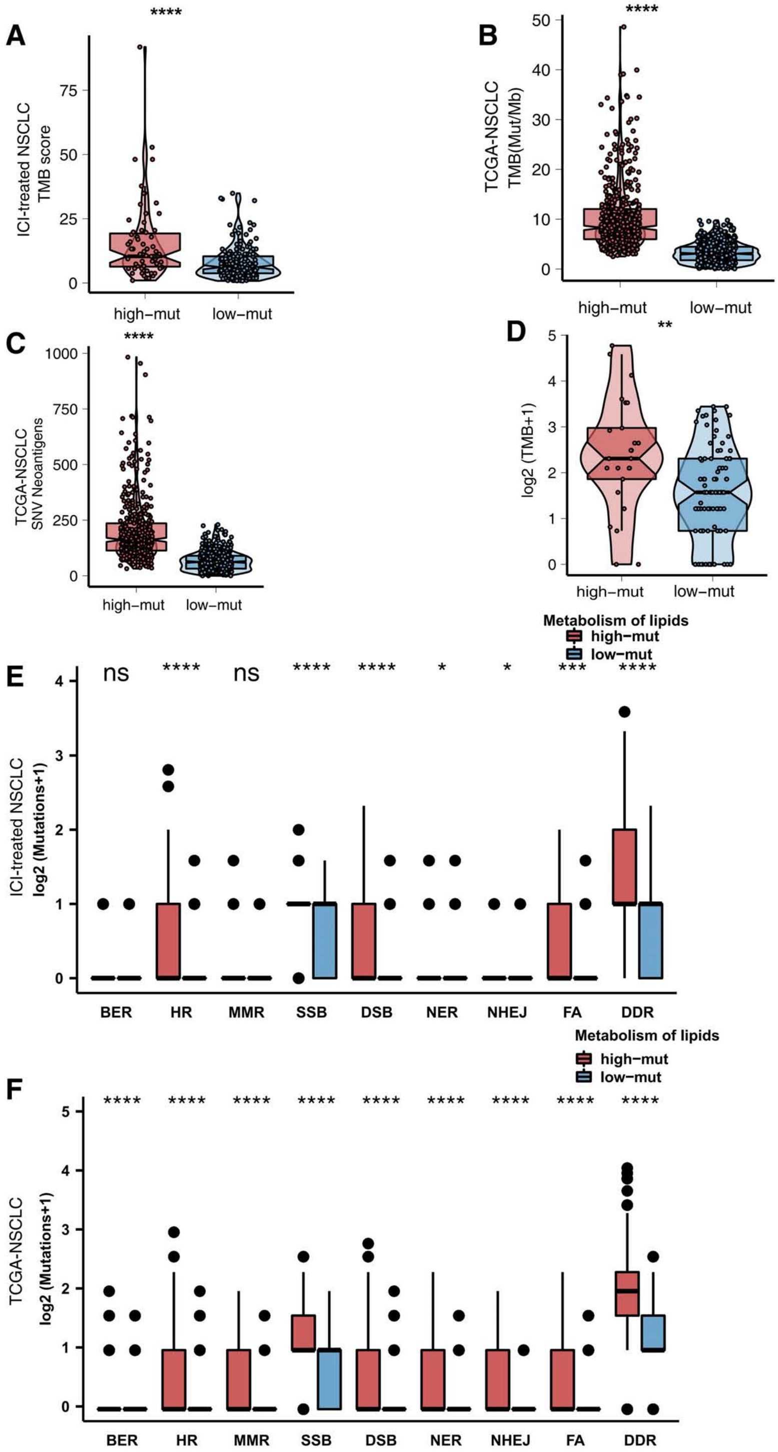


FIGURE 4 | (A) Comparison of TMB scores between the high and low mutation groups in the ICI-treated cohort (Rizvi et al., 2018). **(B)** Comparison of TMB between the high and low mutation groups in the TCGA cohort. **(C)** Comparison of TMB between the high and low mutation groups in the Local-NSCLC cohort. **(D)** Comparison of NAL between the high and low mutation groups in the TCGA cohort. **(E)** Comparison of DNA damage-related gene set alterations between the high and low mutation groups in the ICI-treated cohort (Rizvi et al., 2018). **(F)** Comparison of DNA damage-related gene set alterations between the high and low mutation groups in the TCGA cohort.

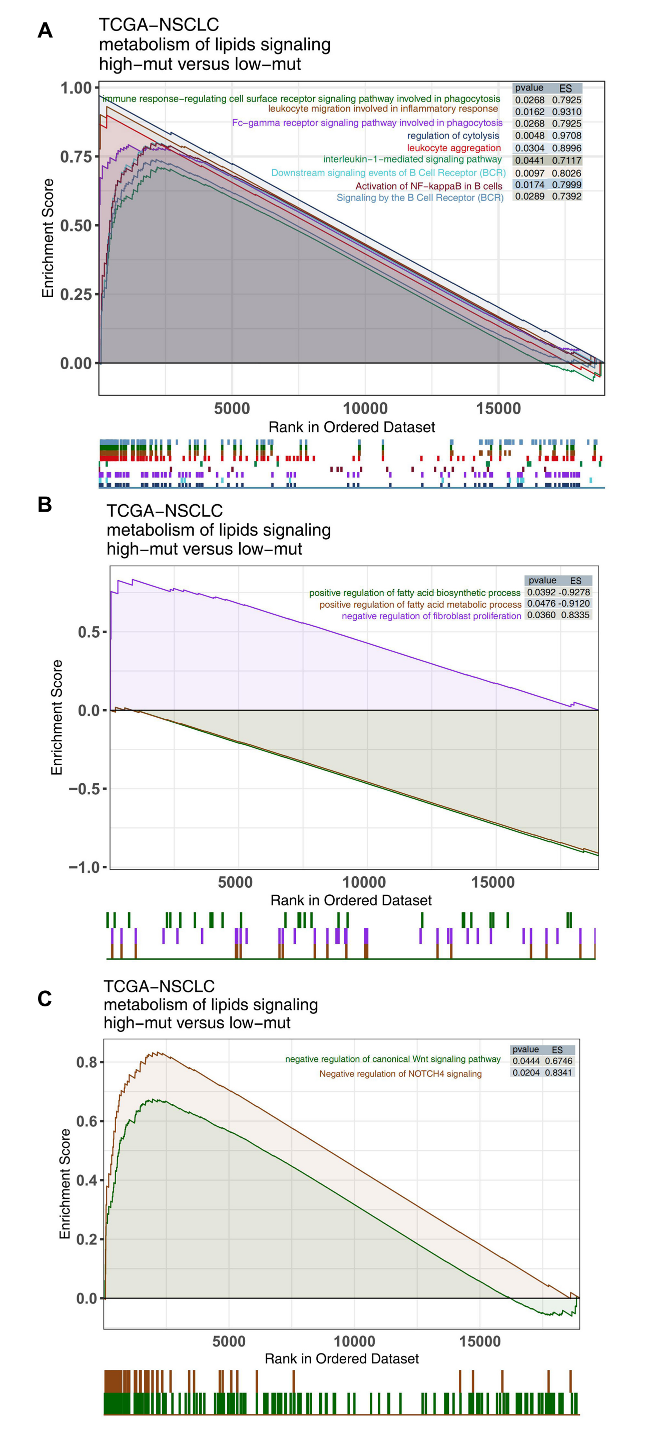


FIGURE 5 | Comparison of GSEA results between the high and low mutation groups in the TCGA cohort. GSEA-identified differences in immune cell **(A)**, exhaustion-related **(B)**, and oncogenic pathway activities **(C)** between the high and low mutation groups in the TCGA cohort.

TMB, NAL, and DDR pathway mutations) and enriched activated immune cells with upregulated inflammatory gene expression profiles.

The inflammatory TME in patients with a high number of mutations in the lipid metabolism pathway may be related to a better prognosis with ICI treatment. Compared with patients with a low number of mutations in the lipid metabolism pathway, patients with a high number of mutations had significantly increased proportions of infiltrating activated immune cells [macrophages (M0- and M1-type), CD4 + T cells (activated memory), CD8 + T cells, Tfh cells, and gamma delta T cells] and upregulated inflammatory expression profiles (IFNG, CD8A, GZMA, GZMB, CXCL9, and CXCL10). Tumor cell necrosis induced by the perforin-granzyme pathway and tumor cell apoptosis induced by the Fas-FasL pathway are regarded as two vital mechanisms by which CD8 + T cells exert antitumor immunity. Additionally, CD8 + T cells can also induce iron-mediated tumor cell death by secreting IFN-γ, which is a newly identified method of cell death that differs from apoptosis and necrosis (Dixon et al., 2012). IFN-γ can downregulate the expression of two subunits of the glutamate-cystine antiporter on the surface of tumor cells, namely, solute carrier family 3 member 2 (SLC3A2) and SLC7A11, thereby inhibiting tumors. Cystine uptake by the cell reduces glutathione synthesis and ultimately leads to insufficient synthesis of glutathione peroxidase 4 (GPX4), which inhibits the cell from effectively removing peroxide. Lipids cause iron-induced death in cells under iron-dependent conditions (Friedmann Angeli et al., 2019; Wang et al., 2019). IFNγ is mainly derived from CD8 + T cells and is also an important cytokine for CD8 + T cells to complete immune-mediated killing. In addition to mediating iron-induced cell death, IFN-γ can also promote antigen presentation and tumor cell killing. IFN-γ can activate the JAK-STAT signaling pathway through interferon receptors acting on tumor cells, thereby upregulating the expression of interferon-stimulated genes (ISGs) and enhancing major histocompatibility complex I (MHC-I) expression on the cell membrane. The expression of MHC-I molecules and intracellular immune proteasomes promotes the recognition of tumor cells by immune cells and simultaneously sensitizes tumor cells to apoptosis signals, which ultimately leads to tumor cell death (Quail and Joyce, 2013; Schneider et al., 2014). M1-type macrophages highly express TNF, inducible nitric oxide synthase (iNOS), MHCII and other proteins, which play an antitumor effect. Chemokines (CXCL9 and CXCL10) play an important role in recruiting CD8 + T cells and NK cells to the TME. The above results suggest the presence of an inflammatory immune microenvironment in the high mutation group (Lin et al., 2019).

The significantly enhanced immunogenicity in patients with a high number of mutations in the lipid metabolism signaling may be associated with a favorable prognosis with ICIs. Mutations in the DDR signaling can contribute to the up-regulation of genome instability and cause accumulated DNA damage, which may be a biomarker for identifying potential ICI responders in multiple cancer types (Teo et al., 2018; Wang et al., 2018). Patients with advanced urothelial cancers with mutations in the DDR pathway had a significantly increased ORR to immunotherapy (67.9% vs. 18.8%; $P < 0.001$) (Teo et al., 2018). Additionally, Wang et al. (2018) found that patients with co-mutations in the DDR pathway had significantly prolonged OS and PFS compared with patients without co-mutations. The TMB has been regarded as a potential molecular marker for predicting ICI response, and its utility has been gradually confirmed (Jessurun et al., 2017). An increased TMB can promote the production of more tumor neoantigens (McGranahan et al., 2016). Neoantigens are presented to DCs, which can promote the transformation of T cells into mature and activated T cells, and high NAL is associated with sensitivity to anti-PD-1/CTLA-4 treatments (Lin A. et al., 2020). In this study, we found that patients with a high number of mutations in the lipid metabolism pathway had a significantly increased TMB, NAL, and mutations of the DDR pathway. Therefore, the above results suggest that up-regulated immunogenicity may be a strategy generating favorable prognoses for NSCLC patients with a high number of mutations in the lipid metabolism pathway. This study analyzed the prognosis of ICI treatment and mutation status of lipid metabolism in patients with non-small cell lung cancer and attempted to elucidate the potential role of a high number of lipid metabolism mutations as a biomarker for screening the predominant population of NSCLC preferred for immunotherapy; however, this study still has several limitations. First, this work included only one ICI-treated cohort of NSCLC, which may introduce bias when screening biomarkers for the prognosis of ICIs of NSCLC. Second, targeted sequencing (MSK-IMPACT) was used to detect somatic mutations in the ICI-treated cohort and included significantly fewer gene mutations compared to whole-exome sequencing (WES). Third, this study cannot separate the effect of the TMB or the mutation counts of DDR signaling from the effect of the mutation status of lipid metabolism on the prognosis of NSCLC patients receiving ICIs. We hope to conduct relevant cell or animal experiments in the future to verify how a high number of lipid metabolism mutations affect the efficacy of immunotherapy and explore their relationship with the TME. We also hope to study NSCLC patients receiving ICIs to separate the effect of the TMB or the mutation counts of DDR signaling.

CONCLUSION

Our study provided solid evidence that high-mutated lipid metabolism signaling was associated with prolonged PFS in NSCLC patients receiving ICIs. Hence, high-mutated lipid metabolism signaling can act as a potential biomarker for ICIs among NSCLC.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

XW: conceptualization. TC: formal analysis, visualization, and writing – original draft. TC, JZ, DL, and GL: writing – review and editing. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.646362/full#supplementary-material

Supplementary Figure 1 | KM survival curves for PFS in NSCLC patients (ICI-treated cohort) grouped according to TMB levels and mutation status of lipids metabolism.

Supplementary Figure 2 | KM survival curves for PFS in NSCLC patients (ICI-treated cohort) grouped according to TMB levels.

Supplementary Figure 3 | The mutual exclusivity analysis of lipid metabolism genes in the high and low mutation groups.

Supplementary Figure 4 | Comparison of mutation frequencies of lipid metabolism genes between the high and low mutation groups in the ICI-treated cohort (Rizvi et al., 2018).

Supplementary Table 1 | The mutation counts of lipid metabolism signaling of NSCLC patients in the TCGA-NSCLC cohort.

REFERENCES

Baek, A. E., Yu, Y.-R. A., He, S., Wardell, S. E., Chang, C.-Y., Kwon, S., et al. (2017). The cholesterol metabolite 27 hydroxycholesterol facilitates breast cancer metastasis through its actions on immune cells. *Nat. Commun.* 8:864. doi: 10.1038/s41467-017-00910-z

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

Brody, R., Zhang, Y., Ballas, M., Siddiqui, M. K., Gupta, P., Barker, C., et al. (2017). PD-L1 expression in advanced NSCLC: insights into risk stratification and treatment selection from a systematic literature review. *Lung Cancer* 112, 200–215. doi: 10.1016/j.lungcan.2017.08.005

Chen, Y., Liu, Q., Chen, Z., Wang, Y., Yang, W., Hu, Y., et al. (2019). PD-L1 expression and tumor mutational burden status for prediction of response to chemotherapy and targeted therapy in non-small cell lung cancer. *J. Exp. Clin. Cancer Res.* 38:193. doi: 10.1186/s13046-019-1192-1

Chowell, D., Morris, L. G. T., Grigg, C. M., Weber, J. K., Samstein, R. M., Makarov, V., et al. (2018). Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* 359, 582–587. doi: 10.1126/science.aao4572

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

DeBerardinis, R. J., Lum, J. J., Hatzivassiliou, G., and Thompson, C. B. (2008). The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab.* 7, 11–20. doi: 10.1016/j.cmet.2007.10.002

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

Friedmann Angeli, J. P., Krysko, D. V., and Conrad, M. (2019). Ferroptosis at the crossroads of cancer-acquired drug resistance and immune evasion. *Nat. Rev. Cancer* 19, 405–414. doi: 10.1038/s41568-019-0149-1

Gaber, T., Strehl, C., and Buttgereit, F. (2017). Metabolic regulation of inflammation. *Nat. Rev. Rheumatol.* 13, 267–279. doi: 10.1038/nrrheum.2017.37

Garassino, M. C., Gadgeel, S., Esteban, E., Felip, E., Speranza, G., Domine, M., et al. (2020). Patient-reported outcomes following pembrolizumab or placebo plus pemetrexed and platinum in patients with previously untreated, metastatic, non-squamous non-small-cell lung cancer (KEYNOTE-189): a multicentre, double-blind, randomised, placebo-controlle. *Lancet Oncol.* 21, 387–397. doi: 10.1016/S1470-2045(19)30801-0

Garon, E. B., Hellmann, M. D., Rizvi, N. A., Carcereny, E., Leighl, N. B., Ahn, M.-J., et al. (2019). Five-Year Overall Survival for Patients With Advanced Non-Small-Cell Lung Cancer Treated With Pembrolizumab: results From the Phase I KEYNOTE-001 Study. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 37, 2518–2527. doi: 10.1200/JCO.19.00934

Garon, E. B., Rizvi, N. A., Hui, R., Leighl, N., Balmanoukian, A. S., Eder, J. P., et al. (2015). Pembrolizumab for the treatment of non-small-cell lung cancer. *N. Engl. J. Med.* 372, 2018–2028. doi: 10.1056/NEJMoa1501824

Gibney, G. T., Weiner, L. M., and Atkins, M. B. (2016). Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet Oncol.* 17, e542–e551. doi: 10.1016/S1470-2045(16)30406-5

Herbst, R. S., Baas, P., Kim, D.-W., Felip, E., Pérez-Gracia, J. L., Han, J.-Y., et al. (2016). Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* 387, 1540–1550. doi: 10.1016/S0140-6736(15)01281-7

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature* 553, 446–454. doi: 10.1038/nature25183

Jessurun, C. A. C., Vos, J. A. M., Limpens, J., and Luiten, R. M. (2017). Biomarkers for Response of Melanoma Patients to Immune Checkpoint Inhibitors: a Systematic Review. *Front. Oncol.* 7:233. doi: 10.3389/fonc.2017.00233

Jiang, L., Fang, X., Wang, H., Li, D., and Wang, X. (2018). Ovarian Cancer-Intrinsic Fatty Acid Synthase Prevents Anti-tumor Immunity by Disrupting Tumor-Infiltrating Dendritic Cells. *Front. Immunol.* 9:2927. doi: 10.3389/fimmu.2018.02927

Jiang, Z., Hsu, J. L., Li, Y., Hortobagyi, G. N., and Hung, M.-C. (2020). Cancer Cell Metabolism Bolsters Immunotherapy Resistance by Promoting an Immunosuppressive Tumor Microenvironment. *Front. Oncol.* 10:1197. doi: 10.3389/fonc.2020.01197

Kassambara, A. (2018). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.7. Available online at: https://CRAN.R-project.org/package=ggpubr.

Langer, C. J., Gadgeel, S. M., Borghaei, H., Papadimitrakopoulou, V. A., Patnaik, A., Powell, S. F., et al. (2016). Carboplatin and pemetrexed with or without pembrolizumab for advanced, non-squamous non-small-cell lung cancer: a randomised, phase 2 cohort of the open-label KEYNOTE-021 study. *Lancet. Oncol.* 17, 1497–1508. doi: 10.1016/S1470-2045(16)30498-3

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Lin, A., Wei, T., Meng, H., Luo, P., and Zhang, J. (2019). Role of the dynamic tumor microenvironment in controversies regarding immune checkpoint inhibitors for the treatment of non-small cell lung cancer (NSCLC) with EGFR mutations. *Mol. Cancer* 18:139. doi: 10.1186/s12943-019-1062-7

Lin, A., Zhang, J., and Luo, P. (2020). Crosstalk Between the MSI Status and Tumor Microenvironment in Colorectal Cancer. *Front. Immunol.* 11:2039. doi: 10.3389/fimmu.2020.02039

Lin, R., Zhang, H., Yuan, Y., He, Q., Zhou, J., Li, S., et al. (2020). Fatty Acid Oxidation Controls CD8(+) Tissue-Resident Memory T-cell Survival in Gastric Adenocarcinoma. *Cancer Immunol. Res.* 8, 479–492. doi: 10.1158/2326-6066.CIR-19-0702

Ma, X., Bi, E., Lu, Y., Su, P., Huang, C., Liu, L., et al. (2019). Cholesterol Induces CD8(+) T Cell Exhaustion in the Tumor Microenvironment. *Cell Metab.* 30, 143–156.e5. doi: 10.1016/j.cmet.2019.04.002

McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 351, 1463–1469. doi: 10.1126/science.aaf1490

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

Quail, D. F., and Joyce, J. A. (2013). Microenvironmental regulation of tumor progression and metastasis. *Nat. Med.* 19, 1423–1437. doi: 10.1038/nm.3394

Rizvi, H., Sanchez-Vega, F., La, K., Chatila, W., Jonsson, P., Halpenny, D., et al. (2018). Molecular Determinants of Response to Anti-Programmed Cell Death (PD)-1 and Anti-Programmed Death-Ligand 1 (PD-L1) Blockade in Patients With Non-Small-Cell Lung Cancer Profiled With Targeted Next-Generation Sequencing. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 36, 633–641. doi: 10.1200/JCO.2017.75.3384

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 160, 48–61. doi: 10.1016/j.cell.2014.12.033

Saleh, R., and Elkord, E. (2019). Treg-mediated acquired resistance to immune checkpoint inhibitors. *Cancer Lett.* 457, 168–179. doi: 10.1016/j.canlet.2019.05.003

Schneider, W. M., Chevillotte, M. D., and Rice, C. M. (2014). Interferon-stimulated genes: a complex web of host defenses. *Annu. Rev. Immunol.* 32, 513–545. doi: 10.1146/annurev-immunol-032713-120231

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

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 23, 3251–3253. doi: 10.1093/bioinformatics/btm369

Sun, Y. (2016). Tumor microenvironment and cancer therapy resistance. *Cancer Lett.* 380, 205–215. doi: 10.1016/j.canlet.2015.07.044

Teo, M. Y., Seier, K., Ostrovnaya, I., Regazzi, A. M., Kania, B. E., Moran, M. M., et al. (2018). Alterations in DNA Damage Response and Repair Genes as Potential Marker of Clinical Benefit From PD-1/PD-L1 Blockade in Advanced Urothelial Cancers. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 36, 1685–1694. doi: 10.1200/JCO.2017.75.7740

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., et al. (2018). The Immune Landscape of Cancer. *Immunity* 48, 812–830.e14. doi: 10.1016/j.immuni.2018.03.023

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–77. doi: 10.5114/wo.2014.47136

Vanderwalde, A., Spetzler, D., Xiao, N., Gatalica, Z., and Marshall, J. (2018). Microsatellite instability status determined by next-generation sequencing and compared with PD-L1 and tumor mutational burden in 11,348 patients. *Cancer Med.* 7, 746–756. doi: 10.1002/cam4.1372

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

Wang, Z., Zhao, J., Wang, G., Zhang, F., Zhang, Z., Zhang, F., et al. (2018). Comutations in DNA Damage Response Pathways Serve as Potential Biomarkers for Immune Checkpoint Blockade. *Cancer Res.* 78, 6486–6496. doi: 10.1158/0008-5472.CAN-18-1814

Warth, A., Körner, S., Penzel, R., Muley, T., Dienemann, H., Schirmacher, P., et al. (2016). Microsatellite instability in pulmonary adenocarcinomas: a comprehensive study of 480 cases. *Virchows Arch.* 468, 313–319. doi: 10.1007/s00428-015-1892-7

Wu, T., and Dai, Y. (2017). Tumor microenvironment and therapeutic response. *Cancer Lett.* 387, 61–68. doi: 10.1016/j.canlet.2016.01.043

Yang, W., Bai, Y., Xiong, Y., Zhang, J., Chen, S., Zheng, X., et al. (2016). Potentiating the antitumour response of CD8(+) T cells by modulating cholesterol metabolism. *Nature* 531, 651–655. doi: 10.1038/nature17412

Yoshida, G. J. (2015). Metabolic reprogramming: the emerging concept and associated therapeutic strategies. *J. Exp. Clin. Cancer Res.* 34:111. doi: 10.1186/s13046-015-0221-y

Conflict of Interest: JZ, DL, and GL were employed by the company, HaploX Biotechnology.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 August 2021
doi: 10.3389/fgene.2021.723725

A Pan-Cancer Analysis of Transcriptome and Survival Reveals Prognostic Differentially Expressed LncRNAs and Predicts Novel Drugs for Glioblastoma Multiforme Therapy

Rongchuan Zhao[1,2], Xiaohan Sa[1,2], Nan Ouyang[1,2], Hong Zhang[3], Jiao Yang[2], Jinlin Pan[1,2], Jinhui Gu[4] and Yuanshuai Zhou[2]**

[1] Division of Life Sciences and Medicine, School of Biomedical Engineering (Suzhou), University of Science and Technology of China, Heifei, China, [2] Jiangsu Key Laboratory of Medical Optics, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China, [3] School of Life Sciences, Shanghai University, Shanghai, China, [4] Department of Anorectum, Suzhou Hospital of Traditional Chinese Medicine, Suzhou, China

OPEN ACCESS

Edited by:
*Chunjie Jiang,
University of Pennsylvania,
United States
Hongyi Zhang,
University of Texas Southwestern
Medical Center, United States*

Reviewed by:
*Guangli Suo,
Suzhou Institute of Nano-tech and
Nano-bionics, Chinese Academy of
Sciences (CAS), China*

***Correspondence:**
*Yuanshuai Zhou
zhouys@sibet.ac.cn
Jinhui Gu
dr.gjh@163.com*

Specialty section:
*This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics*

Received: *11 June 2021*
Accepted: *23 July 2021*
Published: *24 August 2021*

Citation:
*Zhao R, Sa X, Ouyang N, Zhang H,
Yang J, Pan J, Gu J and
Zhou Y (2021) A Pan-Cancer Analysis
of Transcriptome and Survival
Reveals Prognostic Differentially
Expressed LncRNAs and Predicts
Novel Drugs for Glioblastoma
Multiforme Therapy.
Front. Genet. 12:723725.
doi: 10.3389/fgene.2021.723725*

Numerous studies have identified various prognostic long non-coding RNAs (LncRNAs) in a specific cancer type, but a comprehensive pan-cancer analysis for prediction of LncRNAs that may serve as prognostic biomarkers is of great significance to be performed. Glioblastoma multiforme (GBM) is the most common and aggressive malignant adult primary brain tumor. There is an urgent need to identify novel therapies for GBM due to its poor prognosis and universal recurrence. Using available LncRNA expression data of 12 cancer types and survival data of 30 cancer types from online databases, we identified 48 differentially expressed LncRNAs in cancers as potential pan-cancer prognostic biomarkers. Two candidate LncRNAs were selected for validation in GBM. By the expression detection in GBM cell lines and survival analysis in GBM patients, we demonstrated the reliability of the list of pan-cancer prognostic LncRNAs obtained above. By constructing LncRNA-mRNA-drug network in GBM, we predicted novel drug-target interactions for GBM correlated LncRNA. This analysis has revealed common prognostic LncRNAs among cancers, which may provide insights into cancer pathogenesis and novel drug target in GBM.

Keywords: pan-cancer, long noncoding RNA, prognosis, biomarker, glioblastoma multiforme

INTRODUCTION

Non-coding RNAs (ncRNAs), including microRNA (miRNA), circRNA, long non-coding RNA (LncRNA), and many other kind of RNAs, are non-protein coding transcripts, which had been regarded as useless molecules, accounting for more than 95% of human genome (Tao et al., 2015). However, accumulating evidence indicates that ncRNAs have multiple functions in physiological and pathological processes, including cell growth, proliferation and apoptosis (Penna et al., 2015).

LncRNAs are a novel class of ncRNAs that are longer than 200 nucleotides, with no protein-coding capability (Ulitsky and Bartel, 2013). It has been shown that LncRNAs can elicit gene activation or suppression by interacting with proteins, DNAs and RNAs including miRNAs (Kataoka and Wang, 2014). They can also act as molecular signals, decoys, guides, and scaffolds for transcription factors and epigenetic modifiers (Wang and Chang, 2011).

A number of studies have revealed that LncRNAs are dysregulated in many cancer types. Several common LncRNAs have been investigated in cancers and the results revealed that they can function as potential biomarkers associated with tumor initiation, progression, and prognosis. For example, neuroblastoma associated transcript 1 (*NBAT1*) is demonstrated as a tumor-suppressing LncRNA and habitually downregulated in several cancers including neuroblastoma, osteosarcoma, ovarian cancer, and breast cancer. Loss of *NBAT1* induces tumor cell proliferation, differentiation, migration, and invasion through interaction with EZH2 and miR-21, or targeting ERK1/2- and AKT-mediated signaling pathway (Pandey et al., 2014; Hu et al., 2015; Yan et al., 2017; Yang et al., 2017). NBAT1 can also inhibit autophagy by suppressing the transcription of ATG7 in non-small cell lung cancer (Zheng et al., 2018). Colon cancer-associated transcript-1 (*CCAT1*) is found to be consistently elevated in multiple types of cancer and plays a critical role in various biological processes such as proliferation, invasion, migration, drug resistance, and survival (Nissan et al., 2012; He et al., 2014; Kim et al., 2014; Deng et al., 2015; Wang et al., 2019b). *CCAT1* has been demonstrated to enhance the expression of c-Myc (Xiang et al., 2014; Younger and Rinn, 2014). *CCAT1* can also stimulate EGFR expression, thereby activating MEK/ERK1/2 and PI3K/AKT signaling pathways (Jiang et al., 2018). Metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*) plays an important role in the pathogenesis and development of various cancers (Amodio et al., 2018; Liu et al., 2018; Zhao et al., 2018). Previous studies revealed that MALAT1 is upregulated in lung cancer, breast cancer, colorectal cancer, bladder cancer, and hepatocellular carcinoma (Goyal et al., 2021). *MALAT1* epigenetically repress TSC2 transcription *via* recruiting EZH2 to TSC2 promoter regions and thus enhances the apoptosis of cardiomyocytes through autophagy inhibition by regulating TSC2-mTOR signaling (Hu et al., 2019). Although increasing prognostic LncRNAs were identified exclusively in a specific cancer type, a comprehensive pan-cancer analysis is of great significance to be performed for prediction of LncRNAs that may serve as prognostic biomarkers. Identifying new prognostic LncRNA biomarkers is of extreme importance for revealing tumorigenesis underlying mechanisms. Although demonstrating LncRNA's exact function in cancers is difficult at present, it is possible to evaluate their role in prognosis, which is one of the main goals of cancer research.

Glioma is the most common malignant tumor in central nervous system and accounts for approximately 80% of primary intracranial tumors (Zhou et al., 2013). Based on World Health Organization (WHO) classification, glioma is classified into WHO grade I, II, III, and IV (Agnihotri et al., 2013). Among all types of glioma, glioblastoma multiforme (GBM) is the most aggressive type (a WHO grade IV glioma; Lieberman, 2017; Szopa et al., 2017), with a median survival of 15 months. GBM is characterized by chemoradiotherapy resistance and high risk of recurrence (Abbruzzese et al., 2017; Tian et al., 2019). Temozolomide (TMZ) resistance severely limits the efficacy and has become an important cause of poor prognosis. The 5-year recurrence for GBM is nearly universal. Therefore, there is an urgent need to identify novel therapies for GBM (Clarke et al., 2013; Szopa et al., 2017; Tan et al., 2018).

In our study, in order to get more rigorous analysis, we integrated both TANRIC database and ENCORI database, performed a step-by-step filtering and identified a list of 48 pan-cancer prognostic LncRNAs. Through LncRNA expression detection by qPCR and survival analysis on database in GBM, the reliability of our findings is confirmed. Previous pan-cancer analysis commonly aims to discover novel biomarkers across boundaries between tumor types (Weinstein et al., 2013). We took more concentrations on GBM, because there is few LncRNA targeted drugs for GBM therapy. We constructed an LncRNA-mRNA-drug interaction network to give advice to further drug-related LncRNA research and provide guideline for targeted therapeutics.

MATERIALS AND METHODS

Cell Lines

HUVEC, U87MG, and U251MG GBM cell lines were cultured in DMEM (Lot No. 8119284) supplemented with 10% fetal bovine serum (FBS, Lot No. 42G2095K) at 37°C in a humidified air atmosphere containing 5% CO_2. GSC23 GBM stem cell line was cultured in DMEM F-12 (Lot No. RNBG2219) supplemented with EGF (20 ng/ml, Lot No. PHG0311), bFGF (20 ng/ml, Lot No. PHG0368), B27 (1×, Lot No. 17504044), and NEAA (1×, Lot No. 11140050) at 37°C in a humidified air atmosphere containing 5% CO_2. DMEM F-12 was purchased from SIGMA, other reagents were purchased from Gibco.

Data Collection and Preprocessing

Gene expression data (LncRNA sequencing profiles) and corresponding clinical data of 12 cancer types were obtained from the Atlas of ncRNA in Cancer (TANRIC) database based on The Cancer Genome Atlas Data (TCGA) and Cancer Cell Line Encyclopedia (CCLE).[1] We focused on 12 types of cancers, each with more than 10 normal control samples, including bladder urothelial carcinoma (BLCA, 252 tumor samples and 19 normal samples), breast invasive carcinoma (BRCA, 837 tumor samples and 105 normal samples), head and neck squamous cell carcinoma (HSNC, 426 tumor samples and 42 normal samples), kidney chromophobe (KICH, 66 tumor samples and 25 normal samples), kidney renal clear cell carcinoma (KIRC, 448 tumor samples and 67 normal samples), kidney renal papillary cell carcinoma (KIRP, 198 tumor samples and

[1] https://ibl.mdanderson.org/tanric/_design/basic/main.html

30 normal samples), liver hepatocellular carcinoma (LIHC, 200 tumor samples and 50 normal samples), lung adenocarcinoma (LUAD, 488 tumor samples and 58 normal samples), lung squamous cell carcinoma (LUSC, 220 tumor samples and 17 normal samples), prostate adenocarcinoma (PRAD, 374 tumor samples and 52 normal samples), stomach adenocarcinoma (STAD, 285 tumor samples and 33 normal samples), and thyroid carcinoma (THCA, 497 tumor samples and 59 normal samples; **Table 1**). LncRNA ID was annotated according to GENCODE Release 29 (GRCh38.p12).[2]

Identification of Differentially Expressed LncRNAs in Pan-Cancer

The differentially expressed LncRNAs (DELncs) between tumor samples and normal samples were identified using DESeq2 package of R software. The value of p was adjusted by multiple significant tests with Bonferroni method. |log2 fold change (FC) | > 1 and $p < 0.05$ were set as the cutoff criteria. Hierarchical Cluster analysis was performed according to the expression values of DELncs. The heatmaps and volcano maps were plotted based on ggplot2 package of R software.

Survival Analysis of Differentially Expressed LncRNAs in Pan-Cancer

The survival data of the 30 TCGA cancer types and GBM were obtained from ENCORI Pan-Cancer Analysis Platform (Li et al., 2014) and TANRIC database, respectively. Besides the 12 types of cancers in **Table 1**, we also downloaded overall survival information of adrenocortical carcinoma (ACC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), esophageal carcinoma (ESCA), acute myeloid leukemia (LAML), brain lower grade glioma (LGG), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), skin cutaneous melanoma (SKCM), testicular germ cell tumors (TGCT), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC). The survival data of GBM was obtained from The Atlas of ncRNA in Cancer (TANRIC) based on TCGA and Cancer Cell Line Encyclopedia (CCLE). Patients were separated into higher and lower risk groups by median LncRNA expression. By Kaplan–Meier survival analysis, LncRNAs with Log-rank $p < 0.05$ were considered to be significantly associated with prognosis of patients.

Functional Enrichment Analysis of GO Annotation and KEGG Pathways

The Pearson correlation coefficient was used to evaluate co-expression relationship between LncRNA and mRNA. Cluster Profiler v3.8 package of R was used to analyze and visualize functional profiles [Gene Ontology, (GO) annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway] of the co-expressed genes with DELncs. The GO terms and KEGG pathways with $p < 0.05$ was considered as significantly enriched function terms or pathways.

Quantitative RT-PCR (qRT-PCR) Analysis

Total RNA from HUVEC, U87MG, U251MG, and GSC23 cell lines was isolated using RNAiso Plus (TaKaRa, code: 9109). RNA was transcribed to cDNA using PrimeScript™ RT Reagent Kit with gDNA Eraser (TaKaRa) following the manufacturer's instructions. Real-time quantitative PCR (qPCR) was performed using SYBR Green Real-time PCR Master Mix (TOYOBO, Lot No.857300) with primers against selected LncRNAs (primer sequences are listed in **Table 2**). Amplification and real time measurement of PCR products was performed with QuantStudio Real-Time PCR System (Thermo Fisher Scientific). The comparative Ct method was used to quantify the expression levels of LncRNAs. GAPDH gene expression served as an internal control.

Predicting lncRNA-mRNA-Drug Interactions for GBM

In the drug discovery and repositioning process, computational prediction of drug-target interactions (DTIs) plays a key role in identifying putative new drugs or novel targets for existing drugs (Savitski et al., 2014; Chernobrovkin et al., 2015; Franken et al., 2015; Cheng et al., 2016; Guney et al., 2016; Mehmood et al., 2016). Among multiple computational approaches, DTINET is a new computational pipeline, which can integrate heterogeneous information to predict new DTIs and repurpose existing drugs (Luo et al., 2017).

TABLE 1 | LncRNA expression data of 12 cancer types in TANRIC database.

Data source	Cancer type	Normal samples	Tumor samples
TCGA	BLCA	19	252
TCGA	BRCA	105	837
TCGA	HNSC	42	426
TCGA	KICH	25	66
TCGA	KIRC	67	448
TCGA	KIRP	30	198
TCGA	LIHC	50	200
TCGA	LUAD	58	488
TCGA	LUSC	17	220
TCGA	PRAD	52	374
TCGA	STAD	33	285
TCGA	THCA	59	497

TABLE 2 | Primer sequences of LINC0008 and RP11-399O19.9.

LncRNA	Primer
LINC00087	F: 5'-GGCTTGGCGGTTCGGCTGTC-3'
LINC00087	R: 5'-GCACTTGCAGGCGGACGTTGA-3'
RP11-399O19.9	F: 5'-CAGAAGTAGGGCAAGTTAGG-3'
RP11-399O19.9	R: 5'-CTCCACTGTCTTCCTCCC-3'

[2] https://www.gencodegenes.org/human/

RESULTS

The Integrative Pipeline for Identification of Pan-Cancer Prognostic DELncs

Figure 1 shows a scheme of the integrative pipeline containing multi-step of data integration and analysis for the identification of pan-cancer prognostic DELncs, together with its validation and application in GBM. First, we performed differential analysis on LncRNA expression profiles in TANRIC database and found 2,561 DELncs across 12 cancer types. Then we identified 161 of these overall DELncs as common DELncs because of their common changing trends in more than five cancer types. Based on the survival information in ENCORI database, we filtered out more than half of common DELncs with Log-rank $p < 0.05$ in less than six cancer types and acquired a list of 48 pan-cancer prognostic DELncs. Afterward, we validate the reliability of our list in GBM using both qPCR and database analysis. Finally, we construct an LncRNA-mRNA-drug network in GBM and predicted potential LncRNA associated drugs in GBM.

Comparison of Differentially Expressed LncRNAs in Pan-Cancer

To identify common DELncs in different cancer types, we compared the LncRNA expression profiles between tumor samples and paired normal samples in 12 cancer types from TCGA database, including BLCA, BRCA, HSNC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, STAD, and THCA. The result of differential analysis indicated that there were 2,561 DELncs across 12 cancer types altogether, where 859 DELncs were identified in KIRC samples and only 181 DELncs in STAD samples (**Figure 2A**). Among these 2,561 overall DELncs, we found that most of them showed similar tendency in more than one or two cancer types. Here we showed the top list of 10 most common DELncs (**Figure 2B**). The most representative up-regulated LncRNA is FGF14-AS2 (Ensembl ID: ENSG00000272143.1) and the most downregulated LncRNA is RP11-196G18.24 (Ensembl ID: ENSG00000272993.1). FGF14-AS2 was consistently upregulated in nine cancer types, including BRCA, KIRC, BLCA, LIHC, LUAD, LUSC, KICH, HNSC, and PRAD (**Figure 2C**). RP11-196G18.24 showed downregulation in almost all the cancer types except KIRC and THCA (**Figure 2D**).

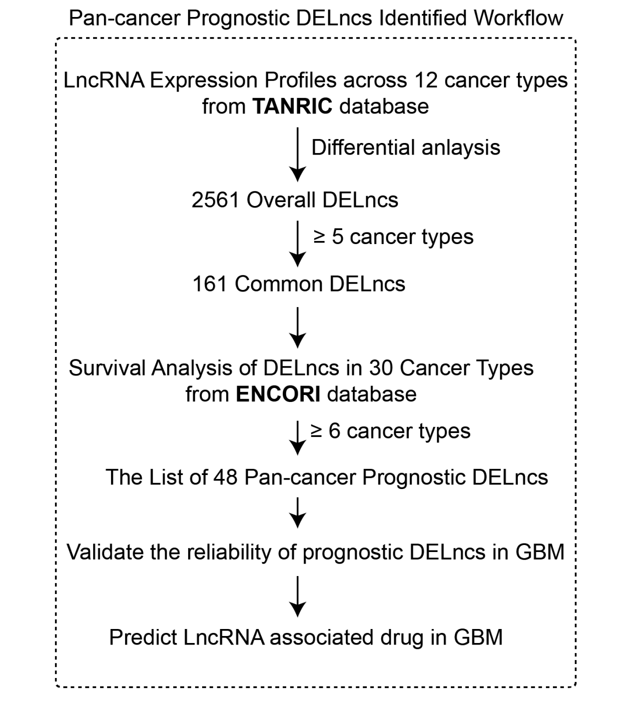

Pan-cancer Prognostic DELncs Identified Workflow
LncRNA Expression Profiles across 12 cancer types from **TANRIC** database
Differential anlaysis
2561 Overall DELncs
≥ 5 cancer types
161 Common DELncs
Survival Analysis of DELncs in 30 Cancer Types from **ENCORI** database
≥ 6 cancer types
The List of 48 Pan-cancer Prognostic DELncs
Validate the reliability of prognostic DELncs in GBM
Predict LncRNA associated drug in GBM

FIGURE 1 | The integrative pipeline for identification of pan-cancer prognostic DELncs.

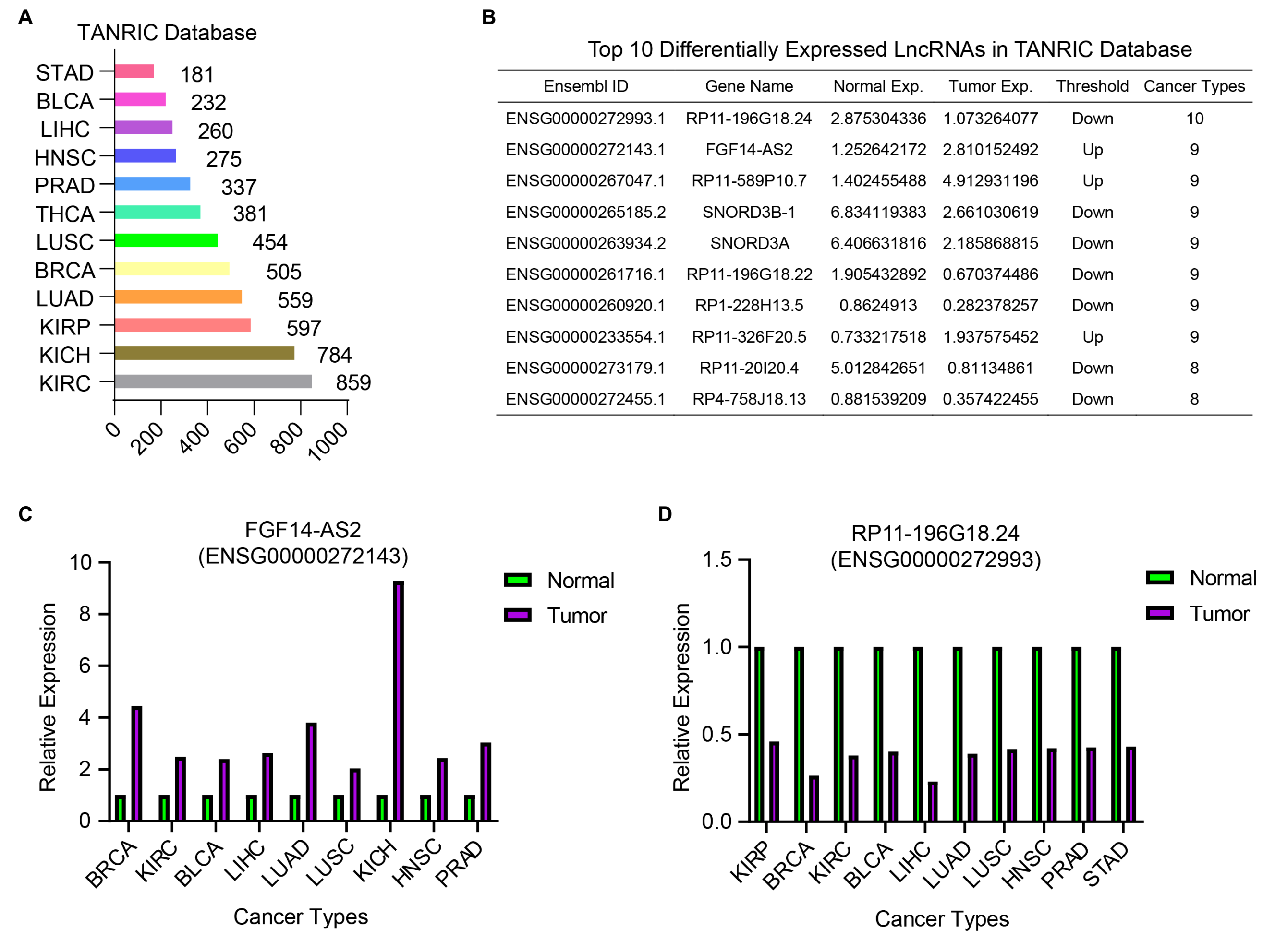
Survival Analysis and Functional Annotation of DELnc in Pan-Cancer

In order to acquire those LncRNAs that may serve as potential prognostic biomarkers of pan-cancer, we perform a step-by-step filtering (**Figure 3A**). In the initial loose screening step, 161 DELncs were selected due to their similar expression trends in more than five among 12 cancer types. Then the survival analysis in 30 cancer types in online tool ENCORI was performed to examine the relationship between 161 LncRNAs and the prognosis of cancer patients. In a more stringent step, 48 DELncs with Log-rank $p < 0.05$ in more than six cancer types were considered to be associated with prognosis of pan-cancer and were selected for further investigation (**Figure 3B**). In this way, we were able to take more LncRNAs in consideration and acquire a concise list of pan-cancer prognotic DELncs for further research.

Cluster analysis and heatmap were performed according to the value of p of Log-rank of the overall survival analysis of these 48 LncRNAs in 30 cancer types (**Figure 3B**). Among the 48 candidates, the top one LncRNA *MIR4435-2HG* (Ensembl ID: ENSG00000172965.10) was significantly associated with the 10 cancer types (Log-rank p value < 0.05), including ACC, COAD, HNSC, KIRC, KIRP, LGG, LIHC, LUAD, MESO, and PAAD. Kaplan–Meier survival estimate in ENCORI Pan-Cancer Analysis Platform revealed that higher expression of *MIR4435-2HG* in 10 cancer types was robustly associated with worse prognosis (**Figure 3C**). In order to uncover the biological functions of *MIR4435-2HG*, we performed the GO annotation and KEGG pathway enrichment analysis. As shown in **Figures 3D,E**, the *MIR4435-2HG* co-expressed genes were associated with the category of morphogenesis of an epithelium, regulation of protein complex assembly, Wnt signaling pathway, and cell-substrate adhesion (**Figure 3D**). KEGG pathway enrichment analysis revealed that the genes associated with *MIR4435-2HG* were mainly enriched in focal adhesion, leukocyte trans-endothelial migration, and pathway in cancer (**Figure 3E**).

Evaluation of DELncs' Expression and Prognostic Value in GBM

To further validate the expression of 48 DELncs from Pan-cancer analysis, we selected two (RP11-399O19.9



Top 10 Differentially Expressed LncRNAs in TANRIC Database

Ensembl ID	Gene Name	Normal Exp.	Tumor Exp.	Threshold	Cancer Types
ENSG00000272993.1	RP11-196G18.24	2.875304336	1.073264077	Down	10
ENSG00000272143.1	FGF14-AS2	1.252642172	2.810152492	Up	9
ENSG00000267047.1	RP11-589P10.7	1.402455488	4.912931196	Up	9
ENSG00000265185.2	SNORD3B-1	6.834119383	2.661030619	Down	9
ENSG00000263934.2	SNORD3A	6.406631816	2.185868815	Down	9
ENSG00000261716.1	RP11-196G18.22	1.905432892	0.670374486	Down	9
ENSG00000260920.1	RP1-228H13.5	0.8624913	0.282378257	Down	9
ENSG00000233554.1	RP11-326F20.5	0.733217518	1.937575452	Up	9
ENSG00000273179.1	RP11-20I20.4	5.012842651	0.81134861	Down	8
ENSG00000272455.1	RP4-758J18.13	0.881539209	0.357422455	Down	8

FIGURE 2 | Differentially expressed LncRNAs analysis in 12 TCGA cancer types between tumor samples and normal samples. **(A)** The number of identified DELncs in each cancer type in TANRIC database. **(B)** A list of top 10 common DELncs in 12 cancer types. **(C)** The relative expression of FGF14-AS2 in BRCA, KIRC, BLCA, LIHC, LUAD, LUSC, KICH, HNSC, and PRAD samples compared with normal samples. **(D)** The relative expression of RP11-196G18.24 in KIRP, BRCA, KIRC, BLCA, LIHC, LUAD, LUSC, HNSC, PRAD, and STAD samples compared with normal samples.

and LINC00087) of them that have not been well studied to assess the reliability of differentially expression analysis in pan-cancer. Here, we chose GBM as our validation set. GBM is highly aggressive grade 4 glioma and is the most common type of malignant glioma, with 10,000 new diagnoses each year. However, there were few LncRNA revealed to be associated with the diagnosis and prognosis of GBM.

The relative expression of RP11-399O19.9 (Ensembl ID: ENSG00000261438.1) and LINC00087 (Ensembl ID: ENSG00000196972.6) in U87MG, U251MG, and GSC23 was up regulated compared with normal cell line HUVEC (**Figures 4A,B**). GO annotation was performed to predict the potential biological processes of RP11-399O19.9 and LINC00087. Based on TANRIC database, RP11-399O19.9 co-expressed genes were correlated with the categorical terms of neutrophil activation, neutrophil degranulation, and neutrophil activation involved in immune response and many other processes of immune system (**Figure 4C**). This indicated that RP11-399O19.9 may play an important role in the regulation of immune system, especially neutrophil activation. Genes associated with LINC00087 were enriched in the modulation of chemical synaptic transmission, the regulation of trans-synaptic signaling, and synaptic vesicle cycle (**Figure 4D**), indicating that LINC00087 might be involved in intercellular signal transmission.

Finally, to investigate the prognostic significance of RP11-399O19.9 and LINC00087 expression in GBM patients, we obtained the overall survival information from TANRIC database. The Kaplan–Meier survival analysis showed that RP11-399O19.9 and LINC00087 are able to separate patients into higher and lower risk groups by median PI, with the value of p of Log-rank of 0.043875 and 0.024661 for GBM. High expression of both RP11-399O19.9 and LINC00087 are significantly associated with poor prognosis for GBM patients (Log-rank $p < 0.05$, **Figures 4E,F**).

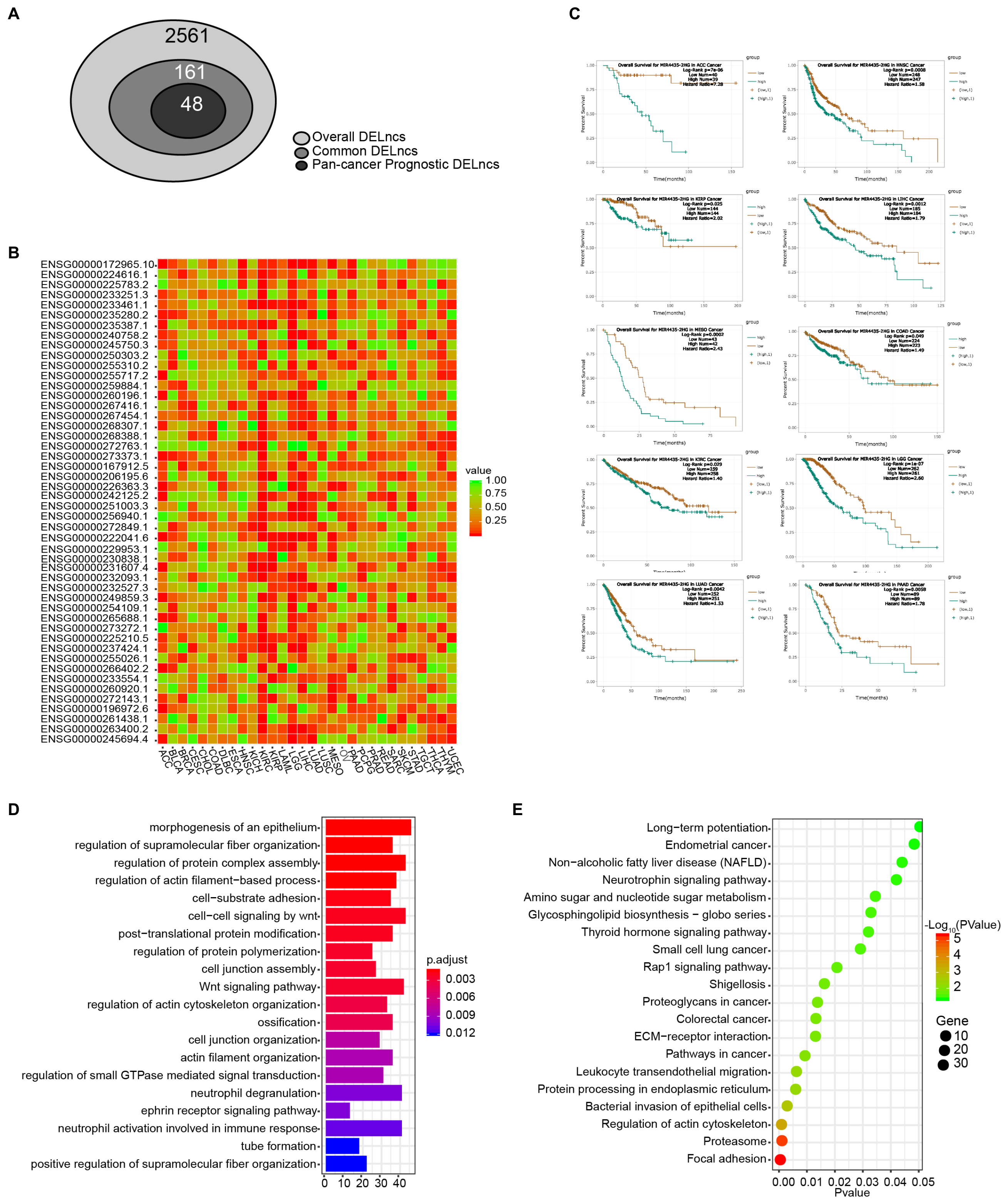


FIGURE 3 | Survival analysis and functional annotation of DELncs. **(A)** Identification of common DELncs and pan-cancer prognostic DELncs. **(B)** Clustering and heatmap of 48 DELncs' prognostic value by the value of *p* of Log-rank survival analysis in 30 cancer types. **(C)** Kaplan–Meier survival analysis of MIR4435-2HG in ACC, COAD, HNSC, KIRC, KIRP, LGG, LIHC, LUAD, MESO, and PAAD. **(D)** Gene Ontology (GO) annotation of MIR4435-2HG by its correlation mRNA expression. **(E)** Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of MIR4435-2HG by its correlation mRNA expression.

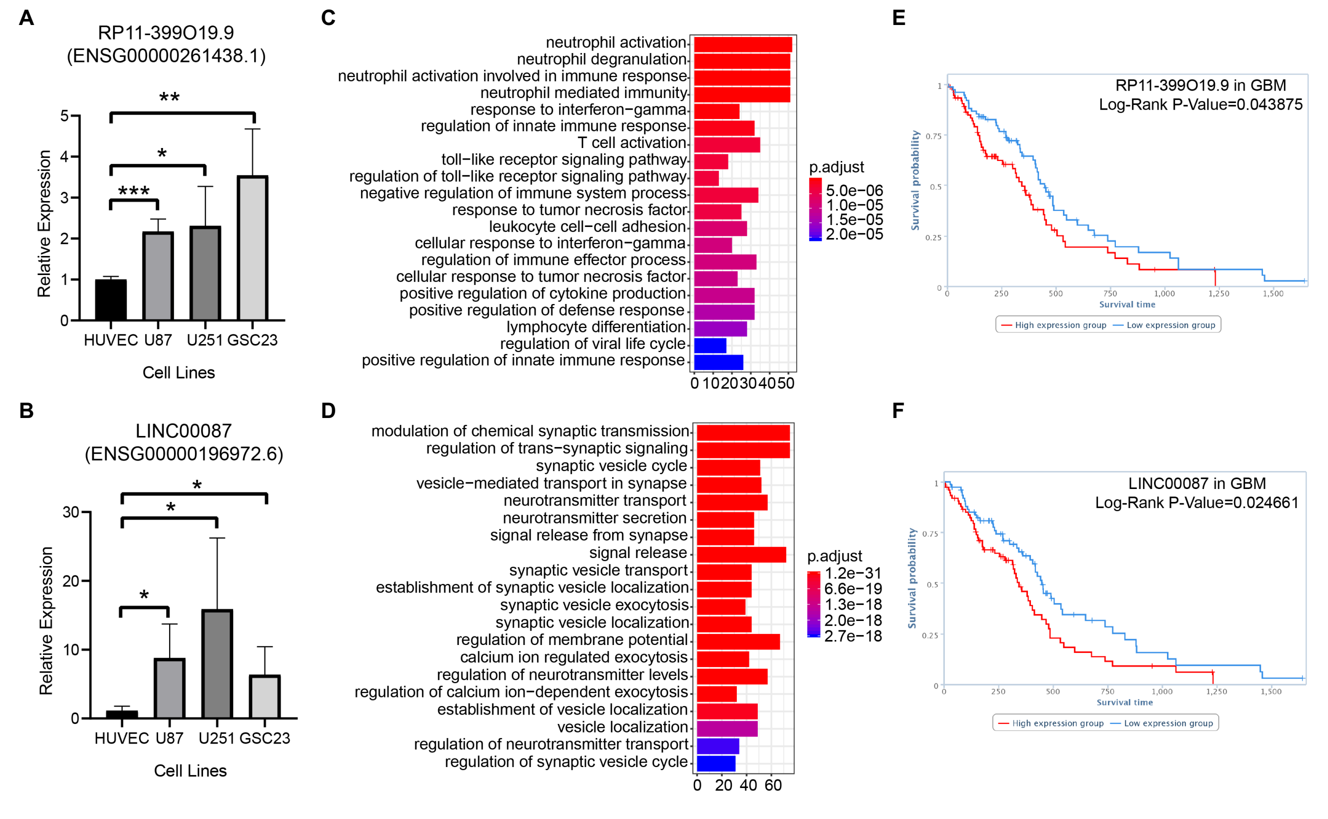


FIGURE 4 | Assess the reliability of candidate LncRNAs by the examination of RP11-399O19.9 and LINC00087 in GBM. **(A)** mRNA expression of RP11-399O19.9 and **(B)** LINC00087 in HUVEC, U87MG, U251MG, and GSC23 cell lines (*means $p < 0.05$, **means $p < 0.01$, and ***means $p < 0.001$). **(C)** GO annotation of RP11-399O19.9 and **(D)** LINC00087. **(E)** Kaplan–Meier survival analysis of RP11-399O19.9 and **(F)** LINC00087 in GBM patients.

Prediction of LncRNAs as Novel Targets for Existing Drugs in GBM Therapy

Based on the list of top 150 novel drug-target interactions (**Supplementary Table 1**) predicted by DTINet, we constructed an LncRNA-mRNA-drug network to identify the GBM correlated drugs. GBM related LncRNA-mRNA-drug network was visualized using Cytoscape software (Version 3.7.1; **Figure 5**). In the network, LINC00087 (Ensembl ID: ENSG00000196972.6) has the most interactions with a large amount of drug targets (mRNA), showing the most relevant with existing drugs including Clozapine, Zolmitriptan, Bethanechol, etc. We find several Extrasynaptic γ-aminobutyric acid type A (GABAA) receptors family (GABR) targets, which have interaction with LINC00087 including GABRA1, GABRB2, GABRD, GABRG1,GABRG2, and GABRG3. GABAA receptor family contributes to memory performance. Dysregulation of GABAA receptor expression, which occurs in some neurological disorders, is associated with memory impairment (Whissell et al., 2016). Their related drug, Clozapine (CZP), a dibenzodiazepine atypical antipsychotic drug, was introduced for treatment of schizophrenia in Europe in 1971, rapidly gaining popularity due to its efficacy and virtual absence of extrapyramidal side effects (Mijovic and MacCabe, 2020). Clozapine may be a potential drug for GBM treatment.

DISCUSSION

Evaluating prognostic value of factors associated with tumorigenesis and progression is an important part of cancer research. Numerous studies have demonstrated that many factors have implication in tumor progression or clinical prognosis in pan-cancer including gene expression, DNA methylation, mutation, etc. Although several LncRNAs have been identified as diagnostic or prognostic markers (Prensner et al., 2011; Sun et al., 2013), a pan-cancer analysis of prognostic LncRNA has rarely been performed. At the same time, there are variations across different cancers in terms of prognosis related LncRNAs, which leads to inconvenience in its utility in clinical oncology. In this study, we analyzed LncRNA expression profiles of 4,848 samples from 12 TCGA cancer types in TANRIC database. We systematically analyzed DELncs between tumor and normal samples in each cancer type and found 2,561 LncRNAs that were simultaneously dysregulated in 12 cancer types. Afterward, we evaluated the prognostic effect of 161 LncRNAs in 30 cancer type and ultimately identified 48 DELncs as our pan-cancer prognostic LncRNAs. *MIR4435-2HG*, as one of the 48 DELncs, showed prognostic importance in 10 cancer types. Previous studies have demonstrated that upregulation of *MIR4435-2HG* is associated with bad prognosis of patients with prostate carcinoma (Zhang et al., 2019), breast cancer (Deng et al., 2016), gastric cancer

(Wang et al., 2019a), lung cancer (Qian et al., 2018), and colorectal cancer (Ouyang et al., 2019). Consistent with these studies, the functional annotation of *MIR4435-2HG* in our study indicates that it may play a leading role in cancer cell metastasis and invasion and thus leads to bad prognosis of patients. Although many other LncRNAs in our list have not been demonstrated to have association with tumorigenesis, the analysis we performed above is helpful to predict LncRNAs as cancer markers and may provide directions in cancer research.

Evaluating gene expression in cancer cell lines and association with patients' prognosis are common methods in cancer research. We did not perform differential expression analysis of LncRNAs in GBM in the first part of results because of lacking LncRNA expression of normal samples of GBM in TANRIC database. We also wanted to acquire common differentially expressed LncRNAs that can give advice to multiple cancer therapies and drug discoveries through pan-cancer analysis. By selecting two of DELncs, detecting their expression in GBM cell lines and analyzing prognosis of GBM patients, we were able to validate the reliability of our 48-DELncs-list. RP11-399O19.9 and LINC00087 have not been well studied, but their dysregulated expressions and prognostic values intimate their importance in tumorigenesis and prognosis. Even if lacking LncRNA transcriptome profile of normal samples of GBM in TCGA database, this study provides a novel method of LncRNA research in GBM.

GBM is considered as incurable intracranial malignant tumor, with a median survival of 15 months following aggressive combination of therapies including maximal-safe surgical resection, adjuvant radiation therapy (RT) with concurrent, and adjuvant temozolomide (TMZ) treatment (Stupp et al., 2009). However, TMZ resistance severely limits the efficacy and has become an important cause of poor prognosis. As TMZ is the only chemotherapy drug available for GBM, it is urgent to look for new drugs or repurpose existing drugs for GBM. Several previous studies indicate that LncRNA may play an important role in GBM. *HOTAIR* could promote glioblastoma cell cycle progression (Zhang et al., 2015); *FOXM1-AS* could enhance self-renewal and tumorigenesis of glioblastoma stem-like cells (Zhang et al., 2017); *H19* could promote glioblastoma cell invasion, angiogenesis, and tube formation (Jia et al., 2016); *MALAT1* could decrease the sensitivity of glioblastoma cells to TMZ (Li et al., 2017). Althogh many LncRNAs have been identified as biomarkers of GBM, there is few LncRNA targeted drugs for GBM therapy. The future of LncRNA-based drug discovery is bright. However, it is still an emerging concept and strategy compared with the traditional drug targets and proteins (Chen et al., 2021). Target selection is a key element of drug development; therefore, identifying the most potential LncRNAs is the first step and the most important process. Further advances in LncRNA-targeted drugs are clearly dependent on the in-depth basic research into the function and mechanisms of LncRNAs. Our study provided a list of 48 LncRNAs by differential expression analysis and survival analysis in pan cancer, which will give advice to the selection of the most potential LncRNAs for further in-depth basic research and LncRNA-based drug discovery. Computational prediction of drug-target interactions (DTIs) is a useful tool for researchers to identify new drugs or novel targets for existing drugs. According to prognostic LncRNA candidates identified, we built LncRNA-mRNA-drug interaction network, which may be beneficial in the treatment of GBM. LINC00087 and Clozapine might be the most valuable LncRNA target and drug in our network for GBM therapy, respectively. In addition, Clomipramine is one of the most widely used tricyclic antidepressants in Western Europe (Balant-Gorgia et al., 1991). Flumazenil appears to act at CNS. It is an antagonist synthesized to competitively block the effects of benzodiazepines on GABAergic pathway-mediated inhibition in the CNS (Votey et al., 1991). Ziprasidone is a recently approved atypical antipsychotic agent (available in oral and short-acting intramuscular formulations) effective in the treatment of schizophrenia in an outpatient setting and in the treatment of acute psychotic episodes (Beedham et al., 2003). These drugs that have been proved to be effective in the treatment of CNS diseases may be effective in GBM treatment.

Since 2012, multiple efforts have launched toward TCGA pan-cancer analysis across many different tumor types (Han et al., 2014; Ching et al., 2016; Luo et al., 2019; Cui et al., 2020). They mainly focused on the mutational landscape (Kandoth et al., 2013). The aim of TCGA pan-cancer initiative is to discover novel intervention strategies, such as discovering novel biomarkers among different tumor samples (Weinstein et al., 2013; Danaher et al., 2018; Gobin et al., 2019). These studies did not make efforts to the LncRNA-based drug discovery. Our research integrated pan cancer analysis with the computational prediction of drug-target interactions together to get 48 DELnc list and its related drugs, which will be of value to both prognostic comments and drug discovery. In our study, we took both LncRNA expression level and prognostic value in consideration and identified a list of 48 pan-cancer prognostic LncRNAs by referring to previous studies. To ensure the reliability of our findings, we validated it in GBM in two aspects: expression level detection by QPCR and survival analysis based on database. We identified these LncRNA not only as biomarkers of pan-cancer but also as novel targets of existing drugs because of their interaction with mRNAs.

In summary, this study provided a list of 48 LncRNAs by differential expression analysis and survival analysis, together with the LncRNA-mRNA-drug interaction network in GBM. The findings also highlighted the prognostic value of LncRNA in pan-cancer research and provided a new perspective for GBM drug target identification. Although we have identified many potential prognostic LncRNAs in multiple cancer types, further research is needed for the evaluation of their function in cancers. Despite limitations of current work, it is a good way to integrate clinical information into LncRNA research in pan-cancer to seek for potential LncRNA targets of cancer therapy and further studies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

RZ: conceptualization, methodology, formal analysis, writing – original draft, and visualization. XS: formal analysis and data curation. NO and HZ: visualization. JY and JP: writing – review and editing. JG: writing – review and editing and supervision. YZ: writing – review and editing, supervision, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (81903028), China Postdoctoral Science Foundation (2020M671599), and the Jiangsu Province Natural Science Foundation for Youths (BK20180223).

ACKNOWLEDGMENTS

We thank all the members of Minxuan Sun laboratory for the helpful suggestions and critical reading of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.723725/full#supplementary-material

REFERENCES

Abbruzzese, C., Matteoni, S., Signore, M., Cardone, L., Nath, K., Glickson, J. D., et al. (2017). Drug repurposing for the treatment of glioblastoma multiforme. *J. Exp. Clin. Cancer Res.* 36:169. doi: 10.1186/s13046-017-0642-x

Agnihotri, S., Burrell, K. E., Wolf, A., Jalali, S., Hawkins, C., Rutka, J. T., et al. (2013). Glioblastoma, a brief review of history, molecular genetics, animal models and novel therapeutic strategies. *Arch. Immunol. Ther. Exp.* 61, 25-41. doi: 10.1007/s00005-012-0203-0

Amodio, N., Raimondi, L., Juli, G., Stamato, M. A., Caracciolo, D., Tagliaferri, P., et al. (2018). MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. *J. Hematol. Oncol.* 11:63. doi: 10.1186/s13045-018-0606-4

Balant-Gorgia, A. E., Gex-Fabry, M., and Balant, L. P. (1991). Clinical pharmacokinetics of clomipramine. *Clin. Pharmacokinet.* 20, 447-462. doi: 10.2165/00003088-199120060-00002

Beedham, C., Miceli, J. J., and Obach, R. S. (2003). Ziprasidone metabolism, aldehyde oxidase, and clinical implications. *J. Clin. Psychopharmacol.* 23, 229-232. doi: 10.1097/01.jcp.0000084028.22282.f2

Chen, Y., Li, Z., Chen, X., and Zhang, S. (2021). Long non-coding RNAs: from disease code to drug role. *Acta Pharm. Sin. B* 11, 340-354. doi: 10.1016/j.apsb.2020.10.001

Cheng, F., Zhao, J., Fooksa, M., and Zhao, Z. (2016). A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes. *J. Am. Med. Inform. Assoc.* 23, 681-691. doi: 10.1093/jamia/ocw007

Chernobrovkin, A., Marin-Vicente, C., Visa, N., and Zubarev, R. A. (2015). Functional identification of target by expression proteomics (FITExP) reveals protein targets and highlights mechanisms of action of small molecule drugs. *Sci. Rep.* 5:11176. doi: 10.1038/srep11176

Ching, T., Peplowska, K., Huang, S., Zhu, X., Shen, Y., Molnar, J., et al. (2016). Pan-cancer analyses reveal long intergenic non-coding RNAs relevant to tumor diagnosis, subtyping and prognosis. *EBioMedicine* 7, 62–72. doi: 10.1016/j.ebiom.2016.03.023

Clarke, J., Penas, C., Pastori, C., Komotar, R. J., Bregy, A., Shah, A. H., et al. (2013). Epigenetic pathways and glioblastoma treatment. *Epigenetics* 8, 785-795. doi: 10.4161/epi.25440

Cui, X., Zhang, X., Liu, M., Zhao, C., Zhang, N., Ren, Y., et al. (2020). A pan-cancer analysis of the oncogenic role of staphylococcal nuclease domain-containing protein 1 (SND1) in human tumors. *Genomics* 112, 3958-3967. doi: 10.1016/j.ygeno.2020.06.044

Danaher, P., Warren, S., Lu, R., Samayoa, J., Sullivan, A., Pekker, I., et al. (2018). Pan-cancer adaptive immune resistance as defined by the tumor inflammation signature (TIS): results from The Cancer genome atlas (TCGA). *J. Immunother. Cancer* 6:63. doi: 10.1186/s40425-018-0367-1

Deng, L. L., Chi, Y. Y., Liu, L., Huang, N. S., Wang, L., and Wu, J. (2016). LINC00978 predicts poor prognosis in breast cancer patients. *Sci. Rep.* 6:37936. doi: 10.1038/srep37936

Deng, L., Yang, S. B., Xu, F. F., and Zhang, J. H. (2015). Long noncoding RNA CCAT1 promotes hepatocellular carcinoma progression by functioning as let-7 sponge. *J. Exp. Clin. Cancer Res.* 34:18. doi: 10.1186/s13046-015-0136-7

Franken, H., Mathieson, T., Childs, D., Sweetman, G. M., Werner, T., Togel, I., et al. (2015). Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protoc.* 10, 1567-1593. doi: 10.1038/nprot.2015.101

Gobin, E., Bagwell, K., Wagner, J., Mysona, D., Sandirasegarane, S., Smith, N., et al. (2019). A pan-cancer perspective of matrix metalloproteases (MMP) gene expression profile and their diagnostic/prognostic potential. *BMC Cancer* 19:581. doi: 10.1186/s12885-019-5768-0

Goyal, B., Yadav, S. R. M., Awasthee, N., Gupta, S., Kunnumakkara, A. B., and Gupta, S. C. (2021). Diagnostic, prognostic, and therapeutic significance of long non-coding RNA MALAT1 in cancer. *Biochim. Biophys. Acta Rev. Cancer* 1875:188502. doi: 10.1016/j.bbcan.2021.188502

Guney, E., Menche, J., Vidal, M., and Barabasi, A. L. (2016). Network-based in silico drug efficacy screening. *Nat. Commun.* 7:10331. doi: 10.1038/ncomms10331

Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J., Edgerton, M. E., et al. (2014). The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat. Commun.* 5:3963. doi: 10.1038/ncomms4963

He, X., Tan, X., Wang, X., Jin, H., Liu, L., Ma, L., et al. (2014). C-Myc-activated long noncoding RNA CCAT1 promotes colon cancer cell proliferation and invasion. *Tumour Biol.* 35, 12181-12188. doi: 10.1007/s13277-014-2526-4

Hu, P., Chu, J., Wu, Y., Sun, L., Lv, X., Zhu, Y., et al. (2015). NBAT1 suppresses breast cancer metastasis by regulating DKK1 via PRC2. *Oncotarget* 6, 32410-32425. doi: 10.18632/oncotarget.5609

Hu, H., Wu, J., Yu, X., Zhou, J., Yu, H., and Ma, L. (2019). Long non-coding RNA MALAT1 enhances the apoptosis of cardiomyocytes through autophagy inhibition by regulating TSC2-mTOR signaling. *Biol. Res.* 52:58. doi: 10.1186/s40659-019-0265-0

Jia, P., Cai, H., Liu, X., Chen, J., Ma, J., Wang, P., et al. (2016). Long non-coding RNA H19 regulates glioma angiogenesis and the biological behavior of glioma-associated endothelial cells by inhibiting microRNA-29a. *Cancer Lett.* 381, 359-369. doi: 10.1016/j.canlet.2016.08.009

Jiang, Y., Jiang, Y. Y., Xie, J. J., Mayakonda, A., Hazawa, M., Chen, L., et al. (2018). Co-activation of super-enhancer-driven CCAT1 by TP63 and SOX2 promotes squamous cancer progression. *Nat. Commun.* 9:3619. doi: 10.1038/s41467-018-06081-9

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, 333-339. doi: 10.1038/nature12634

Kataoka, M., and Wang, D. Z. (2014). Non-coding RNAs including miRNAs and lncRNAs in cardiovascular biology and disease. *Cell* 3, 883-898. doi: 10.3390/cells3030883

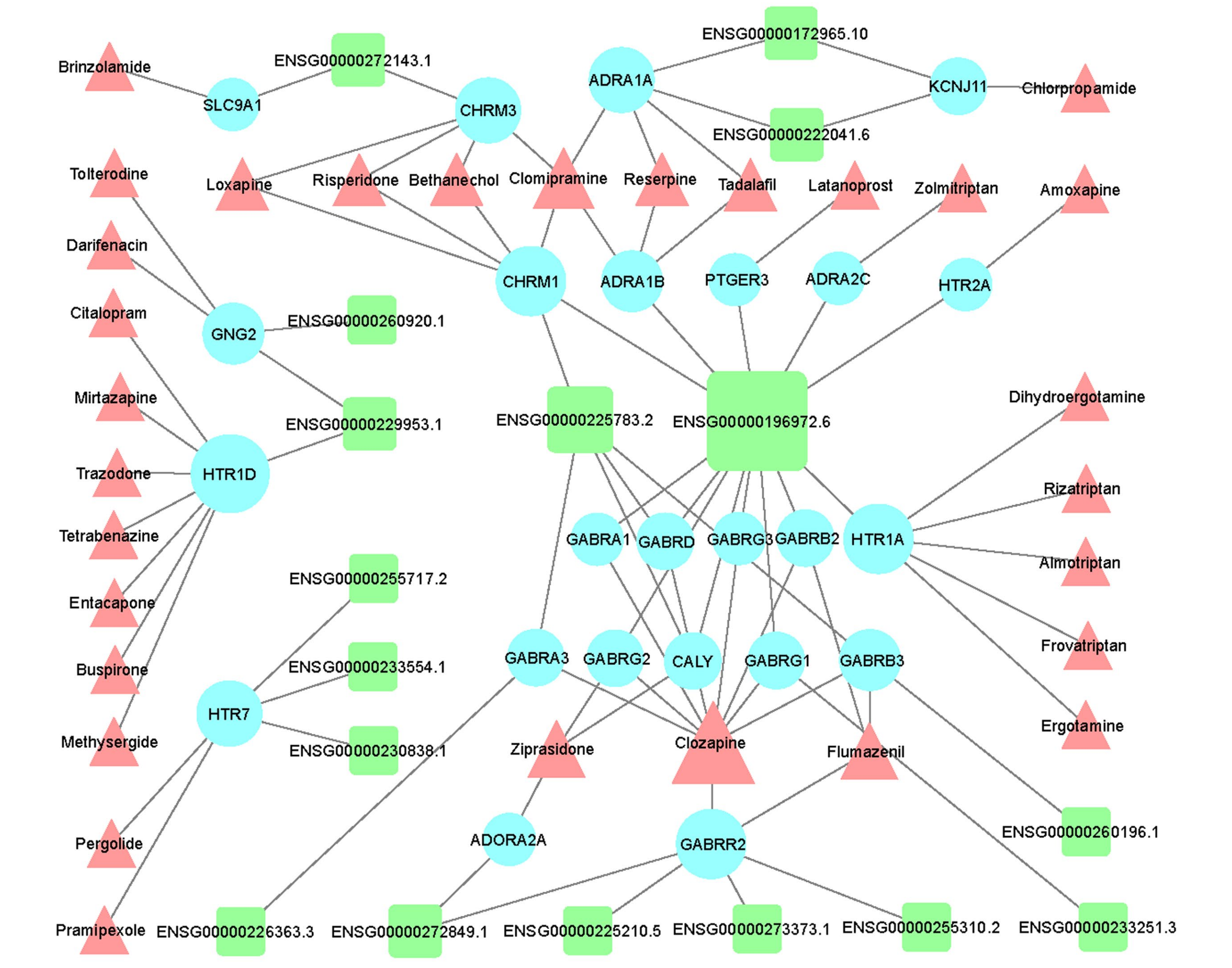


FIGURE 5 | Network visualization of the LncRNA-mRNA-drug interactions in GBM. Visualization of interaction network between differentially expressed LncRNAs and top 150 drug-target interactions (DTIs) predicted by DTINET. LncRNAs, targets (mRNAs), and drugs are shown in green round rectangles, blue circles and pink triangles, respectively.

Kim, T., Cui, R., Jeon, Y. J., Lee, J. H., Lee, J. H., Sim, H., et al. (2014). Long-range interaction and correlation between MYC enhancer and oncogenic long noncoding RNA CARLo-5. *Proc. Natl. Acad. Sci. U. S. A.* 111, 4173-4178. doi: 10.1073/pnas.1400350111

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, H., Yuan, X., Yan, D., Li, D., Guan, F., Dong, Y., et al. (2017). Long non-coding RNA MALAT1 decreases the sensitivity of resistant glioblastoma cell lines to temozolomide. *Cell. Physiol. Biochem.* 42, 1192-1201. doi: 10.1159/000478917

Lieberman, F. (2017). Glioblastoma update: molecular biology, diagnosis, treatment, response assessment, and translational clinical trials. *F1000Res* 6:1892. doi: 10.12688/f1000research.11493.1

Liu, Y., Du, Y., Hu, X., Zhao, L., and Xia, W. (2018). Up-regulation of ceRNA TINCR by SP1 contributes to tumorigenesis in breast cancer. *BMC Cancer* 18:367. doi: 10.1186/s12885-018-4255-3

Luo, Z., Wang, W., Li, F., Songyang, Z., Feng, X., Xin, C., et al. (2019). Pan-cancer analysis identifies telomerase-associated signatures and cancer subtypes. *Mol. Cancer* 18:106. doi: 10.1186/s12943-019-1035-x

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8:573. doi: 10.1038/s41467-017-00680-8

Mehmood, S., Marcoux, J., Gault, J., Quigley, A., Michaelis, S., Young, S. G., et al. (2016). Mass spectrometry captures off-target drug binding and provides mechanistic insights into the human metalloprotease ZMPSTE24. *Nat. Chem.* 8, 1152-1158. doi: 10.1038/nchem.2591

Mijovic, A., and MacCabe, J. H. (2020). Clozapine-induced agranulocytosis. *Ann. Hematol.* 99, 2477-2482. doi: 10.1007/s00277-020-04215-y

Nissan, A., Stojadinovic, A., Mitrani-Rosenbaum, S., Halle, D., Grinbaum, R., Roistacher, M., et al. (2012). Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. *Int. J. Cancer* 130, 1598-1606. doi: 10.1002/ijc.26170

Ouyang, W., Ren, L., Liu, G., Chi, X., and Wei, H. (2019). LncRNA MIR4435-2HG predicts poor prognosis in patients with colorectal cancer. *PeerJ* 7:e6683. doi: 10.7717/peerj.6683

Pandey, G. K., Mitra, S., Subhash, S., Hertwig, F., Kanduri, M., Mishra, K., et al. (2014). The risk-associated long noncoding RNA NBAT-1 controls neuroblastoma progression by regulating cell proliferation and neuronal differentiation. *Cancer Cell* 26, 722-737. doi: 10.1016/j.ccell.2014.09.014

Penna, E., Orso, F., and Taverna, D. (2015). miR-214 as a key hub that controls cancer networks: small player, multiple functions. *J. Invest. Dermatol.* 135, 960-969. doi: 10.1038/jid.2014.479

Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* 29, 742-749. doi: 10.1038/nbt.1914

Qian, H., Chen, L., Huang, J., Wang, X., Ma, S., Cui, F., et al. (2018). The lncRNA MIR4435-2HG promotes lung cancer progression by activating beta-catenin signalling. *J. Mol. Med.* 96, 753-764. doi: 10.1007/s00109-018-1654-5

Savitski, M. M., Reinhard, F. B., Franken, H., Werner, T., Savitski, M. F., Eberhard, D., et al. (2014). Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* 346:1255784. doi: 10.1126/science.1255784

Stupp, R., Hegi, M. E., Mason, W. P., van den Bent, M. J., Taphoorn, M. J., Janzer, R. C., et al. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* 10, 459-466. doi: 10.1016/S1470-2045(09)70025-7

Sun, K., Chen, X., Jiang, P., Song, X., Wang, H., and Sun, H. (2013). iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* 14(Suppl. 2):S7. doi: 10.1186/1471-2164-14-s2-s7

Szopa, W., Burley, T. A., Kramer-Marek, G., and Kaspera, W. (2017). Diagnostic and therapeutic biomarkers in glioblastoma: current status and future perspectives. *Biomed. Res. Int.* 2017:8013575. doi: 10.1155/2017/8013575

Tan, S. K., Pastori, C., Penas, C., Komotar, R. J., Ivan, M. E., Wahlestedt, C., et al. (2018). Serum long noncoding RNA HOTAIR as a novel diagnostic and prognostic biomarker in glioblastoma multiforme. *Mol. Cancer* 17:74. doi: 10.1186/s12943-018-0822-0

Tao, L., Bei, Y., Zhou, Y., Xiao, J., and Li, X. (2015). Non-coding RNAs in cardiac regeneration. *Oncotarget* 6, 42613-42622. doi: 10.18632/oncotarget.6073

Tian, Y., Zheng, Y., and Dong, X. (2019). AGAP2-AS1 serves AS an oncogenic lncRNA and prognostic biomarker in glioblastoma multiforme. *J. Cell. Biochem.* 120, 9056-9062. doi: 10.1002/jcb.28180

Ulitsky, I., and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26-46. doi: 10.1016/j.cell.2013.06.020

Votey, S. R., Bosse, G. M., Bayer, M. J., and Hoffman, J. R. (1991). Flumazenil: a new benzodiazepine antagonist. *Ann. Emerg. Med.* 20, 181–188. doi: 10.1016/S0196-0644(05)81219-3

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904-914. doi: 10.1016/j.molcel.2011.08.018

Wang, H., Wu, M., Lu, Y., He, K., Cai, X., Yu, X., et al. (2019a). LncRNA MIR4435-2HG targets desmoplakin and promotes growth and metastasis of gastric cancer by activating Wnt/beta-catenin signaling. *Aging* 11, 6657-6673. doi: 10.18632/aging.102164

Wang, N., Yu, Y., Xu, B., Zhang, M., Li, Q., and Miao, L. (2019b). Pivotal prognostic and diagnostic role of the long noncoding RNA colon cancerassociated transcript 1 expression in human cancer (review). *Mol. Med. Rep.* 19, 771-782. doi: 10.3892/mmr.2018.9721

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113-1120. doi: 10.1038/ng.2764

Whissell, P. D., Avramescu, S., Wang, D. S., and Orser, B. A. (2016). deltaGABAA receptors are necessary for synaptic plasticity in the hippocampus: implications for memory behavior. *Anesth. Analg.* 123, 1247-1252. doi: 10.1213/ANE.0000000000001373

Xiang, J. F., Yin, Q. F., Chen, T., Zhang, Y., Zhang, X. O., Wu, Z., et al. (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res.* 24, 513-531. doi: 10.1038/cr.2014.35

Yan, C., Jiang, Y., Wan, Y., Zhang, L., Liu, J., Zhou, S., et al. (2017). Long noncoding RNA NBAT-1 suppresses tumorigenesis and predicts favorable prognosis in ovarian cancer. *Onco. Targets. Ther.* 10, 1993–2002. doi: 10.2147/ott.s124645

Yang, C., Wang, G., Yang, J., and Wang, L. (2017). Long noncoding RNA NBAT1 negatively modulates growth and metastasis of osteosarcoma cells through suppression of miR-21. *Am. J. Cancer Res.* 7, 2009-2019.

Younger, S. T., and Rinn, J. L. (2014). 'Lnc'-ing enhancers to MYC regulation. *Cell Res.* 24, 643-644. doi: 10.1038/cr.2014.54

Zhang, H., Meng, H., Huang, X., Tong, W., Liang, X., Li, J., et al. (2019). lncRNA MIR4435-2HG promotes cancer cell migration and invasion in prostate carcinoma by upregulating TGF-beta1. *Oncol. Lett.* 18, 4016-4021. doi: 10.3892/ol.2019.10757

Zhang, K., Sun, X., Zhou, X., Han, L., Chen, L., Shi, Z., et al. (2015). Long non-coding RNA HOTAIR promotes glioblastoma cell cycle progression in an EZH2 dependent manner. *Oncotarget* 6, 537–546. doi: 10.18632/oncotarget.2681

Zhang, S., Zhao, B. S., Zhou, A., Lin, K., Zheng, S., Lu, Z., et al. (2017). m(6) A demethylase ALKBH5 maintains tumorigenicity of glioblastoma stem-like cells by sustaining FOXM1 expression and cell proliferation program. *Cancer Cell* 31, 591-606.e596. doi: 10.1016/j.ccell.2017.02.013

Zhao, M., Wang, S., Li, Q., Ji, Q., Guo, P., and Liu, X. (2018). MALAT1: a long non-coding RNA highly associated with human cancers. *Oncol. Lett.* 16, 19-26. doi: 10.3892/ol.2018.8613

Zheng, T., Li, D., He, Z., Feng, S., and Zhao, S. (2018). Long noncoding RNA NBAT1 inhibits autophagy via suppression of ATG7 in non-small cell lung cancer. *Am. J. Cancer Res.* 8, 1801-1811.

Zhou, X. P., Zhan, W. J., Bian, W. B., Hua, L., Shi, Q., Xie, S., et al. (2013). GOLPH3 regulates the migration and invasion of glioma cells though RhoA. *Biochem. Biophys. Res. Commun.* 433, 338-344. doi: 10.1016/j.bbrc.2013.03.003

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 August 2021
doi: 10.3389/fgene.2021.721258

COL3A1 and MMP9 Serve as Potential Diagnostic Biomarkers of Osteoarthritis and Are Associated With Immune Cell Infiltration

*Shushan Li, Haitao Wang, Yi Zhang, Renqiu Qiao, Peige Xia, Zhiheng Kong, Hongbo Zhao and Li Yin**

Department of Orthopedic Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

OPEN ACCESS

Edited by:
Shengli Li,
Shanghai Jiao Tong University, China

Reviewed by:
Sijia Yan,
Shanghai Jiao Tong University, China
Hong Wang,
China Pharmaceutical University, China

****Correspondence:***
Li Yin
liyin1966zzu@hotmail.com

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

Received: *06 June 2021*
Accepted: *02 August 2021*
Published: *27 August 2021*

Citation:
Li S, Wang H, Zhang Y, Qiao R, Xia P, Kong Z, Zhao H and Yin L (2021) COL3A1 and MMP9 Serve as Potential Diagnostic Biomarkers of Osteoarthritis and Are Associated With Immune Cell Infiltration. Front. Genet. 12:721258. doi: 10.3389/fgene.2021.721258

Background: Osteoarthritis (OA) is one of the most common age-related degenerative diseases. In recent years, some studies have shown that pathological changes in the synovial membrane occur earlier than those in the cartilage in OA. However, the molecular mechanism of synovitis in the pathological process of OA has not been elucidated. This study aimed to identify novel biomarkers associated with OA and to emphasize the role of immune cells in the pathogenesis of OA.

Methods: Microarray datasets were obtained from the Gene Expression Omnibus (GEO) and ArrayExpress databases and were then analyzed using R software. To determine differential immune cell subtype infiltration, the CIBERSORT deconvolution algorithm was used. Quantitative reverse transcription PCR (qRT-PCR) was used to determine the relative expressions of selected genes. Besides, Western blotting was used to assess the protein expression levels in osteoarthritic chondrocytes.

Results: After analyzing the database profiles, two potential biomarkers, collagen type 3 alpha 1 chain (*COL3A1*), and matrix metalloproteinase 9 (*MMP9*), associated with OA were discovered, which were confirmed by qRT-PCR and Western blotting. Specifically, the results revealed that, as the concentration of IL-1β increased, so did the gene and protein expression levels of *COL3A1* and *MMP9*.

Conclusion: The findings provide valuable information and direction for future research into novel targets for OA immunotherapy and diagnosis and aids in the discovery of the underlying biological mechanisms of OA pathogenesis.

Keywords: osteoarthritis, immune cell infiltration, bioinformatics, GEO, diagnostic markers

INTRODUCTION

Osteoarthritis (OA), one of the most common age-related degenerative diseases, is characterized by osteophyte formation, cartilage degeneration, and synovial inflammation (Luo et al., 2018; Wang et al., 2018), which eventually lead to loss of joint function due to the limited repair capacity of the cartilage (Kim et al., 2018). However, the pathology of OA is not fully understood, and there is no treatment available to prevent or slow its progression (Wang et al., 2019). As a result, early diagnosis and treatment are preferred to improve joint function and alleviate joint pain.

According to recent research, the degenerative changes in the synovial membrane in OA occur earlier than those in the cartilage (Sakurai et al., 2019). OA synovitis is most likely caused by an innate immune response and is mediated by the expression of matrix-degrading enzymes, inflammatory cytokines, and chemokines (Gómez et al., 2015; Qadri et al., 2020). In several studies, the degree of synovitis has been validated as a strong predictor of OA, particularly in its early stages (Conaghan et al., 2010; Mathiessen and Conaghan, 2017). Immune responses are widely acknowledged to play an important role in the pathogenesis of OA (Daheshia and Yao, 2008; Han et al., 2018; Jenei-Lanzl et al., 2019). Pro-inflammatory cytokines promote chondrocyte apoptosis and cartilage matrix proteolysis (Utomo et al., 2016; Mobasheri et al., 2017). Furthermore, inflammatory suppression may aid in alleviating cartilage degradation in OA (Kapoor et al., 2011). However, the molecular mechanism of synovitis in the pathological process of OA has not been elucidated.

In the present study, microarray data from synovial membrane and cartilage samples in aged OA patients were integrated and the diagnostic biomarkers of OA were determined. The CIBERSORT algorithm method was then used to analyze immune cell infiltration in "normal" synovial membrane and OA synovial membrane. Furthermore, osteoarthritic chondrocytes (OA-CH) were stimulated with interleukin 1β (IL-1β) to establish a standardized *in vitro* OA model; the relationship between IL-1β and diagnostic biomarkers [collagen type 3 alpha 1 chain (*COL3A1*) and matrix metalloproteinase 9 (*MMP9*)] was determined by quantitative reverse transcription PCR (qRT-PCR) and Western blotting. This study aimed to identify novel biomarkers associated with OA and to emphasize the importance of immune cells in the pathogenesis of OA. The findings of the current study could lead to new OA diagnostic targets.

MATERIALS AND METHODS

Identification of Differentially Expressed Genes

Figure 1 depicts the study workflow. Microarray datasets of synovial membrane (GSE55235 and GSE55457) and cartilage (GSE117999, GSE1919, GSE51588, and E-MTAB-5564) samples were obtained from the Gene Expression Omnibus (GEO)[1] and ArrayExpress[2] databases. The ComBat function in the sva R package[3] was used to correct inter-batch differences in the different datasets. The limma package[4] in R was used to normalize and screen differentially expressed genes (DEGs) by comparing the expression levels in the synovial membrane from normal joints to those from OA joints. DEGs with $|\text{logFC}| > 2$ and an adjusted p-value < 0.05 were considered significantly expressed.

GO and KEGG Pathway Enrichment

The cluster Profiler[5] in R package was used to perform Gene Ontology (GO) annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the enriched DEGs. A value of $p < 0.05$ was considered statistically significant.

Construction of a PPI Network and Analysis of Hub Genes

A protein–protein interaction (PPI) network was established using STRING,[6] an online PPI establishment tool. The genes with a combined score of 0.4 were selected and used to establish the PPI network. Furthermore, the Cytohub plugin in Cytoscape version 3.8.0[7] was used to identify hub genes using the degree method (degree > 4).

CIBERSORT Analysis of Immune Cell Infiltration

The CIBERSORT deconvolution algorithm[8] was used to determine differential immune cell subtype infiltration between normal and OA synovial membrane samples. The difference in immune cell density between the normal and rheumatoid arthritis (RA) groups was visualized using a heatmap package in R version 3.6.0. The Wilcoxon signed-rank test was used to determine the statistical significance of the differences in immune cell infiltration between the two groups as depicted by violin plots.

Ethical Statement

The use of human material was approved by the local ethics committee of The First Affiliated Hospital of Zhengzhou University (reference no. 2021-KY-0338-002), and all patients provided written consent.

IL-1β Stimulation of OA Chondrocytes

Osteoarthritic chondrocytes (2×10^5, passages 2–4) were cultured in six-well plates with DMEM F12 medium [supplemented with 10% normal fetal calf serum (FCS) and 1% penicillin–streptomycin], stimulated with IL-1β (1, 5, and 10 ng/ml) (MAN0004230; Thermo Fisher Scientific, Waltham, MA, United States), and harvested for RNA and protein isolation after 24 and 48 h, respectively.

RNA Extraction and Real-Time PCR Analysis

Total RNA was isolated from the cells using the Absolutely RNA Miniprep Kit (Agilent Technologies, Santa Clara, CA, United States) according to the manufacturer's instructions and reverse-transcribed into complementary DNA (cDNA) using

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

[2] https://www.ebi.ac.uk/arrayexpress/

[3] https://bioconductor.org/packages/release/bioc/html/sva.html

[4] http://bioconductor.org/packages/release/bioc/html/limma.html

[5] https://www.bioconductor.org/help/search/index.html?q=clusterProfiler/

[6] http://string-db.org

[7] https://cytoscape.org/

[8] https://cibersort.stanford.edu/

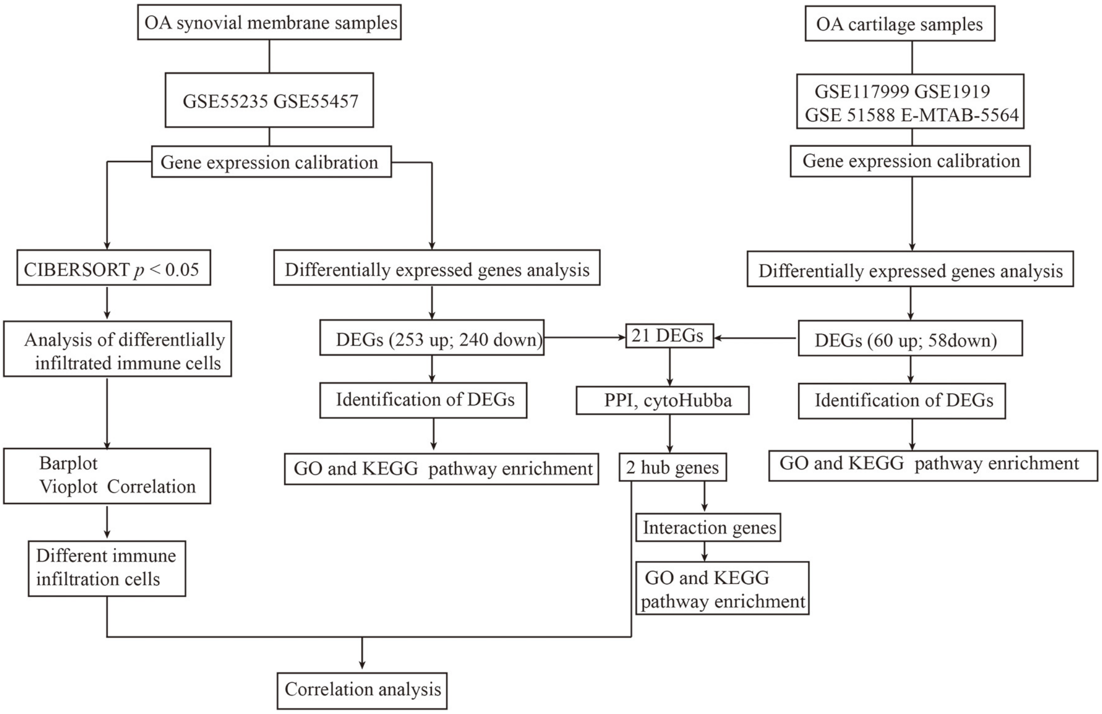

OA synovial membrane samples
GSE55235 GSE55457
Gene expression calibration
CIBERSORT $p < 0.05$
Differentially expressed genes analysis
Analysis of differentllially infiltrated immune cells
DEGs (253 up; 240 down)
Identification of DEGs
GO and KEGG pathway enrichment
Barplot
Vioplot Correlation
Different immune infiltration cells
OA cartilage samples
GSE117999 GSE1919
GSE 51588 E-MTAB-5564
Gene expression calibration
Differentially expressed genes analysis
DEGs (60 up; 58down)
Identification of DEGs
GO and KEGG pathway enrichment
21 DEGs
PPI, cytoHubba
2 hub genes
Interaction genes
GO and KEGG pathway enrichment
Correlation analysis

FIGURE 1 | Workflow of the entire study.

the AffinityScript QPCR cDNA Synthesis Kit (#600559; Agilent Technologies, Santa Clara, CA, United States). Subsequently, an MX3005P QPCR System (Agilent Technologies, Santa Clara, CA, United States) was used to perform real-time PCR for messenger RNA (mRNA) expression with Brilliant III Ultra-Fast SYBR® Green QPCR Master Mix (#600882; Agilent Technologies, Santa Clara, CA, United States). The primer sequences of the target genes were as follows: *MMP9* (Fwd: 5′-GTA CCA CGG CCA ACT ACG AC-3′; Rev: 5′-GCC TTG GAA GAT GAA TGG AA-3′), *COL3A1* (Fwd: 5′-CTTCTCTCCAGCCGAGCTTC-3′; Rev: 5′-TGTGTTTCGTGCAACCATCC-3′), *TBP* (Fwd: 5′-TTGTAC CGCAGCTGCAAA AT-3′; Rev: 5′-TATATTC GGCGTTTCGGGCA-3′), and *GAPDH* (Fwd: 5′-CT GACTTCAACAGCGACACC-3′; Rev: 5′-CC CTGTTGCTGTAGCCAAAT-3′). All genes were analyzed relatively, calibrated to the expression of the control cell culture groups, and normalized to *GAPDH* and *TBP*.

Protein Extraction and Western Blotting Analysis

Osteoarthritic chondrocytes were washed twice with cold phosphate-buffered saline (PBS) and lysed with RIPA buffer (Thermo Fisher Scientific, Waltham, MA, United States) containing proteinase inhibitors (Roche, Basel, Switzerland). The concentration of cellular protein was determined using a BCA protein kit assay. Cell lysates were mixed with sodium dodecyl sulfate (SDS) sample loading buffer (#B7053; Sigma-Aldrich, Taufkirchen, Germany), boiled for 5 min at 95°C, and then subjected to 10% SDS-PAGE. After electrophoretic separation, the proteins were transferred to 0.22-mm polyvinylidene fluoride (PVDF) membranes (Roche, Penzberg, Germany). Blot membranes were blocked with 5% bovine serum albumin (BSA) for 1 h at room temperature and incubated with primary antibodies on a shaker overnight at 4°C. The membranes were then washed and incubated with the appropriate horseradish peroxidase-coupled secondary antibodies (Santa Cruz Biotechnology and Jackson ImmunoResearch, West Grove, PA, United States). The proteins were examined using enhanced chemiluminescence (ECL) detection reagents (Thermo Scientific, Waltham, MA, United States) and signals were normalized to β-actin. The following primary antibodies were used in this study: COL3A1 (1:1,000, #ab838292; Abcam, Cambridge, MA, United States), MMP9 (1:200, #sc-393859; Santa Cruz, Heidelberg, Germany), and β-actin (1:5,000, #ab8227; Abcam, Cambridge, MA, United States).

Statistical Analysis

R version 3.6.0 was used to perform bioinformatics analyses, and a p-value < 0.05 was considered statistically significant. Correlations were determined using Pearson's correlation coefficient, with $|R| < 0.5$ indicating a weak correlation. For qRT-PCR and Western blotting analyses, an unpaired Student's t-test was used for two groups and one-way ANOVA was used for groups of more than two. Each assay was replicated and repeated in at least three independent experiments. A value of $p < 0.05$ was considered statistically significant.

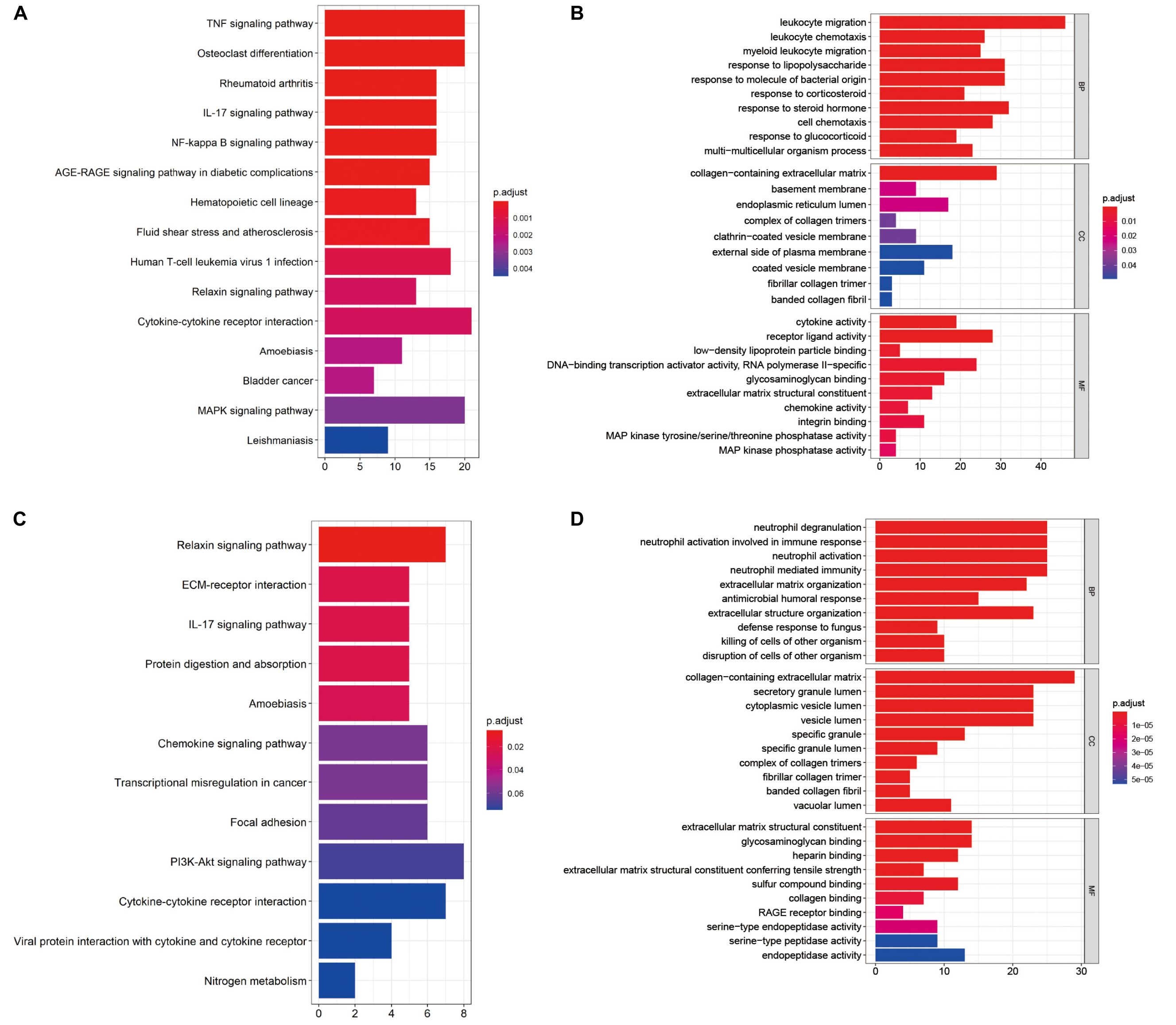


FIGURE 2 | Functional enrichment of the differentially expressed genes (DEGs) in the synovial membrane and cartilage samples. **(A,B)** Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment and Gene Ontology (GO) analysis of the DEGs in synovial membrane samples. **(C,D)** KEGG pathway enrichment and GO analysis of the DEGs in cartilage samples.

RESULTS

Identification of DEGs

Microarray datasets of synovial membrane (GSE55235 and GSE55457) and cartilage samples (GSE117999, GSE1919, GSE51588, and E-MTAB-5564) were obtained from the GEO and ArrayExpress databases. Before analyzing the DEGs, raw data were preprocessed for batch correction and normalization. Gene expression levels with | logFC| > 1 and an adjusted *p*-value < 0.05 were considered differentially expressed. As a result, 253 upregulated and 240 downregulated DEGs were identified in synovial membrane samples when compared to normal samples, while 60 upregulated and 58 downregulated DEGs were identified in cartilage samples, as shown in **Figure 1**.

Function Annotation of DEGs

Kyoto Encyclopedia of Genes and Genomes pathway enrichment and GO functional enrichment of DEGs were performed to investigate the mechanisms involved in the pathogenesis of OA. KEGG pathway enrichment revealed that synovial membrane DEGs were mainly enriched in cytokine–cytokine receptor interaction, mitogen-activated protein kinase (MAPK) pathway, and tumor necrosis factor (TNF) pathway (**Figure 2A**), while DEGs from cartilage samples were enriched in the PI3K/AKT pathway, cytokine–cytokine receptor interaction,

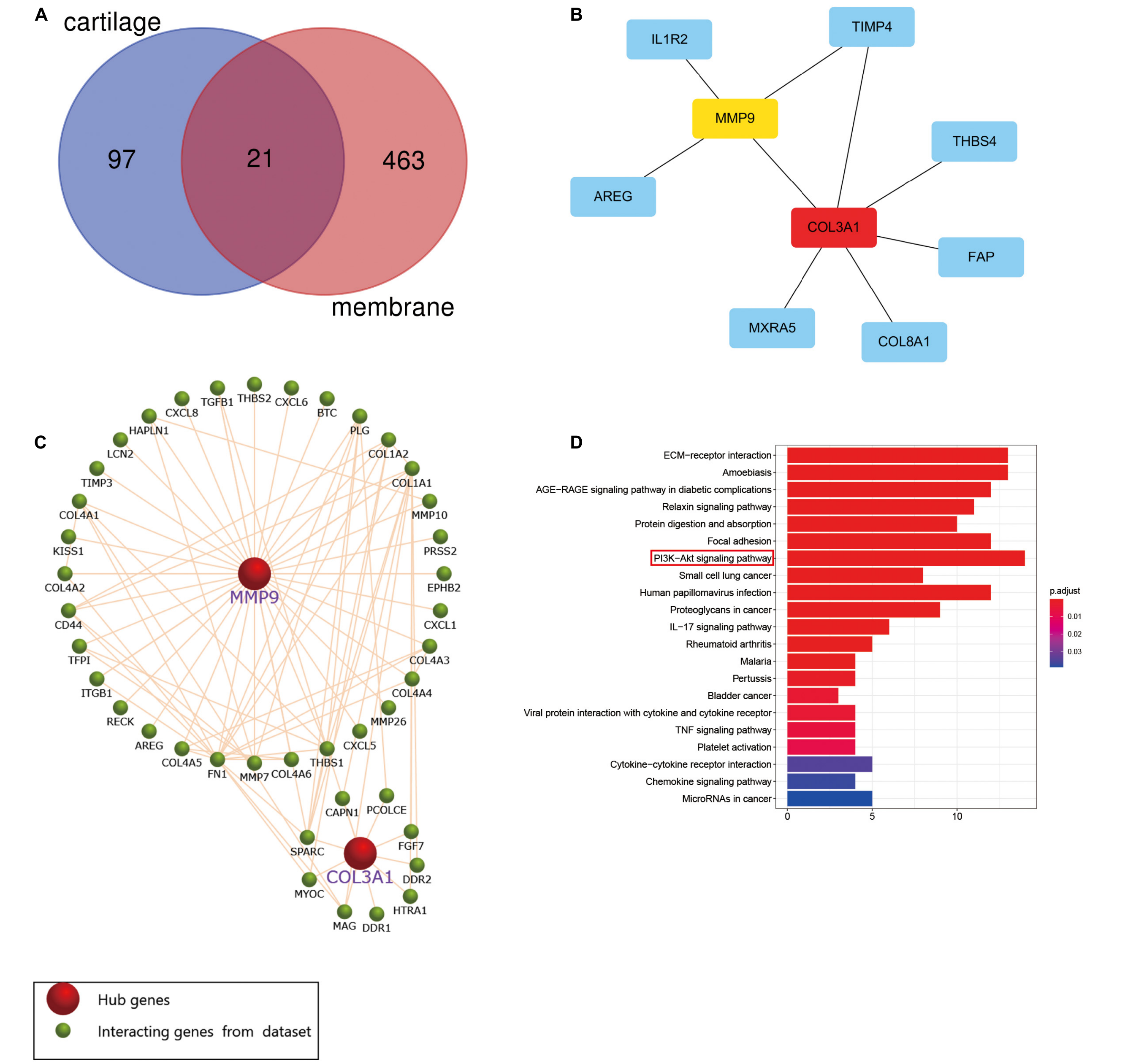


FIGURE 3 | Screening of hub genes and functional analysis. **(A)** Twenty-one differentially expressed genes (DEGs) intersected between the cartilage samples and the synovial membrane samples. **(B)** Protein–protein interaction (PPI) network of the 21 DEGs and two hub genes screened by the degree method (degree > 4) using cytoHubba. A higher ranking is represented by a redder color. **(C)** *MMP9* and *COL3A1* interacting genes indicated using Funrich software. **(D)** Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of *MMP9* and *COL3A1* interacting genes.

and the chemokine pathway (**Figure 2C**). Furthermore, GO functional enrichment analysis revealed that synovial membrane DEGs were mainly involved in leukocyte migration, regulation of inflammatory response, and collagen-containing extracellular matrix (**Figure 2B**), while cartilage DEGs were mainly involved in the collagen-containing extracellular matrix, neutrophil degranulation, and neutrophil activation involved in immune response (**Figure 2D**). These findings suggest that DEGs in both the synovial membrane and cartilage are involved in immune response signaling pathways and that the immune system plays a critical role in the pathological processes of OA.

Screening and Validation of *MMP9* and *COL3A1* Hub Genes

The Venn diagram showed that 21 DEGs from the synovial membrane and cartilage samples overlapped (**Figure 3A**). The

TABLE 1 | Signaling pathway enrichment of *MMP9* and *COL3A1* interacting genes.

ID	Description	*p*-value	p_{adjust}	Count
hsa04151	PI3K–Akt signaling pathway	4.70E–11	6.11E–10	14
hsa04512	ECM–receptor interaction	9.39E–18	8.55E–16	13
hsa05146	Amoebiasis	7.07E–17	3.22E–15	13
hsa04933	AGE–RAGE signaling pathway in diabetic complications	2.89E–15	8.77E–14	12
hsa04510	Focal adhesion	1.38E–11	2.09E–10	12
hsa05165	Human papillomavirus infection	4.48E–09	4.53E–08	12
hsa04926	Relaxin signaling pathway	2.35E–12	5.35E–11	11
hsa04974	Protein digestion and absorption	7.43E–12	1.35E–10	10
hsa05205	Proteoglycans in cancer	1.11E–07	1.01E–06	9
hsa05222	Small cell lung cancer	2.97E–09	3.38E–08	8
hsa04657	IL–17 signaling pathway	2.19E–06	1.81E–05	6
hsa05323	Rheumatoid arthritis	3.88E–05	0.000295	5
hsa04060	Cytokine–cytokine receptor interaction	0.007225	0.034604	5
hsa05206	MicroRNAs in cancer	0.008991	0.03896	5
hsa05144	Malaria	5.26E–05	0.000368	4
hsa05133	Pertussis	0.000271	0.001763	4
hsa04061	Viral protein interaction with cytokine and cytokine receptor	0.000772	0.004389	4
hsa04668	TNF signaling pathway	0.00118	0.006314	4
hsa04611	Platelet activation	0.001719	0.008692	4
hsa04062	Chemokine signaling pathway	0.008221	0.037406	4
hsa05219	Bladder cancer	0.000652	0.003953	3

PPI network between the overlapping DEGs was constructed and two hub genes, *MMP9* and *COL3A1*, were filtered out (**Figure 3B**) by the degree method (degree > 4) using cytoHubba. The Funrich software was used to display the 41 interacting genes to better understand the functions of *MMP9* and *COL3A1* (**Figure 3C**). In addition, KEGG enrichment revealed that 41 *MMP9* and *COL3A1* interacting genes were involved in the PI3K/AKT pathway, IL-17 pathway, TNF pathway, and other immune-related pathways (**Figure 3D** and **Table 1**). These findings imply that MMP9 and COL3A1 are involved in the pathophysiological inflammatory processes that lead to OA.

Analysis of Immune Cell Infiltration in Normal and OA Synovial Membrane Samples

The CIBERSORT algorithm was, for the first time, used to reveal the landscape of the differentially infiltrated immune cells in "*normal*" *versus* OA synovial membrane samples in 22 subpopulations of immune cells. The heatmap shows the proportion of immune cells in the two groups (**Figure 4A**).

The correlation heatmap of the 22 immune cell subtypes showed that two pairs of immune cells [active natural killer (NK) cells and eosinophils, and naive CD4 T cells and resting NK cells] were positively correlated and that two immune cell subtypes (activated mast cells and resting mast cells) were negatively correlated (**Figure 4B**).

Furthermore, the violin plot of the differentially infiltrated immune cells showed that regulatory T cells (Tregs) and resting mast cells had the highest infiltration rates in OA samples compared with "normal" samples, whereas resting CD4^{+} memory T cells, activated NK cells, activated mast cells, and eosinophils were less prominent in OA samples (**Figures 5A**, **6B**).

Correlation Between Hub Genes (*MMP9* and *COL3A1*) and Immune Cell Infiltration

Spearman's correlation analysis was performed to determine the association between the hub genes (*MMP9* and *COL3A1*) and the infiltrated immune cell subtypes in the synovial membranes of both groups (**Figure 5B**). *MMP9* and *COL3A1* were found to be negatively correlated with resting CD4 memory T cells, whereas *MMP9* was found to be positively correlated with M0 macrophages and negatively correlated with activated NK cells.

Validation of Hub Genes (*COL3A1* and *MMP9*) by qRT-PCR and Western Blotting

The fragments per kilobase of exon model per million mapped fragments (FPKM) values of *COL3A1* and *MMP9* were significantly higher in the OA cartilage and synovial membrane compared with those in normal samples (**Figures 6A,B**). To validate the expressions of *COL3A1* and *MMP9* in chondrocytes, qRT-PCR and Western blotting were used to determine the gene and protein expressions in non-osteoarthritic chondrocytes (NCH), OA–CH, and OA–CH treated with different concentrations of IL-1β. As shown in **Figures 6C,D**, the gene expression levels of both *COL3A1* and *MMP9* increased in the OA-CH group, and IL-1β promoted the expressions of *COL3A1* and *MMP9*. Notably, the expressions of *COL3A1* and *MMP9* increased as the

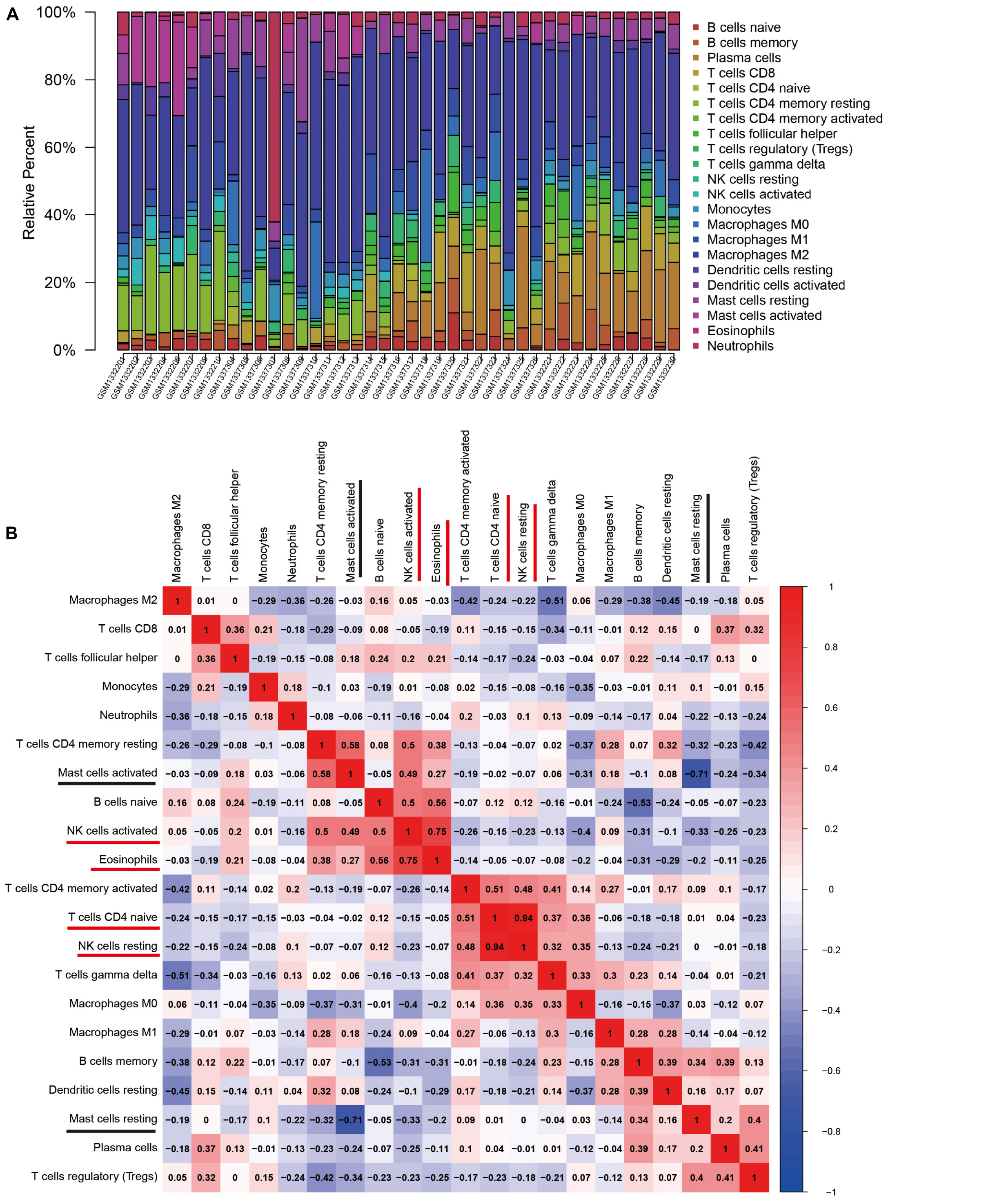


FIGURE 4 | The landscape and correlation heatmap of immune infiltration in synovial membrane samples between the normal and osteoarthritis (OA) groups. **(A)** Relative distribution of 22 immune cells in all samples. **(B)** Correlation heatmap of immune cells in all samples. Red squares indicate positive correlation and blue squares indicate negative correlation; the deeper colored squares indicate stronger correlations.

concentration of IL-1β increased. Furthermore, in the presence of IL-1β, the protein levels of *COL3A1* and *MMP9* increased (**Figures 6E–G**). These findings suggest that the gene and protein expression levels of *COL3A1* and *MMP9* were positively correlated with the degree of inflammation and the inflammatory activity.

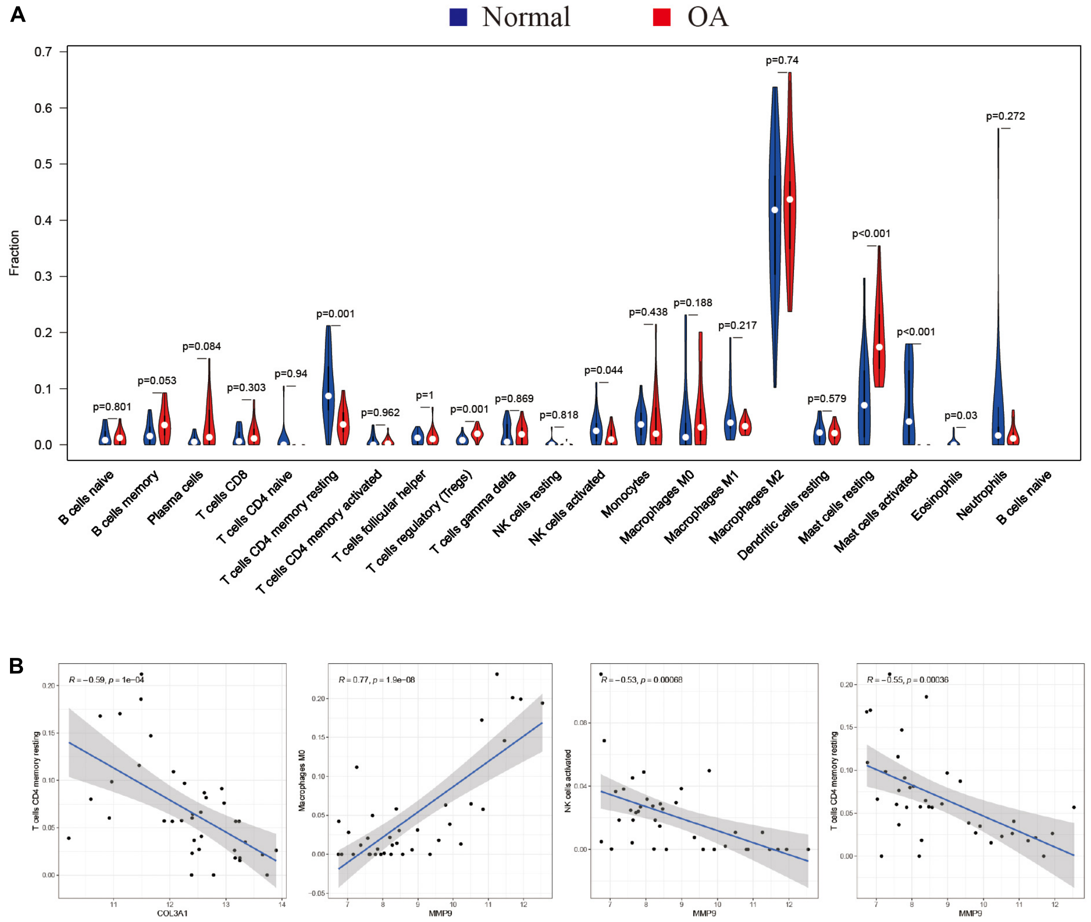


FIGURE 5 | Characterization of immune cell infiltration in normal and osteoarthritis (OA) samples and the correlation between hub gene expression and immune cell infiltration. **(A)** Violin plot showing the differentially infiltrated immune cells of a proportion of the 22 immune cell types. The red underline shows significant difference in the immune cell infiltration between the normal and rheumatoid arthritis (RA) groups. A value of $p < 0.05$ was considered to be statistically significant. **(B)** Correlation coefficient (R) > 0.5; $p < 0.05$ was considered statistically significant.

DISCUSSION

Osteoarthritic is a type of chronic joint disease that is characterized by cartilage degeneration, hyperosteogenia, and synovitis (Xie and Chen, 2019). Accumulating evidence suggests that pro-inflammatory cytokines, such as IL-1β, TNF, and IL-6, play a role in the pathophysiology of OA (Robinson et al., 2016; Urban and Little, 2018). Previous research has focused on the molecular mechanism of OA in the cartilage or chondrocytes while ignoring the synovial membrane. In recent years, an increasing number of studies have shown that synovitis plays a critical role in the pathological process of OA, from the early to the end stages (Atukorala et al., 2016; Huang et al., 2018; Griffin and Scanzello, 2019). Additional research has revealed changes in immune cell infiltration in OA synovial membrane samples (Moradi et al., 2014; Penatti et al., 2017; Rosshirt et al., 2019). However, no study has been conducted to investigate the inflammatory relationship between the synovial membrane and cartilage. In the present study, the gene expression profiles of the synovial membrane and cartilage were combined to identify the important hub genes associated with synovitis in OA.

Differentially expressed genes in the synovial membrane and cartilage were separately analyzed; GO annotation and KEGG pathway enrichment were used to reveal the functions of these DEGs. Our results also showed that both synovial membrane and cartilage DEGs were mainly involved in inflammatory pathways and pathological processes, which was consistent with previous studies (Qin et al., 2012; Hou et al., 2013; Chen et al., 2018). Furthermore, the immune response occurred in the synovial membrane and cartilage, indicating that synovitis plays a critical role in the pathological process of OA.

The hub genes *COL3A1* and *MMP9* were identified and their function validated using Funrich software and by KEGG pathway enrichment, respectively. Besides, *COL3A1* and *MMP9* interacting genes were found to be mainly involved in the PI3K/AKT signaling pathway, extracellular matrix (ECM) receptor interaction, and other inflammatory signaling pathways (IL-17 signaling pathway, cytokine–cytokine receptor interaction, TNF signaling pathway, and chemokine signaling

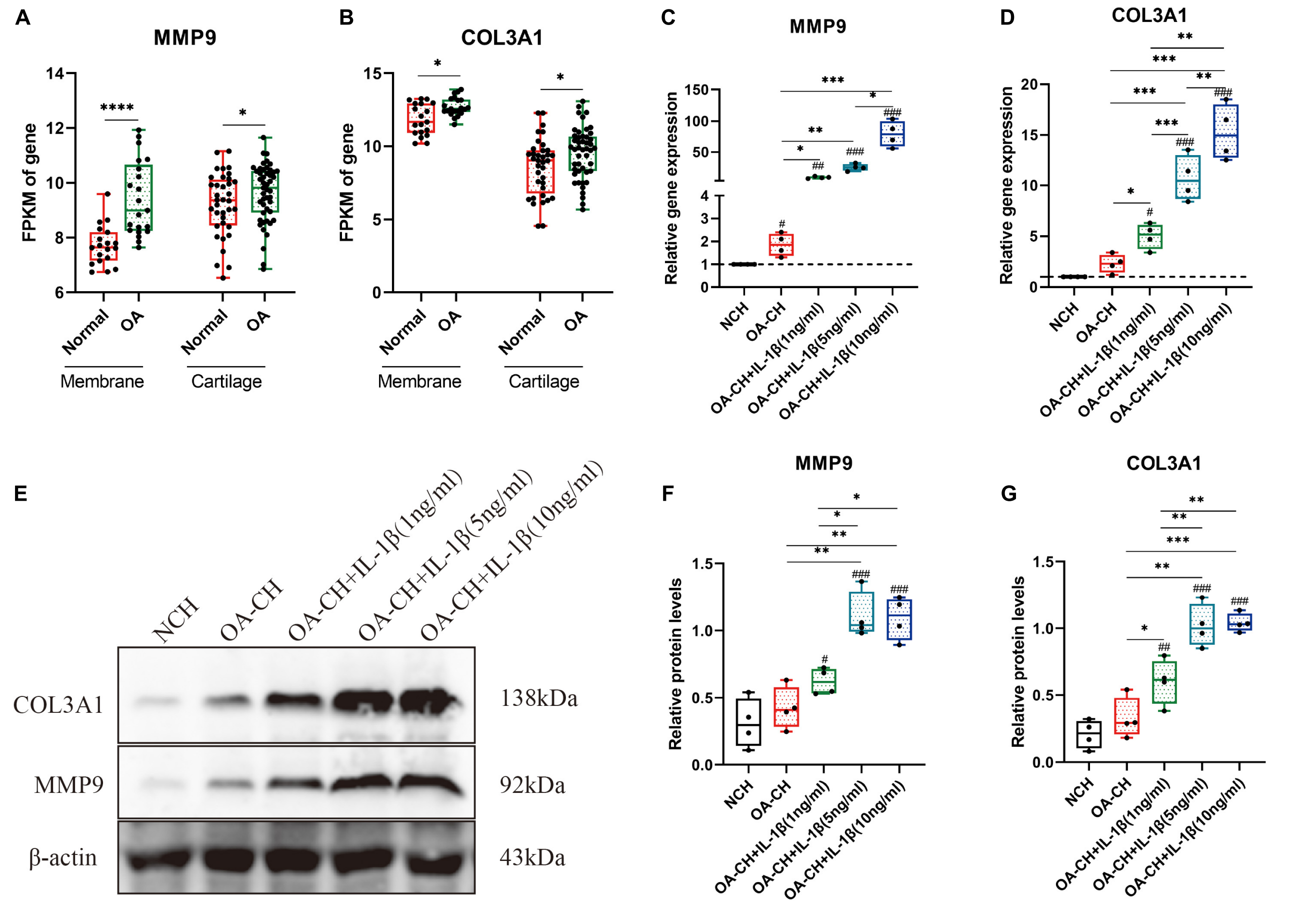


FIGURE 6 | Validation of the hub genes (*COL3A1* and *MMP9*) by quantitative reverse transcription PCR (qRT-PCR) and Western blotting. **(A,B)** Fragments per kilobase of exon model per million mapped fragments (FPKM) of *COL3A1* and *MMP9* in synovial membrane and cartilage samples. *$p < 0.05$; ****$p < 0.001$, unpaired Student's *t*-test. **(C,D)** Gene expression levels of *COL3A1* and *MMP9* in non-osteoarthritic chondrocytes (NCH) and osteoarthritic chondrocytes (OA-CH) treated with different concentrations of IL-1β. **(E–G)** Western blotting was used to determine the protein expression levels of *COL3A1* and *MMP9* in NCH and OA-CH treated with different concentrations of IL-1β. Significant difference to control (NCH): $^{\#}p < 0.05$; $^{\#\#}p < 0.01$; $^{\#\#\#}p < 0.001$. *Significant difference between groups: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. One-way ANOVA with Newman–Keuls multiple comparison test. All values represent the mean ± standard deviation ($n = 4$).

pathway). Numerous studies have reported that inflammatory signaling pathways, including the PI3K/AKT, IL-17, TNF, NF-κB, and MAPK signaling pathways, are involved in the osteoarthritic process (Balabko et al., 2015; Zhang et al., 2018; Han et al., 2019; Li and Zheng, 2019). These findings suggest that *COL3A1* and *MMP9* play important roles in the inflammatory signaling pathways linked to OA.

To further investigate the effect of immune cell infiltration in OA, CIBERSORT was used to perform a comprehensive analysis of OA immune infiltration. The results showed increased infiltration of Tregs and resting mast cells, which contributed to the occurrence and development of OA. Moradi et al. (2014) found that Tregs are enriched in the synovial membrane of OA patients and correlated with the levels of inflammatory factors (IL-10 and TGF-β) (Xia et al., 2017). Resting mast cells were found in high numbers in OA synovial tissue, which is associated with structural damage in OA patients (de Lange-Brokaar et al., 2016). These findings and other related research indicate that Tregs and resting mast cells play an important role in OA. In this study, the relationship between the immune cell subtypes in OA was investigated; the results showed that two pairs of immune cells (activated NK cells and eosinophils, and naive CD4 T cells and resting NK cells) were positively correlated and that two immune cell subtypes (activated and resting mast cells) were negatively correlated. However, the correlation between the immune cell subtypes requires further experimental validation.

The relationship between hub gene expression and immune cell infiltration was also analyzed. The results showed that the expressions of both *MMP9* and *COL3A1* were negatively correlated with resting CD4 memory T cells, while the expression of *MMP9* was positively correlated with M0 macrophages and negatively correlated with activated NK cells. We hypothesized that *MMP9* and *COL3A1* inhibited the immune response by reducing the resting CD4 memory T cells and activated NK cells and that *MMP9* increased M0 macrophages to induce inflammation in the course of OA. However, further research is needed to validate these assumptions on the relationship between hub genes and immune cells.

To investigate the correlation between the hub genes (*MMP9* and *COL3A1*) and OA, qRT-PCR, and Western blotting were used to determine the gene and protein expression levels in chondrocytes. The results indicated that the gene expression levels of *MMP9* and *COL3A1* increased in OA–CH compared with those in NCH. Notably, the gene and protein expression levels of *COL3A1* and *MMP9* increased with an increase in IL-1β concentration. Evidence suggests that the expression levels of *COL3A1* increased in the early stages of OA and decreased in the later stages (Rai et al., 2019). Tang et al. (2018) also discovered that IL-1 increased the protein levels of *COL3A1* in synoviocytes. MMP9, also known as gelatinase B, is an enzyme that degrades the ECM components such as collagen, fibronectin, and laminin. MMP9 was found to be upregulated at the mRNA and protein levels in the cartilage and synovial membrane, as well as in the synovial fluid, and was found to be related to the severity of OA (Bollmann et al., 2021). These findings suggest that COL3A1 and MMP9 could be used as OA diagnostic biomarkers. However, more research is needed to determine the roles of COL3A1 and MMP9 in the progression of OA.

CONCLUSION

In conclusion, the present study identified two potential OA biomarkers, *COL3A1*, and *MMP9*, which were confirmed by qRT-PCR and Western blotting analysis. Notably, the gene and protein expression levels of *COL3A1* and *MMP9* increased with an increase in IL-1β concentration. These findings provide valuable information and direction for future research into novel targets for OA immunotherapy and diagnosis and aid in the discovery of the underlying biological mechanisms of OA pathogenesis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The use of human material has been approved by the local ethics committee (reference number: 2021-KY-0338-002, The First Affiliated Hospital of Zhengzhou University) and the written consent of all patients has been obtained.

AUTHOR CONTRIBUTIONS

SL contributed to sample collection, experiment design, investigation, data curation, writing the original draft, and review and editing. HW and YZ helped with conceptualization and review and editing. RQ, PX, HZ, and ZK helped with the methodology, establishment of qRT-PCR, and bioinformatics analysis. LY made contributions to the conceptualization, review and editing, project administration, and funding acquisition. All authors proofread the final version of the manuscript.

FUNDING

This study was funded by the Key Scientific Research Projects of Colleges and Universities in Henan Province (20A310022).

REFERENCES

Atukorala, I., Kwoh, C. K., Guermazi, A., Roemer, F. W., Boudreau, R. M., Hannon, M. J., et al. (2016). Synovitis in knee osteoarthritis: a precursor of disease? *Ann. Rheum. Dis.* 75, 390–395. doi: 10.1136/annrheumdis-2014-205894

Balabko, L., Andreev, K., Burmann, N., Schubert, M., Mathews, M., Trufa, D. I., et al. (2015). Increased expression of the Th17-IL-6R/pSTAT3/BATF/RorγT-axis in the tumoural region of adenocarcinoma as compared to squamous cell carcinoma of the lung. *Sci. Rep.* 4:7396. doi: 10.1038/srep07396

Bollmann, M., Pinno, K., Ehnold, L. I., Märtens, N., Märtson, A., Pap, T., et al. (2021). MMP-9 mediated Syndecan-4 shedding correlates with osteoarthritis severity. *Osteoarthritis Cartilage* 29, 280–289. doi: 10.1016/j.joca.2020.10.009

Chen, Y., Shou, K., Gong, C., Yang, H., Yang, Y., and Bao, T. (2018). Anti-inflammatory effect of geniposide on osteoarthritis by suppressing the activation of p38 MAPK signaling pathway. *Biomed. Res. Int.* 2018:8384576. doi: 10.1155/2018/8384576

Conaghan, P. G., D'Agostino, M. A., Le Bars, M., Baron, G., Schmidely, N., Wakefield, R., et al. (2010). Clinical and ultrasonographic predictors of joint replacement for knee osteoarthritis: results from a large, 3-year, prospective EULAR study. *Ann. Rheum. Dis.* 69, 644–647. doi: 10.1136/ard.2008.099564

Daheshia, M., and Yao, J. Q. (2008). The interleukin 1beta pathway in the pathogenesis of osteoarthritis. *J. Rheumatol.* 35, 2306–2312. doi: 10.3899/jrheum.080346

de Lange-Brokaar, B. J. E., Kloppenburg, M., Andersen, S. N., Dorjée, A. L., Yusuf, E., Herb-van Toorn, L., et al. (2016). Characterization of synovial mast cells in knee osteoarthritis: association with clinical parameters. *Osteoarthritis Cartilage* 24, 664–671. doi: 10.1016/j.joca.2015.11.011

Gómez, R., Villalvilla, A., Largo, R., Gualillo, O., and Herrero-Beaumont, G. (2015). TLR4 signalling in osteoarthritis–finding targets for candidate DMOADs. *Nat. Rev. Rheumatol.* 11, 159–170. doi: 10.1038/nrrheum.2014.209

Griffin, T. M., and Scanzello, C. R. (2019). Innate inflammation and synovial macrophages in osteoarthritis pathophysiology. *Clin. Exp. Rheumatol.* 37(Suppl. 120), 57–63.

Han, P.-F., Wei, L., Duan, Z.-Q., Zhang, Z.-L., Chen, T.-Y., Lu, J.-G., et al. (2018). Contribution of IL-1β, 6 and TNF-α to the form of post-traumatic osteoarthritis induced by "idealized" anterior cruciate ligament reconstruction in a porcine model. *Int. Immunopharmacol.* 65, 212–220. doi: 10.1016/j.intimp.2018.10.007

Han, Y., Li, X., Yan, M., Yang, M., Wang, S., Pan, J., et al. (2019). Oxidative damage induces apoptosis and promotes calcification in disc cartilage endplate cell through ROS/MAPK/NF-κB pathway: implications for disc degeneration. *Biochem. Biophys. Res. Commun.* 516, 1026–1032. doi: 10.1016/j.bbrc.2017.03.111

Hou, C.-H., Tang, C.-H., Hsu, C.-J., Hou, S.-M., and Liu, J.-F. (2013). CCN4 induces IL-6 production through αvβ5 receptor, PI3K, Akt, and NF-κB singling pathway in human synovial fibroblasts. *Arthritis Res. Ther.* 15:R19. doi: 10.1186/ar4151

Huang, H., Zheng, J., Shen, N., Wang, G., Zhou, G., Fang, Y., et al. (2018). Identification of pathways and genes associated with synovitis in osteoarthritis using bioinformatics analyses. *Sci. Rep.* 8:10050. doi: 10.1038/s41598-018-28280-6

Jenei-Lanzl, Z., Meurer, A., and Zaucke, F. (2019). Interleukin-1β signaling in osteoarthritis–chondrocytes in focus. *Cell. Signal.* 53, 212–223. doi: 10.1016/j.cellsig.2018.10.005

Kapoor, M., Martel-Pelletier, J., Lajeunesse, D., Pelletier, J.-P., and Fahmi, H. (2011). Role of proinflammatory cytokines in the pathophysiology of osteoarthritis. *Nat. Rev. Rheumatol.* 7, 33–42. doi: 10.1038/nrrheum.2010.196

Kim, J.-R., Yoo, J. J., and Kim, H. A. (2018). Therapeutics in osteoarthritis based on an understanding of its molecular pathogenesis. *Int. J. Mol. Sci.* 19:674. doi: 10.3390/ijms19030674

Li, J., and Zheng, J. (2019). Theaflavins prevent cartilage degeneration via AKT/FOXO3 signaling in vitro. *Mol. Med. Rep.* 19, 821–830. doi: 10.3892/mmr.2018.9745

Luo, Y., He, Y., Reker, D., Gudmann, N. S., Henriksen, K., Simonsen, O., et al. (2018). A novel high sensitivity type II collagen blood-based biomarker, PRO-C2, for assessment of cartilage formation. *Int. J. Mol. Sci.* 19:3485. doi: 10.3390/ijms19113485

Mathiessen, A., and Conaghan, P. G. (2017). Synovitis in osteoarthritis: current understanding with therapeutic implications. *Arthritis Res. Ther.* 19:18. doi: 10.1186/s13075-017-1229-9

Mobasheri, A., Rayman, M. P., Gualillo, O., Sellam, J., van der Kraan, P., and Fearon, U. (2017). The role of metabolism in the pathogenesis of osteoarthritis. *Nat. Rev. Rheumatol.* 13, 302–311. doi: 10.1038/nrrheum.2017.50

Moradi, B., Schnatzer, P., Hagmann, S., Rosshirt, N., Gotterbarm, T., Kretzer, J. P., et al. (2014). CD4+CD25+/highCD127low/- regulatory T cells are enriched in rheumatoid arthritis and osteoarthritis joints–analysis of frequency and phenotype in synovial membrane, synovial fluid and peripheral blood. *Arthritis Res. Ther.* 16:R97. doi: 10.1186/ar4545

Penatti, A., Facciotti, F., De Matteis, R., Larghi, P., Paroni, M., Murgo, A., et al. (2017). Differences in serum and synovial CD4+ T cells and cytokine profiles to stratify patients with inflammatory osteoarthritis and rheumatoid arthritis. *Arthritis Res. Ther.* 19:103. doi: 10.1186/s13075-017-1305-1

Qadri, M., Jay, G. D., Zhang, L. X., Richendrfer, H., Schmidt, T. A., and Elsaid, K. A. (2020). Proteoglycan-4 regulates fibroblast to myofibroblast transition and expression of fibrotic genes in the synovium. *Arthritis Res. Ther.* 22:113. doi: 10.1186/s13075-020-02207-x

Qin, J., Shang, L., Ping, A., Li, J., Li, X., Yu, H., et al. (2012). TNF/TNFR signal transduction pathway-mediated anti-apoptosis and anti-inflammatory effects of sodium ferulate on IL-1β-induced rat osteoarthritis chondrocytes in vitro. *Arthritis Res. Ther.* 14:R242. doi: 10.1186/ar4085

Rai, M. F., Tycksen, E. D., Cai, L., Yu, J., Wright, R. W., and Brophy, R. H. (2019). Distinct degenerative phenotype of articular cartilage from knees with meniscus tear compared to knees with osteoarthritis. *Osteoarthritis Cartilage* 27, 945–955. doi: 10.1016/j.joca.2019.02.792

Robinson, W. H., Lepus, C. M., Wang, Q., Raghu, H., Mao, R., Lindstrom, T. M., et al. (2016). Low-grade inflammation as a key mediator of the pathogenesis of osteoarthritis. *Nat. Rev. Rheumatol.* 12, 580–592. doi: 10.1038/nrrheum.2016.136

Rosshirt, N., Hagmann, S., Tripel, E., Gotterbarm, T., Kirsch, J., Zeifang, F., et al. (2019). A predominant Th1 polarization is present in synovial fluid of end-stage osteoarthritic knee joints: analysis of peripheral blood, synovial fluid and synovial membrane. *Clin. Exp. Immunol.* 195, 395–406. doi: 10.1111/cei.13230

Sakurai, Y., Fujita, M., Kawasaki, S., Sanaki, T., Yoshioka, T., Higashino, K., et al. (2019). Contribution of synovial macrophages to rat advanced osteoarthritis pain resistant to cyclooxygenase inhibitors. *Pain* 160, 895–907. doi: 10.1097/j.pain.0000000000001466

Tang, S., Deng, S., Guo, J., Chen, X., Zhang, W., Cui, Y., et al. (2018). Deep coverage tissue and cellular proteomics revealed IL-1β can independently induce the secretion of TNF-associated proteins from human synoviocytes. *J. Immunol.* 200, 821–833. doi: 10.4049/jimmunol.1700480

Urban, H., and Little, C. B. (2018). The role of fat and inflammation in the pathogenesis and management of osteoarthritis. *Rheumatology (Oxford)* 57, iv10–iv21. doi: 10.1093/rheumatology/kex399

Utomo, L., Bastiaansen-Jenniskens, Y. M., Verhaar, J. A. N., and van Osch, G. J. V. M. (2016). Cartilage inflammation and degeneration is enhanced by pro-inflammatory (M1) macrophages in vitro, but not inhibited directly by anti-inflammatory (M2) macrophages. *Osteoarthritis Cartilage* 24, 2162–2170. doi: 10.1016/j.joca.2016.07.018

Wang, Q., Onuma, K., Liu, C., Wong, H., Bloom, M. S., Elliott, E. E., et al. (2019). Dysregulated integrin αVβ3 and CD47 signaling promotes joint inflammation, cartilage breakdown, and progression of osteoarthritis. *JCI Insight* 4:e128616. doi: 10.1172/jci.insight.128616

Wang, R., Xu, B., and Xu, H. (2018). TGF-β1 promoted chondrocyte proliferation by regulating Sp1 through MSC-exosomes derived miR-135b. *Cell Cycle* 17, 2756–2765. doi: 10.1080/15384101.2018.1556063

Xia, J., Ni, Z., Wang, J., Zhu, S., and Ye, H. (2017). Overexpression of lymphocyte activation gene-3 inhibits regulatory T Cell responses in osteoarthritis. *DNA Cell Biol.* 36, 862–869. doi: 10.1089/dna.2017.3771

Xie, C., and Chen, Q. (2019). Adipokines: new therapeutic target for osteoarthritis? *Curr. Rheumatol. Rep.* 21:71. doi: 10.1007/s11926-019-0868-z

Zhang, W., Hsu, P., Zhong, B., Guo, S., Zhang, C., Wang, Y., et al. (2018). MiR-34a enhances chondrocyte apoptosis, senescence and facilitates development of osteoarthritis by targeting DLL1 and regulating PI3K/AKT pathway. *Cell. Physiol. Biochem.* 48, 1304–1316. doi: 10.1159/000492090

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 September 2021
doi: 10.3389/fgene.2021.733211

A Pan-Cancer Analysis of Cystatin E/M Reveals Its Dual Functional Effects and Positive Regulation of Epithelial Cell in Human Tumors

Dahua Xu[1]†, Shun Ding[2]†, Meng Cao[3]†, Xiaorong Yu[3], Hong Wang[1], Dongqin Qiu[2], Zhengyang Xu[2], Xiaoman Bi[1], Zhonglin Mu[2]* and Kongning Li[1,3]**

[1] Key Laboratory of Tropical Translational Medicine of Ministry of Education, College of Biomedical Information and Engineering and Cancer Institute of the First Affiliated Hospital, Hainan Medical University, Haikou, China, [2] Department of Otolaryngology, Head and Neck Surgery, The First Affiliated Hospital, Hainan Medical University, Haikou, China, [3] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Cystatin E/M (CST6), a representative cysteine protease inhibitor, plays both tumor-promoting and tumor-suppressing functions and is pursued as an epigenetically therapeutic target in special cancer types. However, a comprehensive and systematic analysis for CST6 in pan-cancer level is still lacking. In the present study, we explored the expression pattern of CST6 in multiple cancer types across ∼10,000 samples from TCGA (The Cancer Genome Atlas) and ∼8,000 samples from MMDs (Merged Microarray-acquired Datasets). We found that the dynamic expression alteration of CST6 was consistent with dual function in different types of cancer. In addition, we observed that the expression of CST6 was globally regulated by the DNA methylation in its promoter region. CST6 expression was positively correlated with the epithelial cell infiltration involved in epithelial-to-mesenchymal transition (EMT) and proliferation. The relationship between CST6 and tumor microenvironment was also explored. In particular, we found that CST6 serves a protective function in the process of melanoma metastasis. Finally, the clinical association analysis further revealed the dual function of CST6 in cancer, and a combination of the epithelial cell infiltration and CST6 expression could predict the prognosis for SKCM patients. In summary, this first CST6 pan-cancer study improves the understanding of the dual functional effects on CST6 in different types of human cancer.

Keywords: CST6, pan-cancer, DNA methylation, epithelial cell, EMT, tumor microenvironment, prognosis

OPEN ACCESS

Edited by:
Chunjie Jiang,
University of Pennsylvania,
United States

Reviewed by:
Xuexin Yu,
University of Texas Southwestern Medical Center, United States
Shanlan Mo,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS), China
Wei Chen,
Beihang University, China

Correspondence:
Xiaoman Bi
bixiaoman@hainmc.edu.cn
Zhonglin Mu
muzhonglin2@sina.com
Kongning Li
likongning@hainmc.edu.cn

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

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

Received: *30 June 2021*
Accepted: *17 August 2021*
Published: *17 September 2021*

Citation:
Xu D, Ding S, Cao M, Yu X, Wang H, Qiu D, Xu Z, Bi X, Mu Z and Li K (2021) A Pan-Cancer Analysis of Cystatin E/M Reveals Its Dual Functional Effects and Positive Regulation of Epithelial Cell in Human Tumors. Front. Genet. 12:733211. doi: 10.3389/fgene.2021.733211

INTRODUCTION

Cystatin E/M (also known as CST6) is a member of the cystatin superfamily that performs physiological inhibitors of lysosomal cysteine proteases through forming high-affinity reversible complexes (Turk and Bode, 1991). The dysfunction of CST6 contributed to the alterations in proteolysis of tissue architecture, which might accelerate the spread of cancer cells (Shridhar et al., 2004; Keppler, 2006). Increasing evidence has demonstrated the dual functional effects of CST6 in cancer progress (Lalmanach et al., 2021). For instance, the overexpression of CST6 could rescue mice from bone metastasis by suppressing proliferation, migration, and invasion (Jin et al., 2012). Moreover, the loss of CST6 expression has been observed in lung and cervical cancer, and its recovery expression resulted in growth suppression in culture (Zhong et al., 2007; Veena et al., 2008). In contrast to the protective function, Hosokawa et al. (2008) found that the upregulated expression of CST6 promoted tumor growth *in vitro* and *in vivo* in pancreatic ductal adenocarcinoma. CST6 was also shown overexpressed in triple-negative breast cancer and oral cancer, facilitating the tumor metastatic process (Vigneswaran et al., 2003; Li et al., 2018). Collectively, CST6 played crucial and disparate roles in the pathogenesis and development of cancer. However, a comprehensive research about the expression pattern and functional effects of CST6 in pan-cancer level is still lacking.

As an important epigenetically regulatory factor, DNA methylation has been implicated in the dysfunction of CST6. One study has shown that the hypermethylation status of CST6 promoter resulted in CST6 deficiency in glioma tumor-initiating cells, while the promoter was hypomethylated in normal brain tissues (Qiu et al., 2008). The aberrant methylation and downregulated expression of CST6 were also found in breast cancer patients, and the expression could reactivate after DNA demethylating agent treatment (Ai et al., 2006; Schagdarsurengin et al., 2007). However, the relationship between the expression and the DNA methylation of CST6 in pan-cancer level remains unclear.

Numerous studies have shown the important roles of tumor microenvironment in cancer therapy and diagnosis (Wu and Dai, 2017; Hinshaw and Shevde, 2019). CST6 is an epithelium-specific protease inhibitor with essential roles in epidermal differentiation (Zeeuwen et al., 2010). Zhang et al. (2004) found that CST6 was consistently expressed in normal human breast epithelial cells, while it was decreased in breast invasive carcinoma samples, and the expression of CST6 was associated with cell proliferation, migration, and invasion. The CST6 promoter was found highly methylated in cfDNA of breast cancer plasma cells but not in healthy samples (Chimonidou et al., 2013). Moreover, IL-17A, an immunotherapy targeting, could affect keratinocyte differentiation by regulating the expression of CST6 (Sato et al., 2020). However, there has been limited research that comprehensively explored the relationship between CST6 and tumor microenvironment in pan-cancer level.

In this study, we performed a systematic evaluation of the expression pattern of CST6 across cancer types from TCGA and MMDs. Consistent with the known dual role of CST6, we found that there was a broad spectrum of CST6 expression across cancer types. Through DNA methylation analysis, we found that the expression of CST6 was globally regulated by the methylation level of its promoter region. Moreover, the expression of CST6 was related to epithelial cell infiltration, EMT, and proliferation. Finally, the association between the CST6 expression and patient survival was also investigated. Our first pan-cancer study for CST6 provided novel insights into its dual function in the development of cancer.

MATERIALS AND METHODS

Analysis of Gene Expression

We entered CST6 in the "Gene_DE" module of the TIMER2 website[1] (Li et al., 2021) and explored the expression of CST6 between different tumors and adjacent normal tissues in TCGA items. For some tumors with no normal sample or the number of normal tissue specimens was less than 5, we used the "Expression Analysis-Expression DIY" module of the GEPIA2[2] to compare the expression level of CST6 between tumor tissues and GTEx (Genotype-Tissue Expression) datasets (Tang et al., 2019). The gene with $|\log2FC| > 1$ and p-value < 0.05 was considered as significantly differentially expressed by the ANOVA method. In addition, we used the pathological stage module and subtype filter module in GEPIA2 to obtain the expression of CST6 in different tumors at different stages and different subtypes. In order to verify the expression pattern of CST6 in different cancer types, we collected gene expression data of more than 8,000 samples from 11 cancers (**Supplementary Table 1**; Lim et al., 2019). To avoid differences between platforms to the greatest extent, only the dataset generated from the Affymetrix Human Genome U133 Plus 2.0 array was processed to develop the MMDs dataset. All datasets were processed uniformly through RMA normalization, and batch effect were corrected through the Combat R package (Leek et al., 2012). Moreover, the protein level of CST6 between tumor and normal tissue was obtained from the CPTAC analysis module of the UALCAN portal[3] (Chandrashekar et al., 2017). An external validation dataset was obtained with GEO accession GSE46517, which included 73 metastatic and 31 primary melanoma patients.

DNA Methylation Analysis

We downloaded the gene expression and HM450 DNA methylation profiles across cancer types of TCGA Pan-Cancer (PANCAN) cohort through UCSC Xena.[4] The full name, abbreviation, and sample number of cancer types for TCGA are shown in **Supplementary Tables 2**, **3**. The methylation level of CST6 was quantified by averaging the beta values of CpGs located in the promoter region (upstream 2 kb to TSS). Then, Wilcoxon rank test was used to identify differentially methylated CST6 between tumor and normal tissue. In addition, the correlation between DNA methylation and expression of CST6 was calculated using the Pearson correlation method. The results of correlation coefficient less than -0.3 and p-value < 0.05 were identified as significant.

[1] http://timer.cistrome.org/

[2] http://gepia2.cancer-pku.cn/

[3] http://ualcan.path.uab.edu/analysis-prot.html

[4] http://xena.ucsc.edu/

Cystatin E/M-Related Gene Functional Enrichment Analysis

The CST6-related genes were obtained from the "Similar Gene Detection" module of GEPIA2 by Pearson correlation method. We selected the datasets of all TCGA tumors and finally screened out the top 500 CST6-correlated genes. For a specific cancer type in TCGA and MMDs datasets, we used the Pearson correlation method to obtain CST6-related genes (Pearson correlation $>$ 0.3). To explore the potential biological function that CST6 regulated, we passed the CST6-related genes to Metascape[5] with the setting of species ("Homo sapiens") (Zhou et al., 2019).

Estimation of the Relationship Between Cystatin E/M and Tumor Microenvironment, Epithelial-to-Mesenchymal Transition, and Tumor Proliferation

In order to assess the infiltration levels of epithelial cells in diverse cancers, we downloaded the precalculated TCGA data from xCell,[6] a method that yielded cell type enrichment score for 64 immune and stroma cell types, which included epithelial cell infiltration score (Aran et al., 2017).

We estimated the EMT score for an independent cancer type according to the method of Zhang et al. (2020). The epithelial and mesenchymal genes were obtained from a previous study (Mak et al., 2016). Then, the EMT score, which could reflect the EMT level for each sample, was calculated according to the following formula:

$$\mathrm{S}_{EMT} = \sum_{i}^{N} \frac{M^i}{N} - \sum_{j}^{n} \frac{E^j}{n}$$

where N represents the number of mesenchymal genes, and n represents the number of epithelial genes.

The expression level of the proliferation marker ki67 (MKI67) was used to reflect tumor proliferation in TCGA samples. Then, the correlation between CST6, cell infiltration, and proliferation score were estimated through Spearman's rank-order correlation method. The relationship between EMT score and CST6 was calculated by partial correlation through "ggm" R packages considering tumor purity as concomitant variable.

Survival Prognosis Analysis

The survival analysis of OS (overall survival) and RFS (disease-free survival) for CST6 in TCGA tumors were performed through GEPIA2. The median expression of CST6 was used to divide the patients into high-expression and low-expression groups. Then, the OS and RFS of these groups were compared by log-rank test. In addition, the survival interaction between CST6 expression and epithelial cell (positive genes obtained from xCell) was explored using siGCD[7] (Cui et al., 2021).

RESULTS

Gene Expression Analysis of Cystatin E/M in the the Cancer Genome Atlas Datasets

CST6 could serve as a biomarker for tumor diagnosis and play a dual functional effect across cancer types (Lalmanach et al., 2021). Previous studies of CST6 expression in cancer were limited to sample sizes and focused on a single cancer type. To comprehensively characterize the expression pattern of CST6, we first applied the "Gene_DE" module of TIMER2 portal for TCGA datasets. As shown in **Figure 1A**, the expression level of CST6 in bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cervical and endocervical cancer (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), rectum adenocarcinoma (READ), thyroid carcinoma (THCA), and uterine corpus endometrioid carcinoma (UCEC) is significantly higher than in normal samples, while the expression level of CST6 in kidney chromophobe (KICH), kidney clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), and lung squamous cell adenocarcinoma (LUSC) is significantly decreased than adjacent normal samples. Moreover, the patients with HPV infection showed a lower expression level of CST6 in head and neck squamous cell carcinoma (HNSC), and the patients with metastasis status showed a lower expression in skin cutaneous melanoma (SKCM).

Due to the limited normal sample size for several cancer types, we included the normal tissue of the GTEx datasets and evaluated the expression difference of CST6. As shown in **Figure 1B**, we found that the expression of CST6 was significantly different in CESC, ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), and SKCM. These results were consistent with the tumor-promoting function of CST6 in breast cancer (Li et al., 2018), pancreatic cancer (Hosokawa et al., 2008), and papillary thyroid carcinoma (Oler et al., 2008). The tumor-suppressive function of CST6 has also been reported in lung cancer (Zhong et al., 2007), melanoma (Briggs et al., 2010), and renal cell carcinoma (Morris et al., 2010), which agreed with the loss expression of CST6 in these cancer types.

The results of the CPTAC dataset showed that the expression of CST6 total protein in BRCA and KIRC was lower than that of normal tissues (**Figure 1C**, *t*-test, *p*-value $<$ 0.001). To extend the expression pattern of CST6 with tumor pathological information, we applied the "Stage Plot" and "Subtype Filter" functions of GEPIA2. The expression of CST6 was related to the stage of BLCA, COAD, kidney papillary cell carcinoma (KIRP), OV, READ, SKCM, THCA, and uterine carcinosarcoma (UCS) (**Figure 1D**). In addition, the

[5] https://metascape.org/gp/index.html#/main/step1

[6] https://xcell.ucsf.edu/

[7] http://sigcd.idrug.net.cn/Home

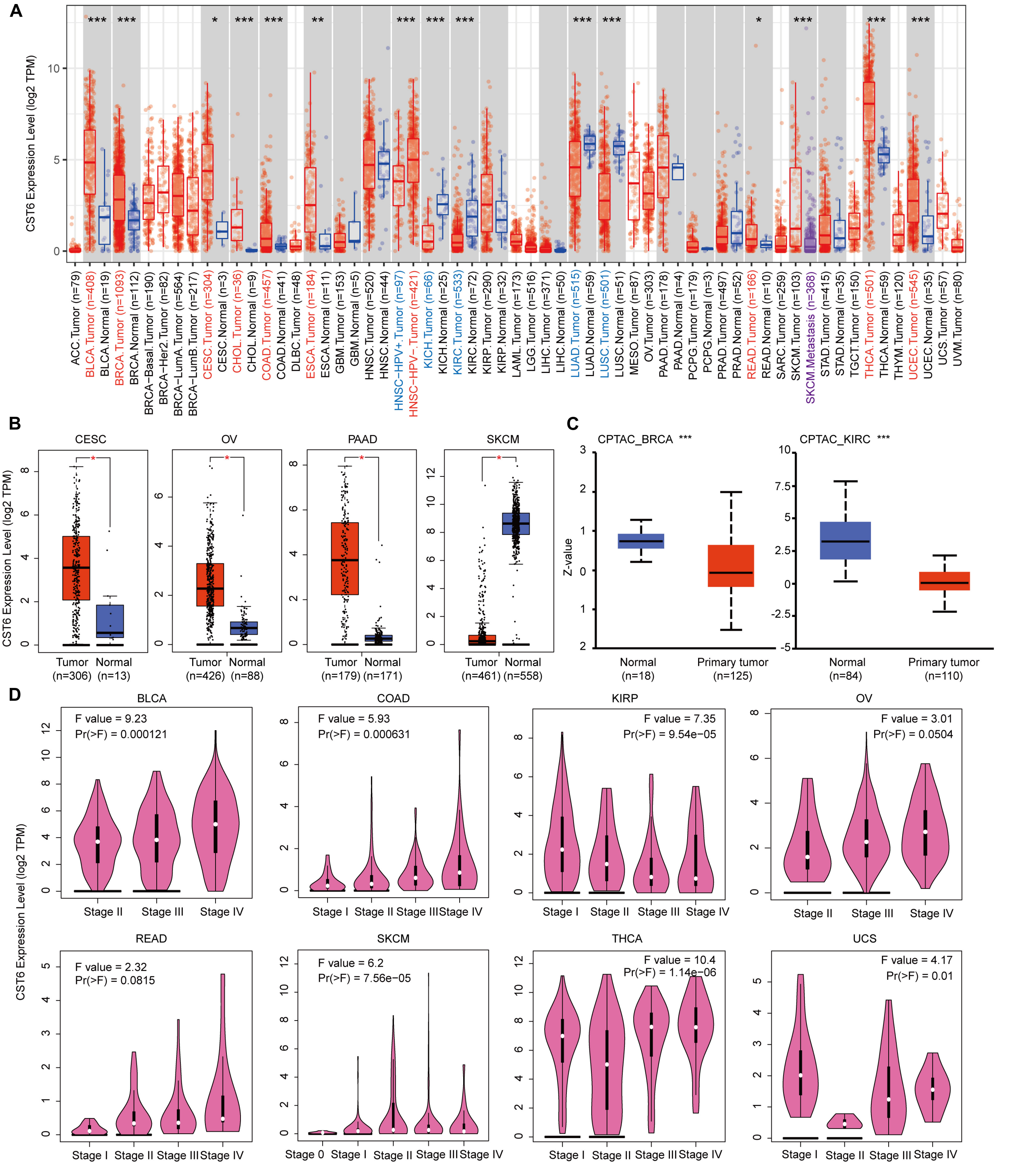


FIGURE 1 | Expression level of cystatin E/M (CST6) in different cancer types and pathological stages. **(A)** The expression level of CST6 in different cancer types from The Cancer Genome Atlas datasets. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. The symbol with red represents upregulated, and the symbol with blue represents downregulated. **(B)** The expression level of CST6 in cervical and endocervical cancer (CESC), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), and skin cutaneous melanoma (SKCM). The corresponding normal samples of the GTEx datasets were included. **(C)** The expression level of CST6 total protein for breast invasive carcinoma (BRCA) and kidney clear cell carcinoma (KIRC) cancer types in the CPTAC dataset. **(D)** The expression level of CST6 was associated with the pathological stages of bladder urothelial carcinoma (BLCA), colon adenocarcinoma (COAD), kidney papillary cell carcinoma (KIRP), OV, rectum adenocarcinoma (READ), SKCM, thyroid carcinoma (THCA), and uterine carcinosarcoma (UCS) cancer types.

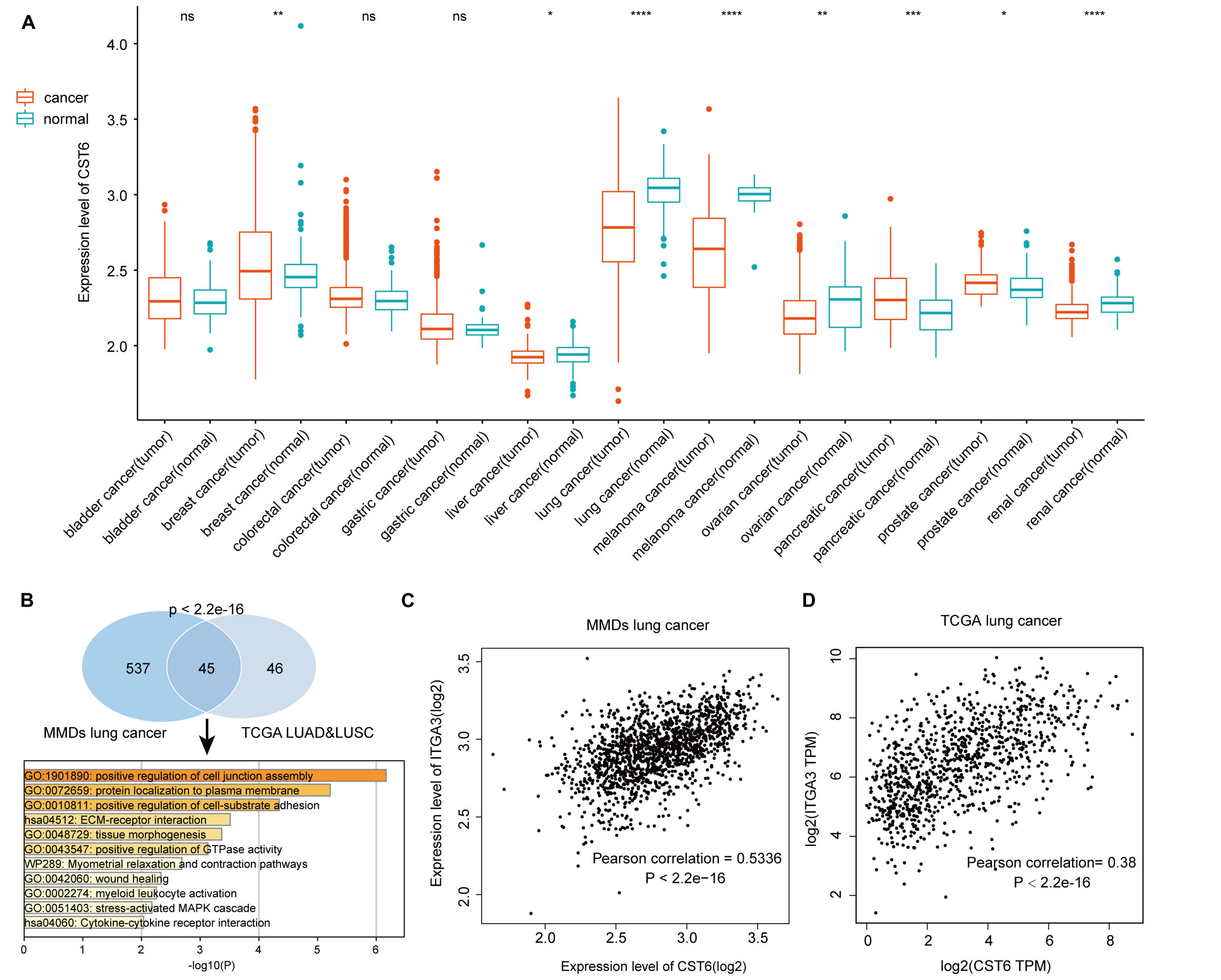


FIGURE 2 | Expression level of CST6 in Merged Microarray-acquired Datasets (MMDs) datasets. **(A)** The expression level of CST6 in different cancer types from MMDs.**p* < 0.05; ***p* < 0.01; ****p* < 0.001; *****p* < 0.0001; ns, not significant. **(B)** The comparison and functional analysis of CST6-related genes between MMDs lung cancer and TCGA lung adenocarcinoma (LUAD) and lung squamous cell adenocarcinoma (LUSC) datasets. **(C,D)** The correlation between the expression of CST6 and integrin α3 (ITGA3) for lung cancer in MMDs and TCGA datasets.

expression of CST6 was significantly different in the subtypes of CESC, HNSC, KIRP, LUAD, LUSC, PAAD, and THCA (**Supplementary Figure 1**).

Gene Expression Analysis of Cystatin E/M in the Merged Microarray-Acquired Datasets

To verify the expression pattern of CST6 across cancer types, we collected the MMDs of more than 7,000 samples from 11 cancers with a standard process. Particularly, we revealed that CST6 showed a higher expression in breast cancer, pancreatic cancer, and prostate cancer, while the expression level of CST6 significantly decreased in liver cancer, lung cancer, melanoma, ovarian cancer, and renal cancer (**Figure 2A**). We found that the expression pattern of CST6 was consistent in breast, pancreatic, lung, melanoma, and renal cancer between TCGA and MMDs datasets. The expression of CST6 was upregulated in BLCA, COAD, and READ for TCGA datasets, while no significantly differently expressed was observed in bladder and colorectal cancer for the MMDs. Moreover, the expression of CST6 was specifically dysregulated in liver, ovarian, and prostate cancer for the MMDs. Regarding the most downregulated cancer type in the two datasets, the coexpression genes of CST6 (Pearson correlation > 0.3) were commonly shared between MMDs lung cancer and TCGA LUAD/LUSC datasets, showing a statistically significant overlap (**Figure 2B**, hypergeometric test, *p*-value < 2.2e-16). We also found that the overlap of CST6-related genes was enriched in the tumor microenvironment-related processes (such as positive regulation of cell junction assembly and protein localization to the plasma membrane, **Figure 2B**).

In addition, integrin α3 (ITGA3) was the most related gene in the two datasets, which showed a positively correlation

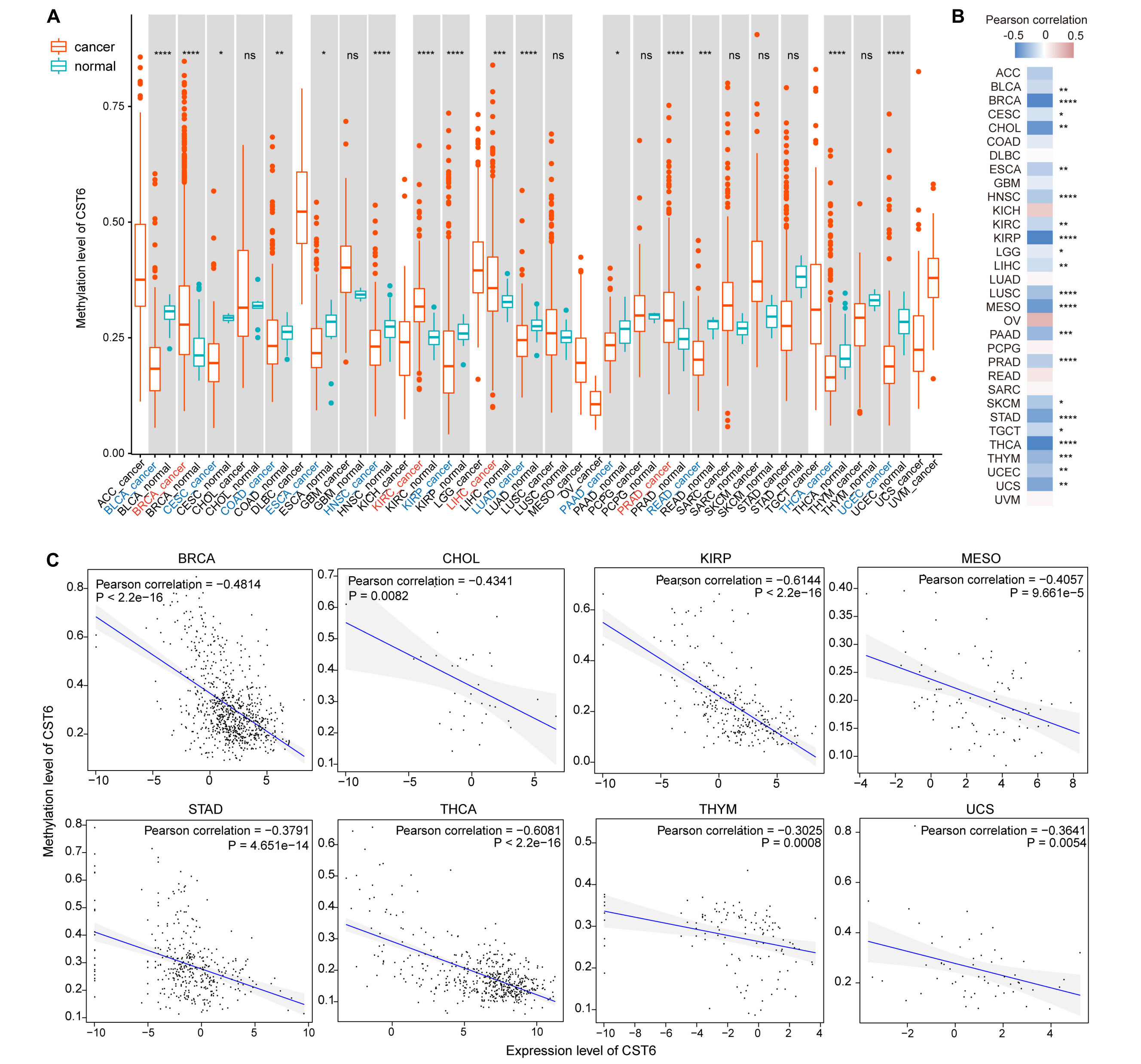


FIGURE 3 | The expression of CST6 was globally regulated by DNA methylation. **(A)** The DNA methylation level of CST6 in different cancer types from TCGA datasets. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; ns, not significant. **(B)** The Pearson correlation coefficients between the expression and methylation level of CST6 in TCGA cancer types. **(C)** The correlation between the expression and methylation level of CST6 in BRCA, CHOL, KIRP, MESO, STAD, THCA, THYM, and UCS cancer types (Pearson correlation coefficient < -0.3, *p*-value < 0.01).

with the expression of CST6 (**Figures 2C,D**). Moreover, a recent study revealed the important prognostic role of ITGA3 in patients with non-small cell lung cancer (Li et al., 2020). Thus, CST6 and ITGA3 may be potential therapeutic targets for lung cancer. Similar results were obtained in melanoma and renal cancer samples. We found that kallikrein-related peptidase 7 (KLK7) and keratin 7 (KRT7) were the most related genes in melanoma and renal cancer separately (**Supplementary Figure 2**). Aberrant expression of KLK7 has found to be related to melanoma aggressiveness by stimulating cell migration and adhesion (Delaunay et al., 2017; Haddada et al., 2018). KRT7 could distinguish the precursor lesions of papillary renal cell tumors, mucinous tubular and spindle cell carcinomas (Szponar and Kovacs, 2012). The detailed information of CST6-related genes list is shown in **Supplementary Table 4**. These results indicated the conservative expression pattern

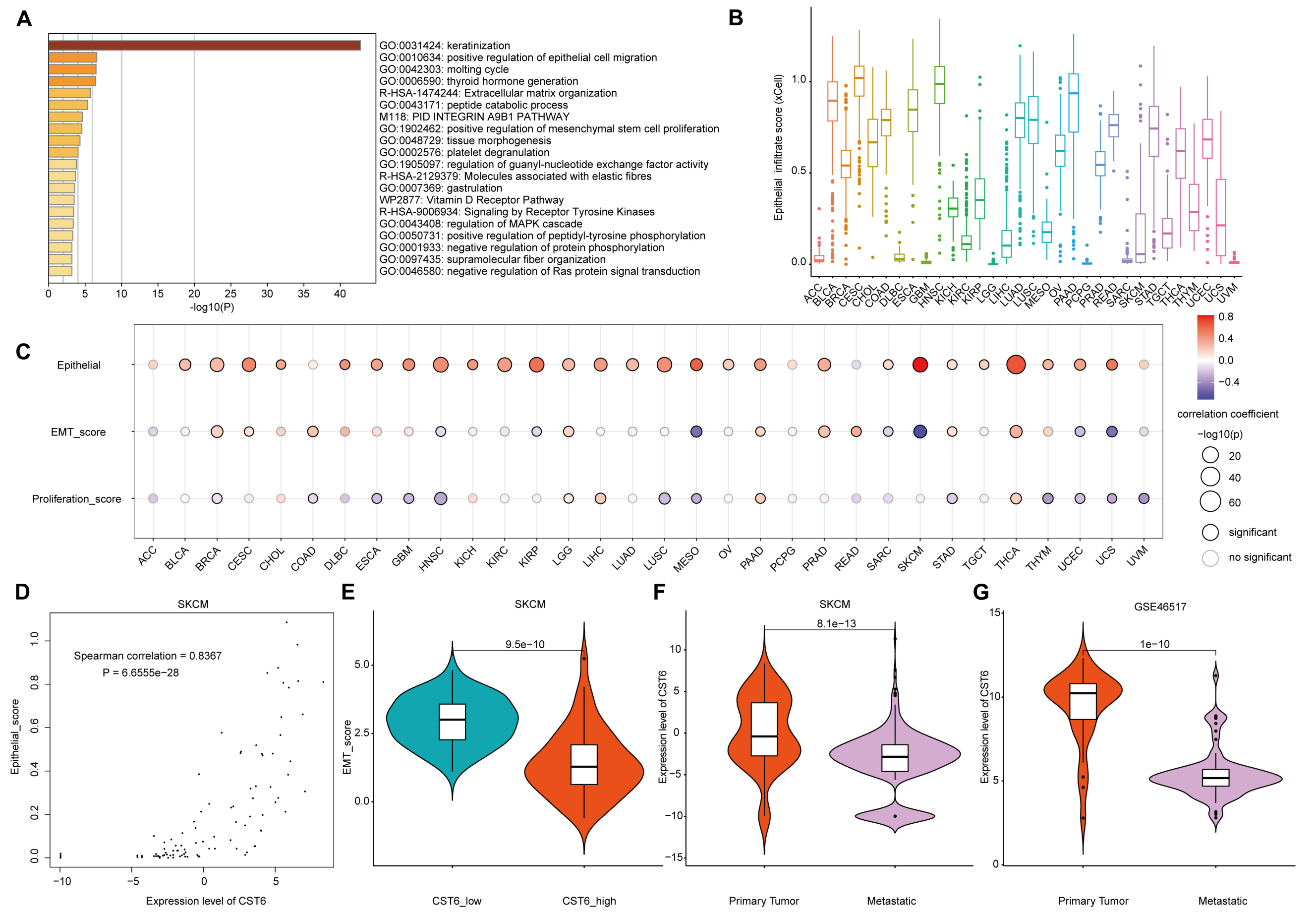


FIGURE 4 | The functional analysis of CST6. **(A)** The functional enrichment analysis based on the top 500 CST6-related genes through Metascape. **(B)** The distribution of epithelial infiltrate score obtained from xCell across TCGA cancer types. **(C)** The Spearman correlation between the expression of CST6 and epithelial cell infiltration, epithelial-to-mesenchymal transition (EMT) score, and proliferation. **(D)** The correlation between the expression of CST6 and the epithelial cell infiltrate score for SKCM in TCGA. **(E)** The violin plot of EMT score between CST6-low and CST6-high groups for SKCM in TCGA. **(F,G)** The violin plot of expression level of CST6 between primary tumor and metastatic samples for SKCM in TCGA and external validation dataset.

and important interactive function of CST6 in different types of cancer.

The Cystatin E/M Expression Was Regulated by DNA Methylation

CST6 expression has been previously associated with its epigenetic regulation by methylation of the promoter region in several cancer types (Rivenbark et al., 2006; Pulukuri et al., 2009; Peters et al., 2014). To explore the relationship between its expression and DNA methylation in TCGA datasets, we first quantified the methylation level of CST6 by averaging the CpG beta value in its promoter region. Then the Wilcoxon test was used to evaluate differentially methylated status between tumor and adjacent normal samples. In contrast to the expression pattern, we found that the methylation level of CST6 was significantly lower in BLCA, CESC, COAD, ESCA, HNSC, KIRP, LUAD, PAAD, READ, THCA, and UCEC tumor samples, while it was higher in BRCA, KIRC, liver hepatocellular carcinoma (LIHC), and prostate adenocarcinoma (PRAD) (Wilcoxon test, p-value < 0.05, **Figure 3A**). Next, we observed a significant negative correlation of CST6 methylation and its expression in more than half (21 out of 32, Pearson correlation < 0 and p-value < 0.05) of the cancer types (**Figure 3B**). The relationship of methylation and expression level of CST6 for the top correlated cancer types (BRCA, CHOL, KIRP, MESO, STAD, THCA, THYM, and USC, Pearson correlation < -0.3 and p-value < 0.01) is shown in **Figure 3C**. All these results further proved the capacity of DNA methylation in regulating the gene expression of CST6.

Functional Analysis of Cystatin E/M in Cancer

Taking advantage of integration of transcriptome and DNA methylation resource, we characterized the expression pattern and DNA methylation regulatory mechanism of CST6. To further investigate the function of CST6 across cancer types, we first obtained 500 CST6-related genes (**Supplementary Table 4**) for all TCGA tumor samples from the GEPIA2 "Correlation Analysis"

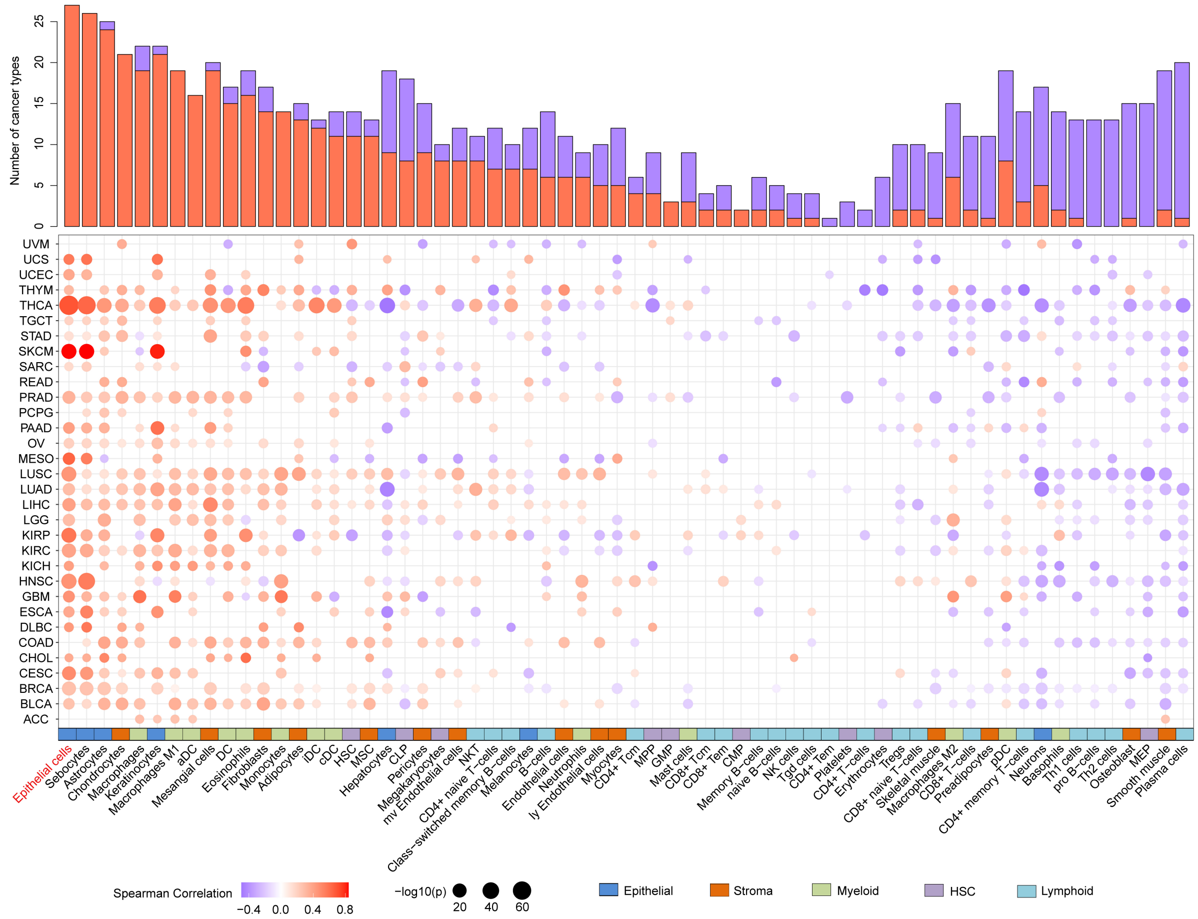


FIGURE 5 | Association analysis between the expression of CST6 and tumor microenvironment in TCGA datasets. Histogram height represents the number of cancer types with significant correlation between CST6 and cell infiltration score. Red- and purple-colored bubbles denote positive and negative correlation. Colored bar on the bottom denotes cell subgroup obtained from xCell.

module. Then, the functional enrichment analysis was performed through Metascape based on the CST6-related gene list. CST6-related genes were significantly enriched in keratinization and positive regulation of epithelial cell migration (**Figure 4A**). Given that the strong correlation between keratin genes and epithelial cell has been reported (Heatley, 2002; Moll et al., 2008), we next explored the relationship of CST6 and epithelial cell in detail. The expression level of CST6 and epithelial cell infiltrate score were varied in cancer types (**Figure 4B** and **Supplementary Figure 3**), while the positive correlation between the two variables was observed in most cancers (**Figure 4C** and **Supplementary Table 5**).

Considering the fact that CST6 has been previously associated with the GO term "positive regulation of mesenchymal stem cell proliferation," we next explored the role that CST6 plays in EMT and proliferation. At first, we downloaded tumor purity for TCGA samples from a previous study (Thorsson et al., 2018). As the EMT score was significantly influenced by tumor purity (**Supplementary Table 6**), we estimated the relationship between EMT and CST6 using partial correlation to remove the confounder. In contrast to epithelial cell infiltration, CST6 showed dual functional effects on EMT and proliferation (**Figure 4C** and **Supplementary Table 5**). Particularly, the correlation coefficients of epithelial cell and EMT score in SKCM were reversed (**Figures 4D,E**), which indicated a potential role of CST6 in the metastasis of melanoma. Given that SKCM has the maximum number of metastatic samples in TCGA, we next compared the expression level of CST6 between metastasis samples and primary tumor tissues. As shown in **Figure 4F**, the expression of CST6 was significantly higher in SKCM primary tumor tissues than that of metastatic samples. Consistent with this finding, a similar pattern was observed in another SKCM metastatic dataset (GSE46517, **Figure 4G**), which further proved the protective effect of CST6 in melanoma metastasis. Taken together, all these results suggest that CST6 was related to epithelial cell infiltration and tumor EMT process.

As evidence has shown the positive correlation between CST6 and epithelial cell, we next explored its relationship with other cells. As shown in **Figure 5**, the expression of CST6 was positively related to most epithelial cells, while the negative correlation between CST6 and plasma cell was observed in most cancer types. A previous study has found that CST6 promoter

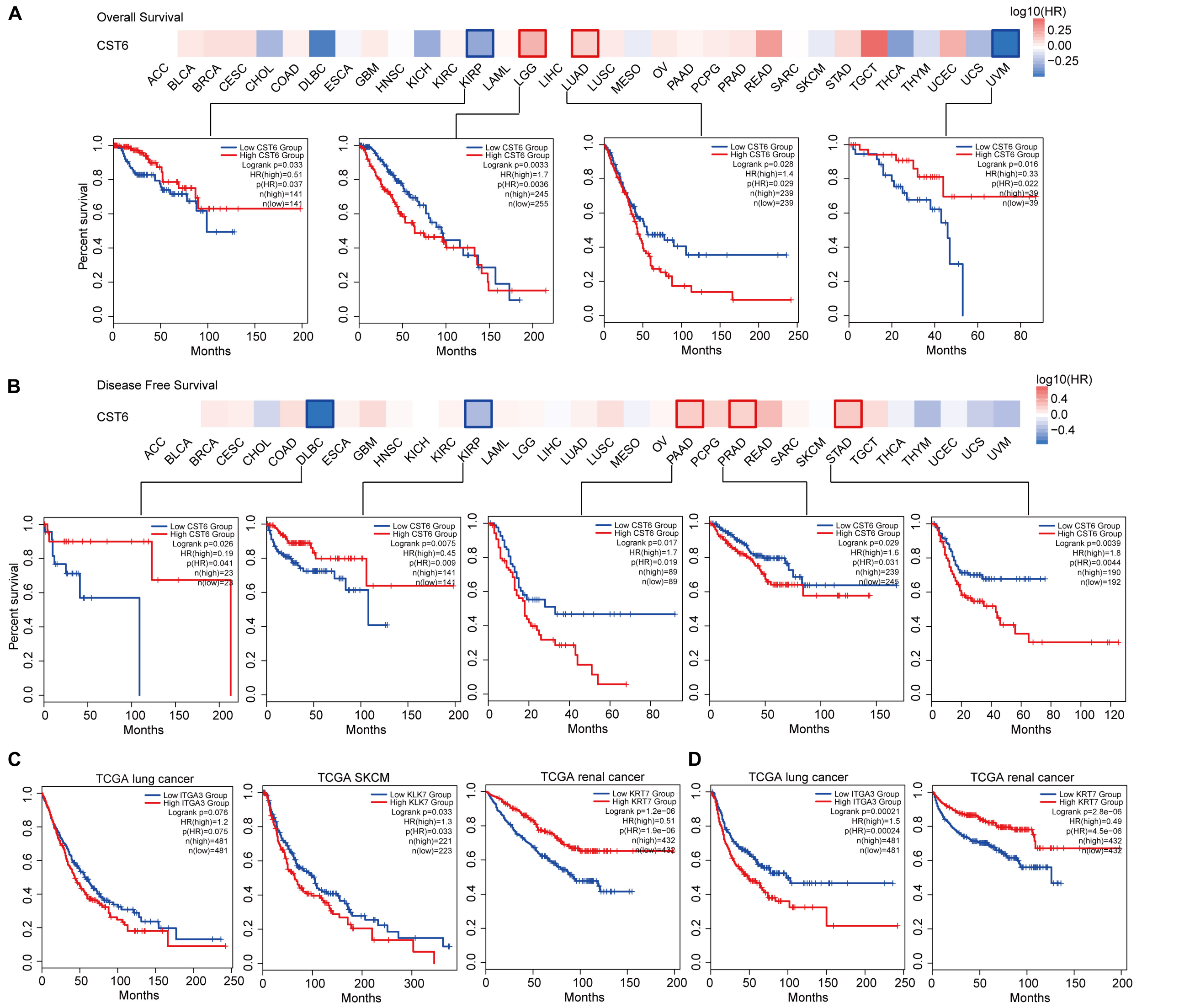


FIGURE 6 | Clinical association analysis between the expression of CST6 and survival prognosis across TCGA cancer types. **(A)** The survival map and Kaplan–Meier estimates of overall survival by CST6 expression in TCGA datasets. **(B)** The survival map and Kaplan–Meier estimates of disease-free survival by CST6 expression in TCGA datasets. **(C)** The Kaplan–Meier estimates of overall survival by ITGA3, KLK7, and KRT7 expression in TCGA lung cancer, melanoma, and renal cancer datasets. **(D)** The Kaplan–Meier estimates of disease-free survival by ITGA3 and keratin 7 (KRT7) expression in TCGA lung and renal cancer datasets.

is highly methylated in cfDNA of BRCA plasma cells but not in healthy samples (Chimonidou et al., 2013). These results suggest the potential regulatory roles of CST6 in the tumor microenvironment.

Clinical Associations of Cystatin E/M in Cancer

We next explored the critical efficiency of CST6 in the survival of tumor patients. Tumor samples were divided into high-expression and low-expression groups based on CST6 expression for each TCGA tumor type. Patients with a higher expression of CST6 had worse survival in brain lower-grade glioma (LGG), LUAD, PAAD, PRAD, and stomach adenocarcinoma (STAD) (HR $>$ 1 and log-rank p $<$ 0.05, **Figures 6A,B**), while they indicated a favorable prognosis in KIRP, uveal melanoma (UVM), and diffuse large B-cell lymphoma (DLBC) (HR $<$ 1 and log-rank p $<$ 0.05, **Figures 6A,B**). The multivariate Cox regression model was also performed with several clinical factors (**Supplementary Figure 4**). In addition, the CST6-related genes mentioned above (ITGA3, LKL7, and KRT7 corresponding to TCGA lung cancer, SKCM, and renal cancer separately) were found to be associated with clinical outcomes (**Figures 6C,D**). These results revealed the dual effects of CST6 on the survival of patient.

The association between epithelial and CST6 for SKCM has been examined herein before. Next, we explored whether these two important elements affected the clinical survival of

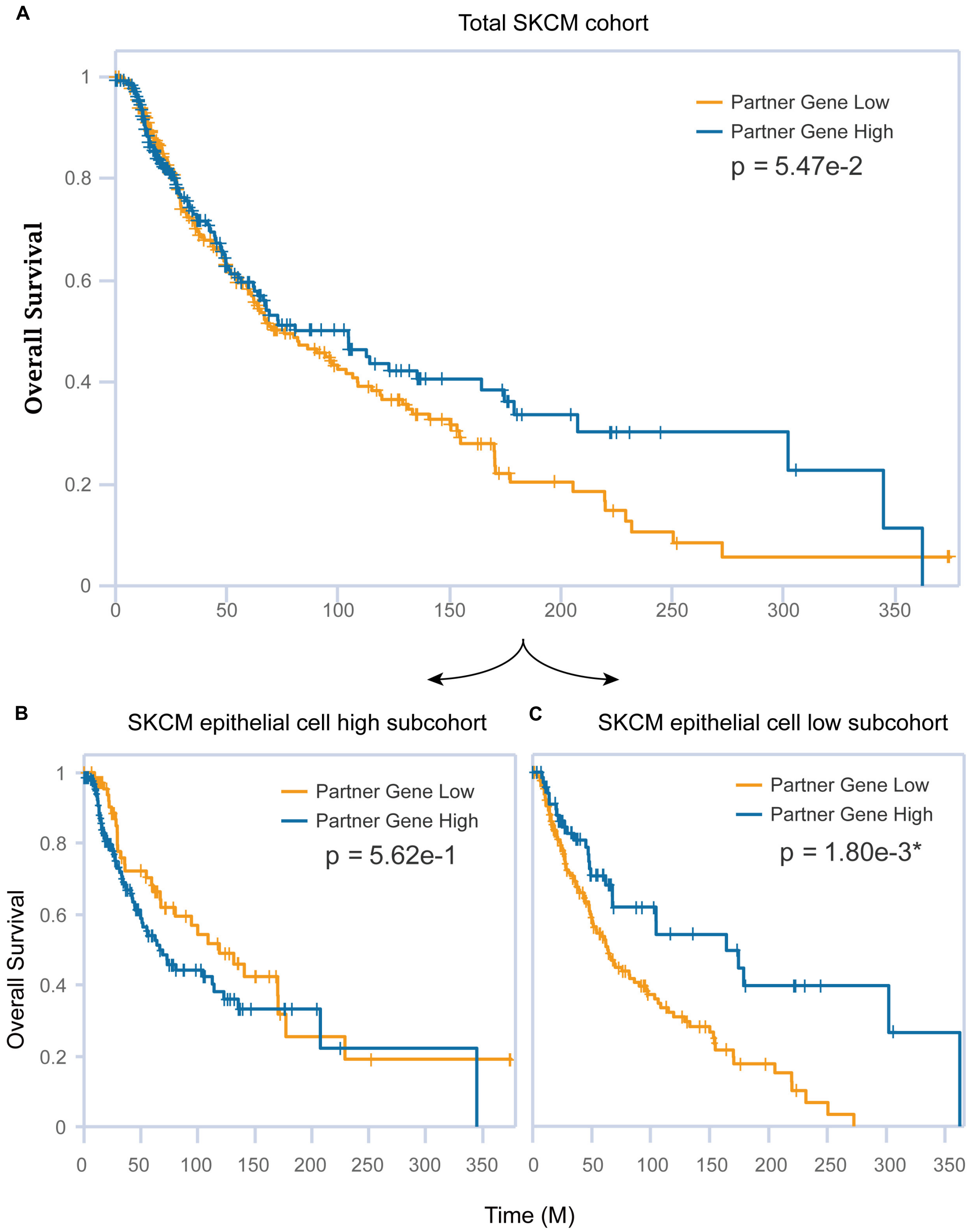


FIGURE 7 | The combination of CST6 expression and epithelial cell infiltration predicted the prognosis for SKCM patients. **(A)** Kaplan–Meier estimates of overall survival (OS) for the total SKCM cohort based on CST6 expression. **(B)** Kaplan–Meier estimates of OS for the SKCM epithelial cell high subcohort based on CST6 expression. **(C)** Kaplan–Meier estimates of OS for the SKCM epithelial cell low subcohort based on CST6 expression.

melanoma patients. The expression of CST6 could not well predict the prognosis of SKCM patients (log-rank p = 0.0547, **Figure 7A**) using siGCD. Meanwhile, we found that the epithelial cell infiltration score of SKCM primary tumor was significantly higher than the metastatic patients (Wilcoxon test, p-value < 0.05, **Supplementary Figure 5A**). Moreover, the expression of CST6 could serve as a protective factor for the clinical survival of metastatic patients, which indicates the important roles of epithelial cell and CST6 in SKCM survival (**Supplementary Figures 5B,C**). Thus, we took the epithelial marker from xCell into consideration. In the case of the epithelial cell low subcohort, the OS of patients with high CST6 expression showed significantly better than those with low scores, while the results did not occur in the epithelial cell high subcohort (**Figures 7B,C**). These results implied that the combination of molecular expression and cell infiltration could better predict the survival of cancer patients.

DISCUSSION AND CONCLUSION

The dual function of CST6 as both tumor suppressing and tumor promoting has been well appreciated (Lalmanach et al., 2021), but its global function and expression pattern in the development of cancer remain largely unknown. Here, we comprehensively characterized the expression pattern of CST6 in cancer from TCGA, and the result was verified from another large sample dataset. Consistent with prior knowledge, the expression of CST6 showed that it was downregulated with tumor-suppressing function, while it showed a reverse level with tumor-promoting function. Evidence has shown the EMT and metastasis functions of ITGA3 in lung cancer, which were similar to the function of CST6 (Li et al., 2020). Meanwhile, we observed a conservative correlation between CST6 and ITGA3 in two datasets (TCGA and MMDs), providing potential therapeutic targets for lung cancer. Apart from ITGA3, we also identified KLK7 and KRT7 as CST6-related genes in melanoma and renal cancer datasets. Overexpression of KLK7 induced a significant reduction in melanoma cell proliferation and colony formation (Delaunay et al., 2017). KRT7 has been proven to be an important biomarker for kidney cancer (Williamson et al., 2017). Taken together, these results indicate the conservative expression pattern and essential interactive function of CST6 in human tumors.

Given that the expression of CST6 exhibited epigenetic inactivation in special cancer types, we explored the relationship between DNA methylation and its expression across cancer types. We found that the expression of CST6 was globally regulated by DNA methylation, especially in BRCA, KIRP, MESO, STAD, THCA, and THYM cancer types. Although we revealed the essential roles of DNA methylation in regulating CST6 expression, we cannot explain its differential expression in some cancer types. Previously, two SNPs in the 5’UTR region of CST6 have been found to be associated with fluconazole susceptibility through a genome-wide association study (Guo et al., 2020). Thus, we counted the number of SNVs located in the CST6 gene region from cBioPortal. Thirteen cancer types with CST6 mutation were identified. Among them, SKCM has the highest alteration frequency (range from 0.19 to 1.36%, **Supplementary Figure 6A**). To explore the effect of SNV on the alteration of CST6 expression, we search the PancanQTL to obtain the CST6-related *cis*-eQTLs (within 1 Mb from the gene transcriptional start site) (Gong et al., 2018). We found more than 30 *cis*-eQTLs in STAD and THCA, and the alternation of rs619701 could improve the expression level of CST6 in THCA (**Supplementary Figures 6B,C** and **Supplementary Table 7**). Integration of SNV data may provide a novel insight for understanding the regulatory mechanism of CST6 in cancer.

Moreover, we found that the expression of CST6 was globally positively correlated with epithelial cell infiltration, suggesting its important roles in the epithelium. Rivenbark et al. (2007) have observed a strong immunostaining phenomenon for CST6 in normal breast epithelial and myoepithelial cells, while it was negative in primary breast tumors. The relevance between CST6 and epithelium encourages us to further explore its relationship with EMT. As the EMT process plays an essential role in cancer metastasis (Brabletz et al., 2018), we found the protective function of CST6 in melanoma metastasis considering the negative correlation between CST6 and EMT score. Although the low-level internalization of CST6 that could affect the migration of melanoma cell has been proven (Wallin et al., 2017), we first revealed the potential mechanism between CST6 and EMT in the melanoma metastasis. Besides, we also found that the combination of epithelium infiltration and CST6 expression could well predict the survival of SKCM patients. Our results suggested the necessity to consider molecular and tumor microenvironment in tumor prognostics.

Our results have been partially limited by the nature of the datasets. Although the variation trend in expression and DNA methylation of CST6 was adverse in most cancer types, a few discordant events were also observed. For instance, the expression and DNA methylation level of CST6 were all upregulated in BRCA tumor samples. This may due to the unbalanced sample size between tumor and normal samples. The discordant of CST6 expression in transcriptome and proteome has also been observed (KIRC was the only cancer type that CST6 showed to be downregulated in both TCGA and CPTAC datasets). Meanwhile, there was a considerable number of lncRNA involved in the CST6-related genes. Thus, the posttranscriptional regulation like non-coding RNA may be another explanation, and this will be the further direction that we will analyze.

In summary, our comprehensive analysis of the expression pattern and dual functional effects of CST6 in pan-cancer level reveals its essential roles. The expression of CST6 was globally regulated by DNA methylation and related to epithelium infiltration. Particularly, CST6 performed a protective function in melanoma metastasis. Dysfunction of CST6 has also shown dual effects in clinical survival in different cancer types.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

KL, ZM, and XB conceived, designed the experiments, finalized, and submitted the manuscript. DX, SD, MC, XY, HW, DQ, and

ZX analyzed the data. DX, SD, and MC drafted the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the Hainan Provincial Natural Science Foundation of China (820RC637 and 820RC627), National Natural Science Foundation of China (32160152), Major Science and Technology Program of Hainan Province (ZDKJ202003), and Hainan Province Clinical Medical Center.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.733211/full#supplementary-material

REFERENCES

Ai, L., Kim, W. J., Kim, T. Y., Fields, C. R., Massoll, N. A., Robertson, K. D., et al. (2006). Epigenetic silencing of the tumor suppressor cystatin M occurs during breast cancer progression. *Cancer Res.* 66, 7899–7909. doi: 10.1158/0008-5472.CAN-06-0576

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 18, 1–14. doi: 10.1186/s13059-017-1349-1

Brabletz, T., Kalluri, R., Nieto, M. A., and Weinberg, R. A. (2018). EMT in cancer. *Nat. Rev. Cancer* 18, 128–134. doi: 10.1038/nrc.2017.118

Briggs, J. J., Haugen, M. H., Johansen, H. T., Riker, A. I, Abrahamson, M., Fodstad, Ø, et al. (2010). Cystatin E/M suppresses legumain activity and invasion of human melanoma. *BMC Cancer* 10:17. doi: 10.1186/1471-2407-10-17

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 (United States)* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Chimonidou, M., Tzitzira, A., Strati, A., Sotiropoulou, G., Sfikas, C., Malamos, N., et al. (2013). CST6 promoter methylation in circulating cell-free DNA of breast cancer patients. *Clin. Biochem.* 46, 235–240. doi: 10.1016/j.clinbiochem.2012.09.015

Cui, X., Han, L., Liu, Y., Li, Y., Sun, W., Song, B., et al. (2021). siGCD: a web server to explore survival interaction of genes, cells and drugs in human cancers. *Brief. Bioinform.* bbab058. doi: 10.1093/bib/bbab058 [Epub ahead of print].

Delaunay, T., Deschamps, L., Haddada, M., Walker, F., Soosaipillai, A., Soualmia, F., et al. (2017). Aberrant expression of kallikrein-related peptidase 7 is correlated with human melanoma aggressiveness by stimulating cell migration and invasion. *Mol. Oncol.* 11, 1330–1347. doi: 10.1002/1878-0261.12103

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., et al. (2018). PancanQTL: systematic identification of cis -eQTLs and trans -eQTLs in 33 cancer types. *Nucleic Acids Res.* 46, D971–D976. doi: 10.1093/nar/gkx861

Guo, X., Zhang, R., Li, Y., Wang, Z., Ishchuk, O. P., Ahmad, K. M., et al. (2020). Understand the genomic diversity and evolution of fungal pathogen *Candida glabrata* by genome-wide analysis of genetic variations. *Methods* 176, 82–90. doi: 10.1016/j.ymeth.2019.05.002

Haddada, M., Draoui, H., Deschamps, L., Walker, F., Delaunay, T., Brattsand, M., et al. (2018). Kallikrein-related peptidase 7 overexpression in melanoma cells modulates cell adhesion leading to a malignant phenotype. *Biol. Chem.* 399, 1099–1105. doi: 10.1515/hsz-2017-0339

Heatley, M. K. (2002). Keratin expression in human tissues and neoplasms [1]. *Histopathology* 41, 365–366. doi: 10.1046/j.1365-2559.2002.15261.x

Hinshaw, D. C., and Shevde, L. A. (2019). The tumor microenvironment innately modulates cancer progression. *Cancer Res.* 79, 4557–4566. doi: 10.1158/0008-5472.CAN-18-3962

Hosokawa, M., Kashiwaya, K., Eguchi, H., Ohigashi, H., Ishikawa, O., Furihata, M., et al. (2008). Over-expression of cysteine proteinase inhibitor cystatin 6 promotes pancreatic cancer growth. *Cancer Sci.* 99, 1626–1632. doi: 10.1111/j.1349-7006.2008.00869.x

Jin, L., Zhang, Y., Li, H., Yao, L., Fu, D., Yao, X., et al. (2012). Differential secretome analysis reveals CST6 as a suppressor of breast cancer bone metastasis. *Cell Res.* 22, 1356–1373. doi: 10.1038/cr.2012.90

Keppler, D. (2006). Towards novel anti-cancer strategies based on cystatin function. *Cancer Lett.* 235, 159–176. doi: 10.1016/j.canlet.2005.04.001

Lalmanach, G., Kasabova-Arjomand, M., Lecaille, F., and Saidi, A. (2021). Cystatin m/e (Cystatin 6): a janus-faced cysteine protease inhibitor with both tumor-suppressing and tumor-promoting functions. *Cancers (Basel).* 13, 1–18. doi: 10.3390/cancers13081877

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

Li, Q., Ma, W., Chen, S., Tian, E. C., Wei, S., Fan, R. R., et al. (2020). High integrin a3 expression is associated with poor prognosis in patients with non-small cell lung cancer. *Transl. Lung Cancer Res.* 9, 1361–1378. doi: 10.21037/tlcr-19-633

Li, Q., Zheng, Z. C., Ni, C. J., Jin, W. X., Jin, Y. X., Chen, Y., et al. (2018). Correlation of cystatin E/M with clinicopathological features and prognosis in triple-negative breast cancer. *Ann. Clin. Lab. Sci.* 48, 40–44.

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2021). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 48, W509–W514. doi: 10.1093/NAR/GKAA407

Lim, S. B., Chua, M. L. K., Yeong, J. P. S., Tan, S. J., Lim, W.-T., and Lim, C. T. (2019). Pan-cancer analysis connects tumor matrisome to immune response. *NPJ Precis. Oncol.* 3:15. doi: 10.1038/s41698-019-0087-0

Mak, M. P., Tong, P., Diao, L., Cardnell, R. J., Gibbons, D. L., William, W. N., et al. (2016). A patient-derived, pan-cancer EMT signature identifies global molecular alterations and immune target enrichment following epithelial-to-mesenchymal transition. *Clin. Cancer Res.* 22, 609–620. doi: 10.1158/1078-0432.CCR-15-0876

Moll, R., Divo, M., and Langbein, L. (2008). The human keratins: biology and pathology. *Histochem. Cell Biol.* 129, 705–733. doi: 10.1007/s00418-008-0435-6

Morris, M. R., Ricketts, C., Gentle, D., Abdulrahman, M., Clarke, N., Brown, M., et al. (2010). Identification of candidate tumour suppressor genes frequently methylated in renal cell carcinoma. *Oncogene* 29, 2104–2117. doi: 10.1038/onc.2009.493

Oler, G., Camacho, C. P., Hojaij, F. C., Michaluart, P., Riggins, G. J., and Cerutti, J. M. (2008). Gene expression profiling of papillary thyroid carcinoma identifies transcripts correlated with BRAF mutational status and lymph node metastasis. *Clin. Cancer Res.* 14, 4735–4742. doi: 10.1158/1078-0432.CCR-07-4372

Peters, I., Dubrowinskaja, N., Abbas, M., Seidel, C., Kogosov, M., Scherer, R., et al. (2014). DNA methylation biomarkers predict progression-free and overall survival of metastatic renal cell cancer (mRCC) treated with antiangiogenic therapies. *PLoS One* 9:e91440. doi: 10.1371/journal.pone.0091440

Pulukuri, S. M., Gorantla, B., Knost, J. A., and Rao, J. S. (2009). Frequent loss of cystatin E/M expression implicated in the progression of prostate cancer. *Oncogene* 28, 2829–2838. doi: 10.1038/onc.2009.134

Qiu, J., Ai, L., Ramachandran, C., Yao, B., Gopalakrishnan, S., Fields, C. R., et al. (2008). Invasion suppressor cystatin E/M (CST6): high-level cell type-specific expression in normal brain and epigenetic silencing in gliomas. *Lab. Invest.* 88, 910–925. doi: 10.1038/labinvest.2008.66

Rivenbark, A. G., Jones, W. D., and Coleman, W. B. (2006). DNA methylation-dependent silencing of CST6 in human breast cancer cell lines. *Lab. Invest.* 86, 1233–1242. doi: 10.1038/labinvest.3700485

Rivenbark, A. G., Livasy, C. A., Boyd, C. E., Keppler, D., and Coleman, W. B. (2007). Methylation-dependent silencing of CST6 in primary human breast tumors and metastatic lesions. *Exp. Mol. Pathol.* 83, 188–197. doi: 10.1016/j.yexmp.2007.03.008

Sato, E., Yano, N., Fujita, Y., and Imafuku, S. (2020). Interleukin-17A suppresses granular layer formation in a 3-D human epidermis model through regulation of terminal differentiation genes. *J. Dermatol.* 47, 390–396. doi: 10.1111/1346-8138.15250

Schagdarsurengin, U., Pfeifer, G. P., and Dammann, R. (2007). Frequent epigenetic inactivation of cystatin M in breast carcinoma. *Oncogene* 26, 3089–3094. doi: 10.1038/sj.onc.1210107

Shridhar, R., Zhang, J., Song, J., Booth, B. A., Kevil, C. G., Sotiropoulou, G., et al. (2004). Cystatin M suppresses the malignant phenotype of human MDA-MB-435S cells. *Oncogene* 23, 2206–2215. doi: 10.1038/sj.onc.1207340

Szponar, A., and Kovacs, G. (2012). Expression of KRT7 and WT1 differentiates precursor lesions of Wilms' tumours from those of papillary renal cell tumours and mucinous tubular and spindle cell carcinomas. *Virchows Arch.* 460, 423–427. doi: 10.1007/s00428-012-1209-z

Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* 47, W556–W560. doi: 10.1093/nar/gkz430

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., et al. (2018). The immune landscape of cancer. *Immunity* 48, 812–830.e14. doi: 10.1016/j.immuni.2018.03.023

Turk, V., and Bode, W. (1991). The cystatins: protein inhibitors of cysteine proteinases. *FEBS Lett.* 285, 213–219. doi: 10.1016/0014-5793(91)80804-C

Veena, M. S., Lee, G., Keppler, D., Mendonca, M. S., Redpath, J. L., Stanbridge, E. J., et al. (2008). Inactivation of the cystatin E/M tumor suppressor gene in cervical cancer. *Genes Chromosomes Cancer* 47, 740–754. doi: 10.1002/gcc.20576

Vigneswaran, N., Wu, J., and Zacharias, W. (2003). Upregulation of cystatin M during the progression of oropharyngeal squamous cell carcinoma from primary tumor to metastasis. *Oral Oncol.* 39, 559–568. doi: 10.1016/S1368-8375(03)00038-1

Wallin, H., Apelqvist, J., Andersson, F., Ekström, U., and Abrahamson, M. (2017). Low-level internalization of cystatin E/M affects legumain activity and migration of melanoma cells. *J. Biol. Chem.* 292, 14413–14424. doi: 10.1074/jbc.M117.776138

Williamson, S. R., Gadde, R., Trpkov, K., Hirsch, M. S., Srigley, J. R., Reuter, V. E., et al. (2017). Diagnostic criteria for oncocytic renal neoplasms: a survey of urologic pathologists. *Hum. Pathol.* 63, 149–156. doi: 10.1016/j.humpath.2017.03.004

Wu, T., and Dai, Y. (2017). Tumor microenvironment and therapeutic response. *Cancer Lett.* 387, 61–68. doi: 10.1016/j.canlet.2016.01.043

Zeeuwen, P. L. J. M., van Vlijmen-Willems, I. M. J. J., Cheng, T., Rodijk-Olthuis, D., Hitomi, K., Hara-Nishimura, I., et al. (2010). The cystatin M/E-cathepsin L balance is essential for tissue homeostasis in epidermis, hair follicles, and cornea. *FASEB J.* 24, 3744–3755. doi: 10.1096/fj.10-155879

Zhang, J., Shridhar, R., Dai, Q., Song, J., Barlow, S. C., Yin, L., et al. (2004). Cystatin m: a novel candidate tumor suppressor gene for breast cancer. *Cancer Res.* 64, 6957–6964. doi: 10.1158/0008-5472.CAN-04-0819

Zhang, Z., Jing, J., Ye, Y., Chen, Z., Jing, Y., Li, S., et al. (2020). Characterization of the dual functional effects of heat shock proteins (HSPs) in cancer hallmarks to aid development of HSP inhibitors. *Genome Med.* 12, 1–16. doi: 10.1186/s13073-020-00795-6

Zhong, S., Fields, C. R., Su, N., Pan, Y. X., and Robertson, K. D. (2007). Pharmacologic inhibition of epigenetic modifications, coupled with gene expression profiling, reveals novel targets of aberrant DNA methylation and histone deacetylation in lung cancer. *Oncogene* 26, 2621–2634. doi: 10.1038/sj.onc.1210041

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523. doi: 10.1038/s41467-019-09234-6

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 October 2021
doi: 10.3389/fgene.2021.722835

Genetic and Epigenetic Impact of Chronic Inflammation on Colon Mucosa Cells

Jia He[1,2†], Jimin Han[2†], Jia Liu[2†], Ronghua Yang[1], Jingru Wang[1], Xusheng Wang[2] and Xiaodong Chen[1*]*

[1]Department of Burn Surgery, The First People's Hospital of Foshan, Foshan, China, [2]School of Pharmaceutical Sciences (Shenzhen), Sun Yat-sen University, Guangzhou, China

Chronic inflammation increases cancer risk, and cancer development is characterized by stepwise accumulation of genetic and epigenetic alterations. During chronic inflammation, infectious agents and intrinsic mediators of inflammatory responses can induce genetic and epigenetic changes. This study tried to evaluate both the genetic and epigenetic influence of chronic inflammation on colon mucosa cells. Repetitive dextran sulfate sodium (DSS) treatment induced chronic colitis model. Whole-exome sequencing (WES) (200× coverage) was performed to detect somatic variations in colon mucosa cells. With the use of whole-genome bisulfite sequencing (BS) at 34-fold coverage (17-fold per strand), the methylome of both the colitis and control tissue was comparatively analyzed. Bioinformatics assay showed that there was no significant single-nucleotide polymorphism/insertion or deletion (SNP/InDel) mutation accumulation in colitis tissue, while it accumulated in aged mice. Forty-eight genes with SNP/InDel mutation were overlapped in the three colitis tissues, two (Wnt3a and Lama2) of which are in the cancer development-related signaling pathway. Differentially methylated region (DMR) assay showed that many genes in the colitis tissue are enriched in the cancer development-related signaling pathway, such as PI3K–AKT, Ras, Wnt, TGF-beta, and MAPK signaling pathway. Together, these data suggested that even though chronic inflammation did not obviously increase genetic mutation accumulation, it could both genetically and epigenetically alter some genes related to cancer development.

Keywords: chronic colitis, chronic inflammation, SNP/indel, DNA methylation, cancer, aging

OPEN ACCESS

Edited by:
Shengli Li,
Shanghai Jiao Tong University, China

Reviewed by:
Jianhui Chen,
The First Affiliated Hospital of Sun Yat-sen University, China
Huimei Wang,
Fudan University, China

****Correspondence:***
Xiaodong Chen
cxd234@163.com
Xusheng Wang
wangxsh27@mail.sysu.edu.cn

†These authors have contributed equally to this work

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

Received: *09 June 2021*
Accepted: *13 September 2021*
Published: *26 October 2021*

Citation:
He J, Han J, Liu J, Yang R, Wang J, Wang X and Chen X (2021) Genetic and Epigenetic Impact of Chronic Inflammation on Colon Mucosa Cells. Front. Genet. 12:722835. doi: 10.3389/fgene.2021.722835

INTRODUCTION

Chronic inflammation has been indicated as an important risk factor for cancer; one of the best examples of the association between chronic inflammation and cancer is found in the heightened predisposition for cancer of patients suffering from ulcerative colitis (UC) and Crohn's disease of the colon, the major forms of idiopathic inflammatory bowel disease (McLarnon, 2011; Risques et al., 2011). Extensive and chronic UC leads to a 19-fold increase in risk for colon cancer, and about 5% of UC patients develop tumors (Gillen et al., 1994). And pancreatic inflammation is a key risk factor for pancreatic cancer (Maisonneuve and Lowenfels, 2002; Raimondi et al., 2010). Another major disease linked to chronic inflammation is gastric cancer, the second leading cause of cancer death worldwide (Qadri et al., 2014; Senol et al., 2014), in which the predisposing inflammation is most often caused by colonization of the gastric epithelium by *Helicobacter pylori*, and chronically infected individuals have an increased risk of developing gastric cancer (Helicobacter and Cancer C, 2001; Meira et al., 2008).

During inflammation, there are increased levels of reactive oxygen and nitrogen species (RONS), which can induce cytotoxic and mutagenic DNA lesions, including abasic sites, oxidized bases (e.g., 8oxoG), deaminated bases (e.g., uracil and hypoxanthine), and ethenoadenine (εA) (Lonkar and Dedon, 2011). In addition to base damage, RONS could also induce DNA double-strand breaks (DSBs). DSBs are among the most toxic of DNA lesions. Moreover, DSBs can also be potently mutagenic due to the potential loss of vast stretches of chromosomes if not accurately repaired (Hoeijmakers, 2009; Maynard et al., 2009). As an unwanted result, these wide range of genomic alterations, including point mutations, copy number changes, and rearrangements, can lead to the development of cancer (Meyerson et al., 2010).

Genomic sequencing has developed to be an effective alternative to locus-specific and gene-panel tests in a research setting for establishing a new genetic basis of disease (de Ligt et al., 2012; Yang et al., 2013). Whole-exome sequencing (WES) is a next-generation technology to determine the variations of the coding regions (exons) of a genome. WES provides coverage of more than 95% exons, which contains 85% disease-causing mutations in many disease-predisposing single-nucleotide polymorphisms (SNPs) throughout the genome (Kaname et al., 2014; Rabbani et al., 2014). Somatic mutations in tumor genome are extensively explored, while the characterization of chronic inflammation-induced somatic mutation via WES approaches is not deeply explored yet, especially the quantitative expansion of different types of genomic alteration during a certain period of inflammation.

Epigenetic alterations, in particular alterations in DNA methylation, are involved in inflammation-associated carcinogenesis (Hartnett and Egan, 2012). Studies have found hypermethylation for p14^{ARF}, p16^{INK4a}, estrogen receptor, and many other genes in human patients with colitis-associated cancers (Tominaga et al., 2005; Dhir et al., 2008; Wang et al., 2008; Yu et al., 2014). However, how DNA methylation alterations contribute to inflammation-associated carcinogenesis is still unclear. Reports showed that methylation in the promoter region of the upstream area could inhibit gene expression, while gene body methylation was positively correlated with gene expression, which prevented transcription from being too active and related to gene disorder (Ehrlich, 2009; Ndlovu et al., 2011). Therefore, it is valuable to understand the mechanism of upstream cell signal and gene body methylation crosstalk. It can be a guide to understand the pathological and regulation mechanism in the process of inflammation to carcinogenesis.

Here, we present the characteristic of the somatic variation of exome in colon mucosa cells of three chronic colitis mice via WES approaches, the colitis was induced via dextran sulfate sodium (DSS) repeatedly for 40 weeks, and age-matched and no-DSS-treated mice serve as the control mice. In addition, the exome of 8-week-old mouse colon mucosa cells was also sequenced, which is compared with the exome of the 56-week-old (56W) mice to evaluate accumulation of somatic mutation through aging. Our data showed that there was no significant difference in quantification of SNP/insertion or deletion (InDel) mutation between colitis and the control mice, and the similar result appeared in the copy number variation (CNV) number. Furthermore, we found that the SNP/InDel number was obviously elevated in the older mice compared with the young mice, suggesting that aging could make significant contribution to accumulation of somatic mutation. We also performed the DNA methylation sequencing between control and DSS group in 56W mice to explore how DNA methylation alterations in chronic inflammation act upon carcinogenesis. Differentially methylated region (DMR) assay indicated methylation in the upstream, downstream, and gene body regions was significant different. Subsequently, functional analysis showed that differentially methylated genes in chronic inflammation are enriched in the signal pathways related to carcinogenesis. These all suggest that a part of genes related to cancer development appears to have both genetic and epigenetic alterations by chronic inflammation.

METHODS

Dextran Sulfate Sodium-Induced Chronic Colitis

Distilled water with 2.5% DSS replaced the distilled water for 7 days, during which colitis was induced and the mouse body weight decreased by about 18–20%. Then at the 8 day, DSS water was replaced by distilled water for another 7 days, and then the distilled water was replaced by 2.5% DSS water again. This kind of procedure (7 days of water with 2.5% DSS/ 7 days of distilled water) was repeatedly performed for 20 times, and the mouse body weight is monitored weekly. Eight weeks after the last DSS treatment procedure, mice were sacrificed for colon mucosa cells and muscularis mucosae isolation.

DNA Quantification and Qualification

DNA degradation and contamination were monitored on 1% agarose gels. Then DNA purity was checked using the NanoPhotometer spectrophotometer (IMPLEN, Westlake Village, CA, United States). Subsequently, DNA concentration was measured using Qubit DNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, Carlsbad, CA, United States). Fragment distribution of DNA library was measured using the DNA Nano 6000 Assay Kit of Agilent Bioanalyzer 2,100 system (Agilent Technologies, Santa Clara, CA, United States).

Whole-Exome Sequencing Library Generation

A total amount of 1 μg of genomic DNA per sample was used as input material for the DNA sample preparation. Sequencing libraries were generated using Agilent SureSelect Mouse All Exon Kit (Agilent Technologies, Santa Clara, CA, United States) following the manufacturer's recommendations, and index codes were added to attribute sequences to each sample. Briefly, fragmentation was carried out by hydrodynamic shearing system (Covaris, Woburn, MA, United States) to generate fragments of 180–280 bp. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities, and enzymes were removed. After adenylation of 3′ ends of DNA fragments, adapter oligonucleotides were ligated. DNA fragments with ligated adapter molecules on both ends were selectively enriched in a PCR. After PCR, library hybridize with liquid phase with a biotin-labeled probe and then use magnetic beads with streptomycin to capture the 220,000 exons within 24,000 genes. Captured libraries were enriched in a PCR to add index tags to prepare for hybridization. Products were purified using AMPure XP system (Beckman Coulter, Beverly, MA, United States) and quantified using the Agilent high-sensitivity DNA assay on the Agilent Bioanalyzer 2,100 system.

Whole-Exome Sequencing and Quality Control

The original image data generated by the sequencing machine were converted into sequence data via base calling (Illumina pipeline CASAVA v1.8.2) and then subjected to quality control (QC) procedure to remove unusable reads: 1) the reads contain the Illumina library construction adapters; 2) the reads contain more than 10% unknown bases (N bases); and 3) one end of the read contains more than 50% of low-quality bases (sequencing quality value ≤ 5).

Whole-Exome Sequencing Read Mapping

Sequencing reads were aligned to the reference genome using BWA with default parameters. Subsequent processing, including duplicate removal was performed using samtools and PICARD (http://picard.sourceforge.net).

Variant Detection and Annotation

The raw SNP/InDel sets are called by samtools with the parameters as "-q 1 -C 50 -m 2 -F 0.002 -d 1,000." Then we filtered these sets using the following criteria: 1) the mapping quality >20 and 2) the depth of the variate position >4. BreakDancer and CNVnator were used for structural variation (SV) and CNV detections, respectively. ANNOVAR was used for functional annotation of variants. The UCSC known genes were used for gene and region annotations.

Library Construction and Methylated Sequencing

After extraction of genomic DNAs of samples, first, the sample is tested for quality. After the sample quality is qualified, the DNA libraries for bisulfite sequencing were carried out. Specific steps are as follows: the genomic DNA first ultrasound interrupted into the 100–300 bp by Sonication (Covaris) and purified with MiniElute PCR Purification Kit (QIAGEN, Valencia, CA, United States). DNA fragment terminal repaired, 3′ end plus "A" nucleotide base connect the sequencing connector. Methylated sequencing adapters ligate to the genomic fragments. Using ZYMO EZ DNA Methylation-Gold kit ligates methylated sequencing adapters. After desalting treatment, the adhesive is recycled, and the library fragment size selection was performed. Select library fragment size again after PCR amplification. After the construction of the library was completed, the quality inspection of the library was performed. The qualified library will be used for sequencing, which uses Illumina HiSeq™ 2,500 by Gene Denovo Biotechnology Co (Guangzhou, China). The original reads were filtered based on the following rules for getting high-quality clean reads: 1) if reads contain more than 10% unknown nucleotides (N), they will be removed; 2) if reads contain more than 40% of low-quality (Q-value ≤20) bases, low-quality reads will be removed.

Methylation Level Analysis

The standard clean reading map obtained by BSMAP software (version 2.90) was mapped to the mouse reference genome. Self-defined Perl script to call methylated cytosine and the methylation level was calculated based on the percentage of methylated cytosine in the whole genome, in each chromosome (CG) and in different regions of the genome in each sequence context (CHG and CHH). Subsequently, a 2-kb region methylation profile was drawn based on the average methylation level of each 100-bp interval in order to evaluate different methylation patterns in different genomic regions.

Differentially Methylated Region Analysis

DMRs for each sequence context (CG, CHG, and CHH) between two samples were identified according to the following stringent criteria: 1) more than five methylated cytosines in at least one sample; 2) more than 10 read's coverage for each cytosine, and more than four reads for each methylated cytosine; 3) region length is between 40 bp and 10 kb; 4) the distance between adjacent methylated sites <200 bp; 5) the fold change of the average methylation level >2; and 6) Pearson's chi-square test (χ^2) value $p \le 0.05$. The putative DMRs overlapping at adjacent 2 kb (upstream or downstream) or body regions of genes or transposable elements (TEs) were sorted out for further study.

Enrichment Analysis of Functional Differently Methylated Region-Related Genes

Significant enrichment analysis was based on Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (http://www.kegg.jp/), and hypergeometric test was applied to find the pathway of significant enrichment in the DMR-related genes compared with the whole genome background. After multiple examination and correction, the pathway of Q-value 0.05 was defined as the pathway of significant enrichment in the differential expression gene. Pathway

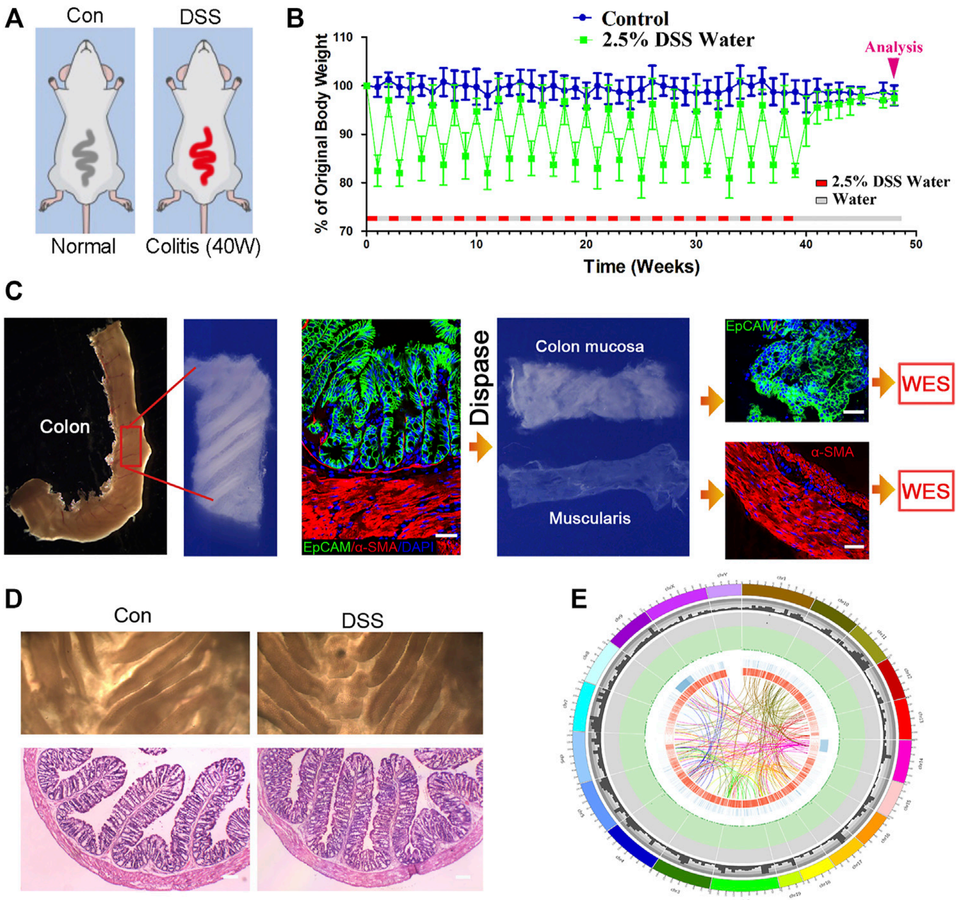


FIGURE 1 | Establishing chronic inflammation model for genomic sequencing assay. **(A, B)** Schematic for dextran sulfate sodium (DSS)-induced colitis, and the procedure of DSS treatment to induce chronic colitis. *n* = 12 for DSS-treated mice and *n* = 10 for control mice. **(C)** The procedure of colon mucosa and muscularis isolation, and the isolated mucosa and muscularis were subjected to whole-exome sequencing (WES). Scale bar: 50 μm. **(D)** Colons were separated in both DSS-treated mice and the age-matched control mice; no obvious morphological difference was found the two groups with dissecting microscope, neither in histopathology. *n* = 6 for both control and DSS-treated mice, and representative results are shown. Scale bar: 50 μm. **(E)** Circos imaging of the overview result of WES, showing (from outside to inside) the length of the genome, the number of genes within 10 M, the frequency of single-nucleotide polymorphisms (SNPs) within 1M (red squares over 0.0015 and green triangles below 0.0005), INS (insertion), INV (inversion), ITX (intrachromosomal translocation), and CTX (interchromosomal translocation).

(p-value ≤0.05, q-value <0.05) was used to analyze whether specific DMRs affect gene's enrichment.

RESULT

Establishing Chronic Inflammation Model and Strategies for Genomic Sequencing Assay

DSS-induced colitis is a wildly used model for the colitis study, in which the chronic DSS colitis could last over 2 months (Wirtz et al., 2007). While a much longer period of chronic colitis that could last over 10 months was desired in this study, to achieve this kind of chronic colitis, we adjusted DSS administration procedure, as follows: distilled water with 2.5% DSS replaced the distilled water for 7 days, during which colitis was induced and the mouse body weight decreased about 20%; and at the eighth day, DSS water was replaced by distilled water for another 7 days, the mouse body weight was restored, and the distilled water was replaced by 2.5% DSS water again. This kind of procedure (7 days of water with 2.5% DSS/7 days of distilled water) was repeatedly performed for 20 times, the mouse body weight was monitored weekly, and 48 weeks later, the mice were sacrificed for colon mucosa and muscularis mucosae isolation (**Figures 1A,B**). Colons were separated in both DSS-treated mice and the age-matched control mice. To separate the colon mucosa tissue from the muscularis mucosae tissue, we digested the colon tissue in 0.4% Dispase II at 37°C for 1.5 h and separated the mucosa and muscularis mucosae layer via dissecting forceps. Then tissue section and immunofluorescence staining were performed on the isolated mucosa and muscularis mucosae layer to validate that the mucosa and muscularis mucosae layer are isolated completely and clearly (**Figure 1C**). No obvious morphological difference was found between the two groups assayed by dissecting microscope, neither

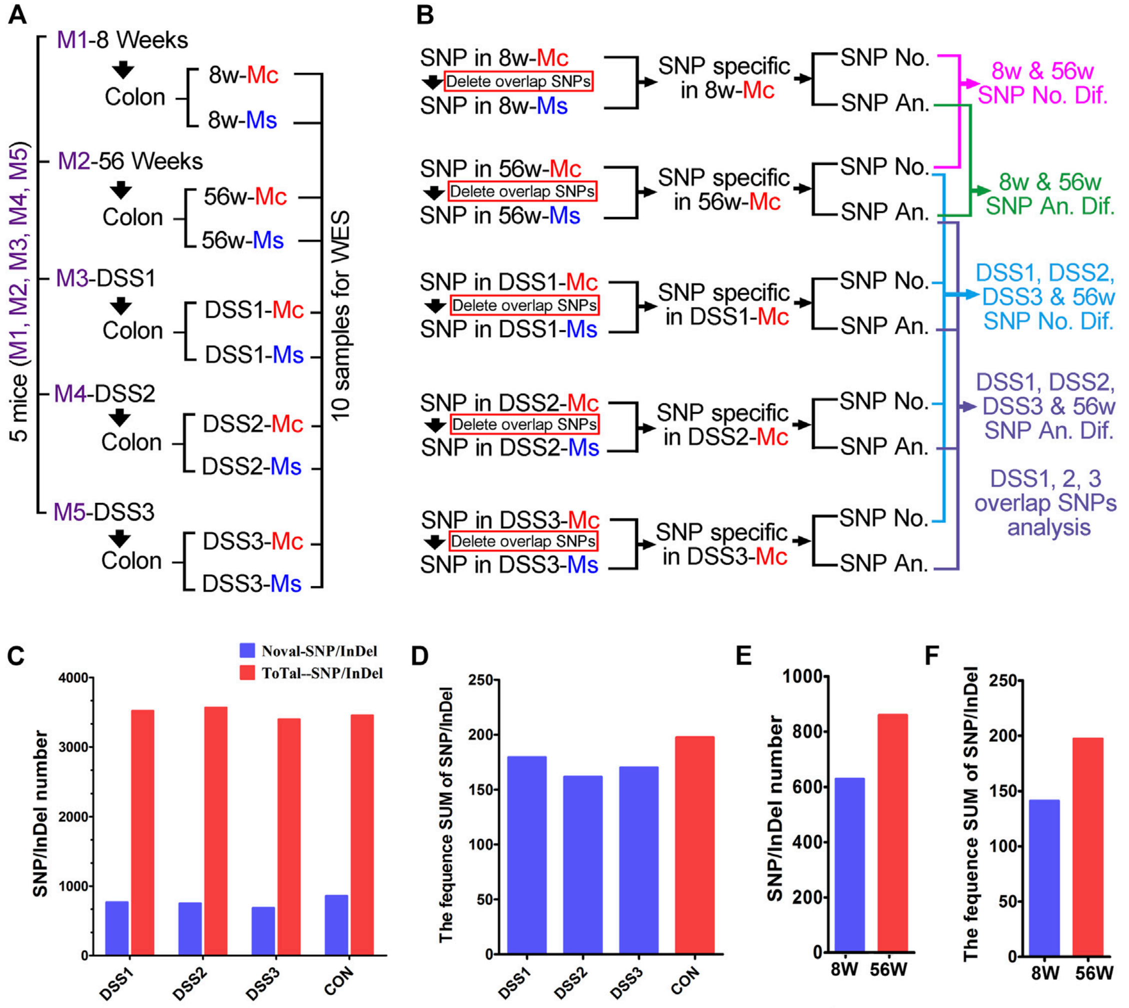
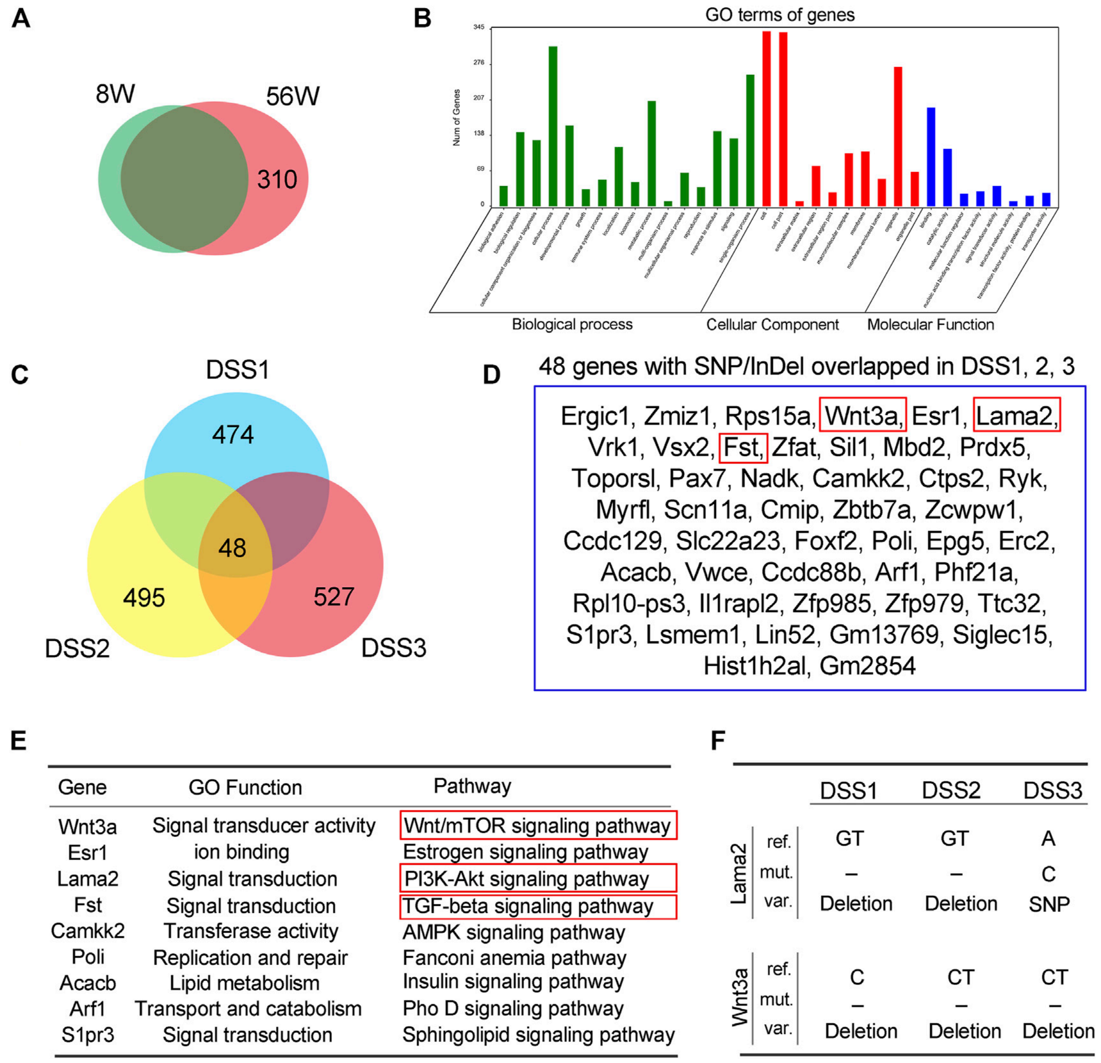


FIGURE 2 | Comparative and quantitative assays of single-nucleotide polymorphism/insertion or deletion (SNP/InDel) in the colitis tissue. **(A)** Ten samples from five mice were used for whole-exome sequencing (WES), mucosa (Mc) and muscularis (Ms) isolation and sequenced individually. **(B)** Comparative assay strategy for the WES result; SNP/InDel from the Mc and Ms of same mice was compared to wipe off the germline mutation. Both statistical and annotative assays were performed on the SNP/InDel specific in Mc from different mice. **(C)** The total SNP/InDel number in the mucosa of three dextran sulfate sodium (DSS)-induced colitis mice is 3,521/ 3,570/3,400; for the no colitis mouse, the SNP/InDel number is 3,456. The novel somatic SNP/InDel number in the three DSS-induced colitis mice is 769/753/687; in age-matched (56W) no-colitis control mouse, the SNP/InDel number is 860. **(D)** SNP/InDel frequency SUM is 179/162/170 in three DSS-induced colitis mice and is 195 in the control mice. **(E)** The SNP/InDel number is obviously elevated in the older mouse (56 weeks, SNP/InDel: 860) compared with the young mouse (8 weeks, SNP/ InDel: 629). **(F)** The SNP/InDel frequency SUM in 56W mouse was obviously increased compared with that in the 8W mouse.

histopathologically with H and E assay (**Figure 1D**). And then the tissues were subjected to DNA isolation and WES. Approximately 20G clean data were generated for each sample, and the average depth is over 200 (>200X). And then the generated reads were compared with reference genome (mm9 mouse genome, related to Agilent mouse exon kit), to evaluate the genomic variations in each individual. The genomic variations of individuals include SNPs, small InDels, and larger-scale variations; and CNV, which generally refers to large-scale (>1 kb) chromosomal copy number changes, e.g., amplifications or deletions, were compared with a reference genome. A overall analysis graph is presented in **Figure 1E**, which includes the SNP frequency in 1 M, and also the INS (insertion), INV (inversion), ITX (intrachromosomal translocation), and CTX (interchromosomal translocation).

Quantitative Assay of Single-Nucleotide Polymorphism/Insertion or Deletion in the Colitis Tissue

Ten samples from five mice were prepared for WES in this study; the five mice (close breeding and same generation) included an 8-week-old mouse (young mouse), a 56W mouse (old mouse, also served as the age-matched mouse for colitis mice), and three colitis mice (56W). The mucosa (Mc) and muscularis mucosae



48 genes with SNP/InDel overlapped in DSS1, 2, 3

Ergic1, Zmiz1, Rps15a, Wnt3a, Esr1, Lama2, Vrk1, Vsx2, Fst, Zfat, Sil1, Mbd2, Prdx5, Toporsl, Pax7, Nadk, Camkk2, Ctps2, Ryk, Myrfl, Scn11a, Cmip, Zbtb7a, Zcwpw1, Ccdc129, Slc22a23, Foxf2, Poli, Epg5, Erc2, Acacb, Vwce, Ccdc88b, Arf1, Phf21a, Rpl10-ps3, Il1rapl2, Zfp985, Zfp979, Ttc32, S1pr3, Lsmem1, Lin52, Gm13769, Siglec15, Hist1h2al, Gm2854

Gene	GO Function	Pathway
Wnt3a	Signal transducer activity	Wnt/mTOR signaling pathway
Esr1	ion binding	Estrogen signaling pathway
Lama2	Signal transduction	PI3K-Akt signaling pathway
Fst	Signal transduction	TGF-beta signaling pathway
Camkk2	Transferase activity	AMPK signaling pathway
Poli	Replication and repair	Fanconi anemia pathway
Acacb	Lipid metabolism	Insulin signaling pathway
Arf1	Transport and catabolism	Pho D signaling pathway
S1pr3	Signal transduction	Sphingolipid signaling pathway

		DSS1	DSS2	DSS3
Lama2	ref.	GT	GT	A
	mut.	–	–	C
	var.	Deletion	Deletion	SNP
Wnt3a	ref.	C	CT	CT
	mut.	–	–	–
	var.	Deletion	Deletion	Deletion

FIGURE 3 | Annotation of single-nucleotide polymorphism/insertion or deletion (SNP/InDel) accumulated in the old and dextran sulfate sodium (DSS)-treated mice. **(A)** There were 310 SNP/InDel unique in 56W mice mucosa. **(B)** Gene Ontology (GO) enrichment showed that these genes were classified into the cellular function of biological process, cellular component, and molecular function. **(C)** There were 769/753/687 somatic (novel) SNP/InDel in the three colitis mice, and this SNP/InDel was annotated into 474/495/527 genes; among these genes from different mice, 48 genes were overlapped. **(D, E)** The 48 genes that were overlapped in three colitis mice. Among these 48 genes, Wnt3a, Lama2, and Fst exert signal transduction function in the Wnt/mTOR, PI3K–AKT, and TGF-beta signaling pathways, respectively. **(F)** Lama2 mutation among the three DSS-treated mice mucosa, two of them were GT deletion mutation, and the other was SNP, while all the Wnt3a mutations in three DSS-treated mucosa were deletion mutation.

(Ms) tissue were separated from the colon of the abovementioned five mice (**Figure 2A**). And concerning the germline mutation contained in the total SNP/InDel, the exome of muscularis mucosae was also sequenced to exclude the germline mutation and then to highlight the somatic mutation generated in the mucosa (the mutations existing in both the mucosa and muscularis were taken as the germline mutation). The SNP/InDel in the mucosa excluded the SNP/InDel muscularis of the same mice and was defined as the SNP/InDel specific in the mucosa (**Figure 2B**). And then the number and annotation of SNP/InDel specific in different mucosae were comparatively assessed (**Figure 2B**). The total SNP/InDel number (compared with the reference sequence) in the mucosa of three colitis mice is 3,521, 3,570, and 3,400, respectively, while in the age-matched no-colitis mouse, the SNP/InDel number is 3,456, indicating that there was no significant quantitative difference in total SNP/InDel mutation between the colitis and control tissues (**Figure 2C**). The somatic (novel) SNP/InDel number in the three DSS-induced colitis mice is 769/753/687, while in the age-matched (56W) control mouse, the SNP/InDel number is 860 (**Figure 2C**). To show an SNP/

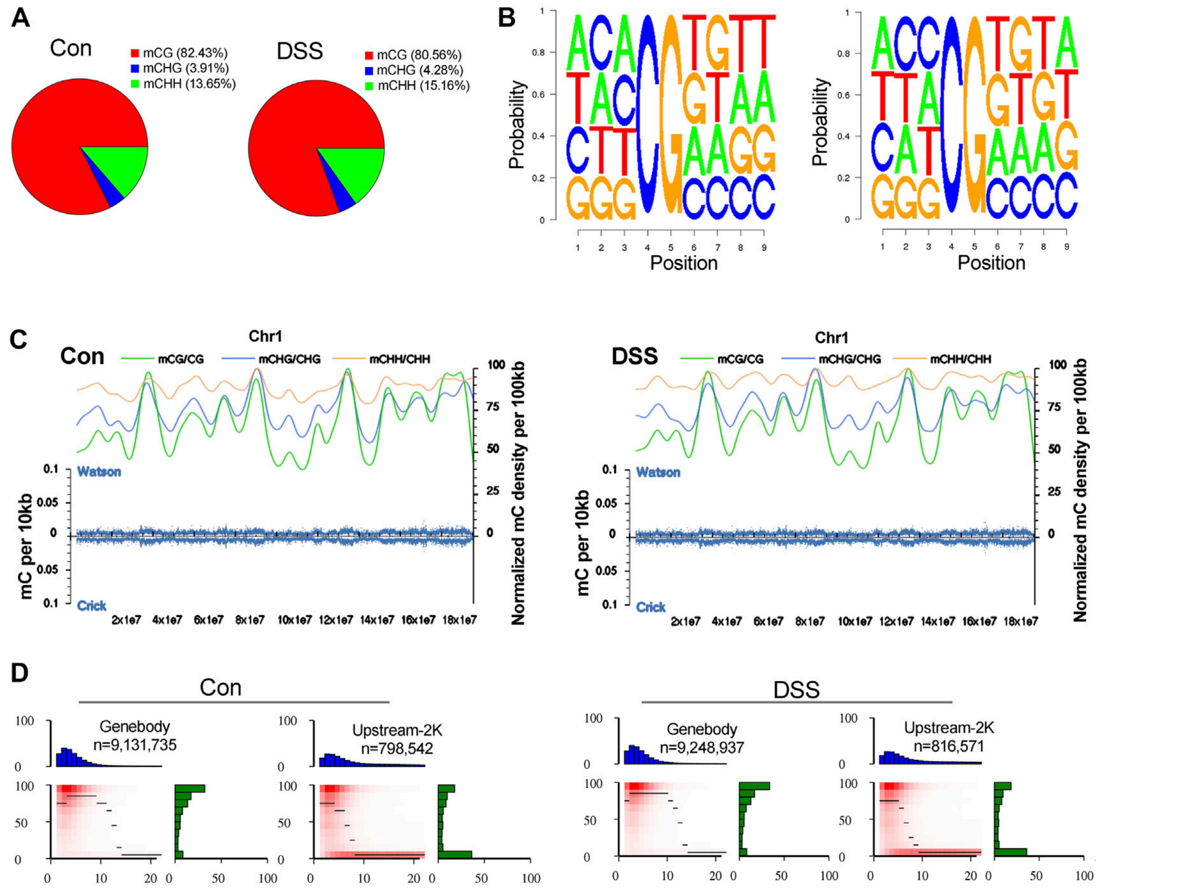


FIGURE 4 | The impact of chronic inflammation upon the DNA methylation profile in colon mucosa. **(A)** The ratio of different types of mC (mCG, mCHG, and mCHH) in dextran sulfate sodium (DSS)-treated and control mice were assessed, in which mCG was 82.43 and 80.56% in control and DSS-treated mice, respectively. **(B)** The sequence characterization of the nucleotides near methylation C of CG, by counting the 9-bp base near CG bases and comparing the base of CG and mCG in the whole genome. **(C)** mC density in control and DSS-treated mice chromosome 1. **(D)** The characteristics of methylation patterns in different gene regions are represented by heat map.

InDel profile that could take the SNP/InDel frequency into account, we calculated the summation (SUM) of all the somatic SNP/InDel frequencies in each individual, and we compared the SNP/InDel frequency SUM between different individuals. Similar to the SNP/InDel number, the SNP/InDel frequency SUM in three DSS-induced colitis mice was not obviously different from that of the control mouse, as SNP/InDel frequency SUM is 179/162/170 in three DSS-induced colitis mice and 195 in the control mice (**Figure 2D**). These data suggest that chronic inflammation could not significantly increase the accumulation of SNP/InDel number as well as its frequency. Similarly, the somatic CNV number in the mucosa of the colitis mice was also not significantly different from that of the control mice (**Figure 2E**).

Even though somatic SNP/InDel/CNV number was not increased by chronic inflammation, we found that the SNP/InDel number was obviously elevated in the older mouse (56W, SNP/InDel: 860) compared with the young mouse (8 weeks old, SNP/InDel: 629) (**Figure 2F**). Thus, this result suggests that aging makes significant contribution to the accumulation of somatic mutation in the individuals.

Annotation of Single-Nucleotide Polymorphism/Insertion or Deletion Accumulated in the Old and Dextran Sulfate Sodium-Treated Mice

To get a further insight on the functional profile of SNP/InDel that uniquely accumulated in old (56W) or DSS-treated mouse colon mucosa, further assay was performed. Comparative assay of the SNP/InDel showed that there were 310 SNP/InDel unique in 56W mouse mucosa; Gene Ontology (GO) enrichment showed that these genes were classified into the cellular function of biological process, cellular component, and molecular function (**Figures 3A,B**). There were 769/753/687 somatic (novel) SNP/InDel in the three colitis mice, and this SNP/InDel was annotated into 474/495/527 genes; among these genes from different mice, 48 genes were overlapped (**Figure 3C**), listed in **Figure 3D**.

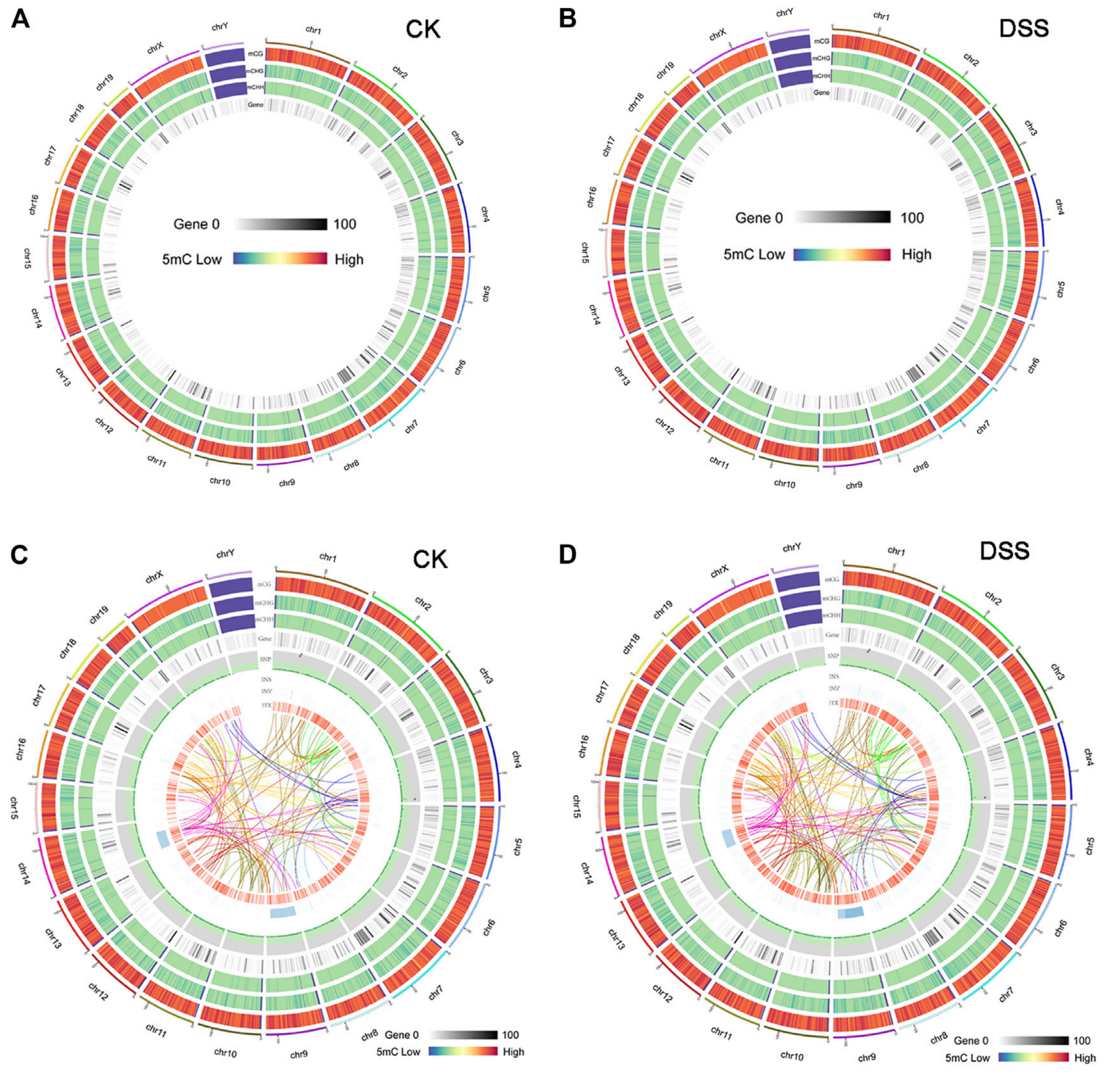


FIGURE 5 | The overview of DNA methylation in control and dextran sulfate sodium (DSS)-treated mice and the conjoint analysis of DNA methylation and gene mutation. **(A, B)** The genomic distribution of mCG, mCHG, and mCHH in two different groups of mice was presented in a Circos graphs. **(C, D)** The conjoint analysis of the DNA methylation and whole-exome sequencing (WES) result.

Various signaling pathways regulate cellular proliferation, differentiation, and immortalization of colorectal cancer, especially Wnt/β-catenin, PI3K/AKT/mTOR and TGF-beta/Smad signaling (Pandurangan et al., 2018). Among these 48 genes, Wnt3a, Lama2, and Fst exert signal transduction function in the Wnt/mTOR, PI3K–AKT, and TGF-beta signaling pathways, respectively (**Figure 3E**). Lama2 is a tumor suppressor by changes in its expression and methylation patterns and can modulate PTEN to exert effects on PI3K/AKT signaling (Wang et al., 2019). For Lama2 mutation among the three DSS-treated mouse mucosae, two of them were GT deletion mutation, and the other was SNP (A–C). Wnt3a is a Wnt protein that activates the canonical Wnt pathway and promotes colon cancer progression (Clevers, 2006; Qi et al., 2014), while all the Wnt3a mutations in three DSS-treated mucosae were deletion mutation (one is C deletion and other two is CT deletion) (**Figure 3F**).

The Impact of Chronic Inflammation Upon the DNA Methylation Profile in Colon Mucosa

To explore the epigenetic profile in the colon mucosa with chronic inflammation, bisulfite sequencing was performed. In the genomic DNA, C bases can be classified into three groups according to their sequence features as CG, CHG, and CHH. In methylated C, the proportions of these three sequence types vary among species. Thus, the number of each type of mC (mCG, mCHG, and mCHH), and their share of total mC sites, could reflect the characteristics of the genome-wide methylation profile.

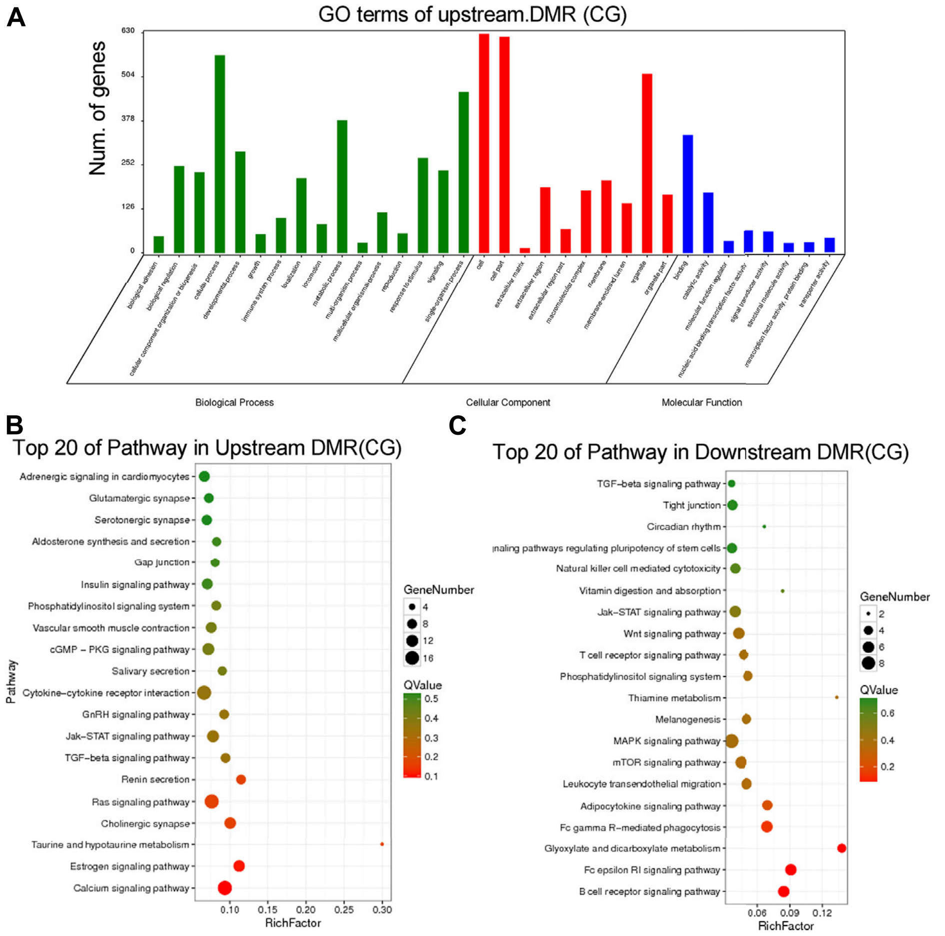


FIGURE 6 | Annotation of the differently methylated regions (DMRs) in chronic inflammation mice genomic. **(A)** Gene Ontology (GO) enrichment result showed the DMRs enriched in biological process, cellular component, and molecular function. **(B)** The top 20 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway of upstream DMR (CG)-related genes. **(C)** The top 20 KEGG pathway of downstream DMR (CG)-related genes.

The ratio of different types of mC in DSS-treated and control mice was assessed, in which mCG was 82.43 and 80.56% in control and DSS-treated mice, respectively (**Figure 4A**). The sequence features of the bases near the methylation sites within the whole genome are instructive to reflect the sequence bias of methylation. Counting the 9-bp base near CG bases and comparing the base of CG and mCG in the whole genome enabled to obtain the sequence bias characteristic of methylation. The sequence characterization of the nucleotides near methylation C of CG is presented in **Figure 4B**. In addition, by analyzing the density distribution of mC at the chromosome level, we obtained the centralization bias of the methylation at the macro level. The mC density in each chromosome was assessed individually, and the result of mC density in control and DSS-treated mouse chromosome 1 is presented (**Figure 4C**). Different genomic regions have different biological functions; to get further insight on the mC distribution feature, the distribution of methylation status is represented by heat map, which visualized the characteristics of methylation patterns in different gene regions (**Figure 4D**). The genomic distribution of mCG, mCHG, and mCHH in two different groups of mice is presented in circle graphs (**Figures 5A,B**). In addition, the conjoint analysis of the DNA methylation and WES result was performed; the profile of both DNA methylation and gene mutation in control and DSS-treated mice is presented (**Figures 5C,D**).

Annotation of the Differently Methylated Regions in Chronic Inflammation Mouse Genomic

To further interpret the DNA methylation result, analysis of the DMR was performed. The DMRs that meet a certain condition in the same position of both control and DSS-treated mice were searched, and the difference in methylation level in this region is greater than 2 as DMRs in this study. Finally, according to the DMR and gene region (including 2 kb upstream of the gene and 2 kb downstream of the gene) on each chromosome, the genes related to differential methylation were determined, and GO and pathway enrichment analysis of these genes were performed. GO

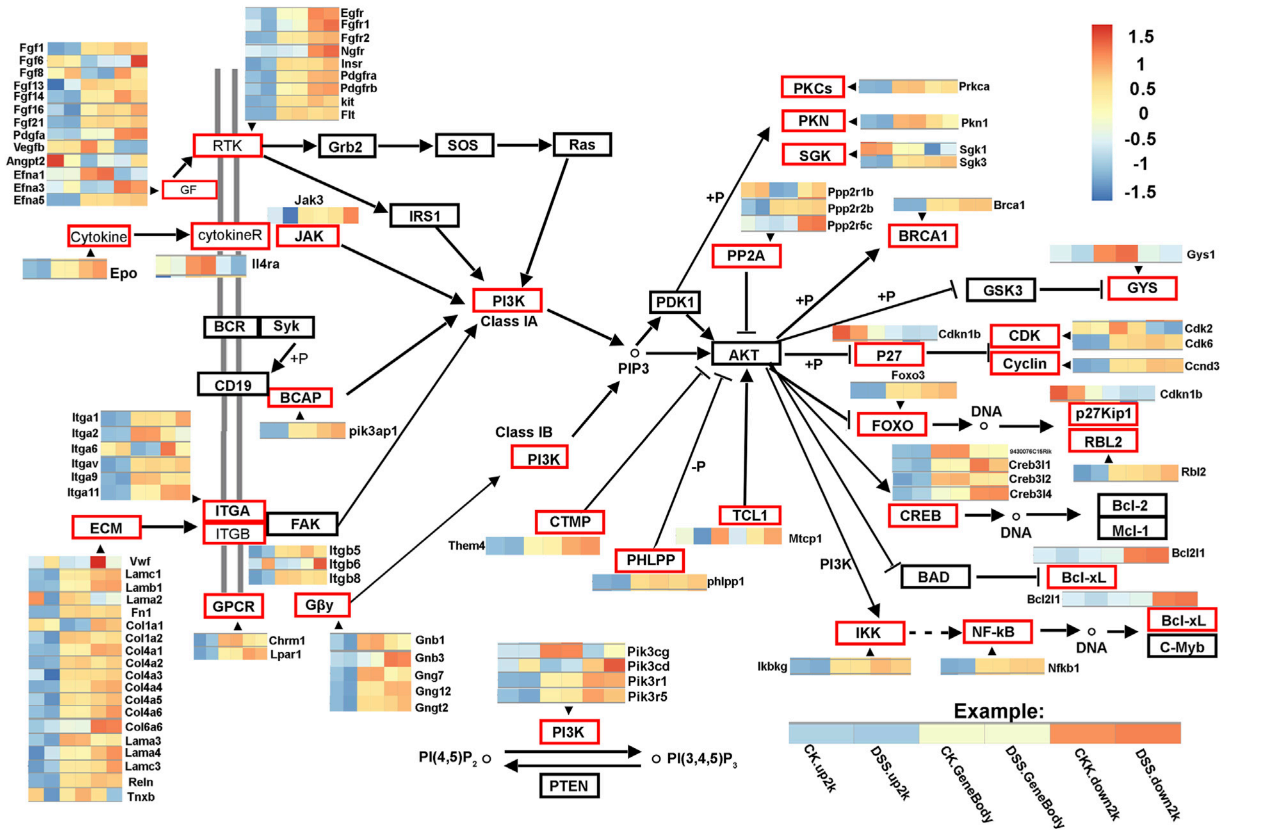


FIGURE 7 | DNA methylation in different genes of PI3K/AKT signaling pathways. The network shows differentially DNA methylated genes of PI3K/AKT signaling pathways. The gene regions (CK.gene body and DSS. gene body), 2 kb upstream of the gene regions (CK.up2k and DSS. up2k), and downstream of the gene regions (CK.down2k and DSS. down2k) in control and dextran sulfate sodium (DSS)-treated mice are analyzed. The differentially DNA methylated genes are marked in red boxes.

enrichment result showed the genes that most significantly enriched in biological process including cellular process, metabolic process, and single-organism process. And the genes in cell, cell part, and organelle were significantly enriched in cellular component. Genes of binding and catalytic activity were significantly enriched in molecular function (**Figure 6A**). Different genes interact with each other to executive certain biological function, pathway assay helps understand gene function, and KEGG is the major public database for pathway assay. KEGG pathway enrichment could determine biochemical and metabolic participation of DMR-related genes. In this study, the KEGG assay showed that upstream DMR (CG)-related genes are enriched to pathways including Jak-STAT signaling pathway, TGF-beta signaling pathway, and Ras signaling pathway (**Figure 6B**). And the downstream DMR (CG)-related genes are enriched to pathways including Wnt signaling pathway, MAPK signaling pathway, and mTOR signaling pathway (**Figure 6C**). The KEGG enrichment of DMR-related genes in different signaling pathways was assessed independently. DNA methylation in different genes of PI3K/AKT signaling pathways and network of gene regulation is shown in **Figure 7**. In conclusion, the data of this study suggest that chronic inflammation showed little influence on genetic stability; no significant mutations were accumulated in chronic tissue, while the chronic inflammation did have a certain impact on the DNA methylation of colon mucosa tissue.

DISCUSSION

The relationship between inflammation and cancers has been studied for over 150 years, and accumulated researches support that chronic inflammatory diseases are related to cancers (Balkwill and Mantovani, 2001; Coussens and Werb, 2002; Philip et al., 2004). As early as 1,863, Virchow indicated that cancers tended to occur at sites of chronic inflammation. Lately, it turned out that acute inflammation contributed to the regression of cancer. However, accumulated epidemiologic studies support that chronic inflammatory diseases are frequently associated with increased risk of cancers. Our understanding of the association between chronic inflammation and cancer is mostly illustrated by inflammatory bowel disease and colon carcinogenesis. Previous report shows that there was an 18-fold increase in the risk of developing colorectal cancer in extensive Crohn's colitis and a 19-fold increase in risk in extensive UC when compared with the general population, matched for age, sex, and years at risk (Gillen et al., 1994). And increased cancer incidence is associated with increased duration of the inflammation. On basis of the toxic effect of inflammation/RONS on DNA, in this study, we hypothesize that chronic inflammation would result in somatic mutation accumulation, which thereby increases the risk of carcinogenesis. While inconsistent with highly increased cancer development risk in inflammation, there was no somatic mutation increase observed during chronic inflammation, thus increasing the somatic mutation dislike to be the mediator of inflammation-induced cancer developing risk.

Previous reports indicate that the number of ε-base lesions and 8oxoG increased in the colons of mice following a single DSS treatment; in addition, consistent with the ability of Aag (alkyladenine DNA glycosylase) to excise both εA and 8oxoG (Bartsch and Nair, 2002; Meira et al., 2008), inflammation-induced εA, εC, and 8oxoG increased more dramatically in the Aag-deficient mice, since Aag could recognize the DNA damage and initiate base excision repair. Moreover, in the mice with deficiency in three DNA repair proteins ($Aag^{-/-}$/$Alkbh2^{-/-}$/$Alkbh3^{-/-}$ triple-knockout), a single cycle of DSS-induced colitis resulted in absolute lethality of these mice (Calvo et al., 2012). These studies indicate the crucial role of the DNA repair proteins in both tumor suppression and tissue homeostasis. Despite the increased levels of toxic and mutagenic εA and 1, N^2-εG in colon mucosa cells following DSS treatment, our data suggest that these DNA lesions do not ultimately contribute to the accumulation of somatic genomic alteration, including the SNP/InDel and CNV; presumably, there is redundant potential of DNA repair proteins, which is enough to compensate the increased DNA damage/repair activity during the inflammation, thus finally not resulting in the obviously increased DNA mutation. Collectively, increasing the accumulation of somatic gene mutations does not seem to be the major mechanism of chronic inflammation in promoting the neoplasia; and we proposed that the DNA repair mechanism is efficient enough to repair the increased DNA damage induced by inflammation, thus eliminating the risk of DNA mutation accumulation.

The methylation CpG islands in the promoter region of tumor suppressor genes can silence gene expression and lead to tumorigenesis (Baylin and Jones, 2011). In this process, the activation status of several cancer-related pathways are changed, including Ras, Wnt/β-catenin, PI3K/AKT, and MAPK signaling pathways (Jones and Baylin, 2007; Ying and Tao, 2009; Fattahi et al., 2020). In our study, the differentially methylated genes and gene region (including 2 kb upstream of the gene and 2 kb downstream of the gene) were determined, and GO and KEGG enrichment analysis of these genes were performed. We found that upstream DMR (CG) and downstream DMR (CG)-related genes were enriched to different cancer-related pathways, which pointed their different functions on carcinogenesis.

Tumor formation is a complex process involving many genes and procedures. The process of chronic inflammatory models may be a transition between inflammation and cancer. Our research presents how chronic inflammation transitions to a tumor, what happens on the exon, what goes on in the epigenetic methylation. And our results provide a train of thoughts. The process of studying methylation can predict in advance which proteins will be disorganized. We may be able to reverse the transformation cells by correcting DNA methylation abnormalities. In conclusion, to study methylation and exon sites SNPs in the human disease model is significant for the diagnosis and treatment of chronic inflammation and tumors.

DATA AVAILABILITY STATEMENT

The whole-exome sequencing data presented in this study have been deposited in the European Variation Archive repository (https://www.ebi.ac.uk/eva), accession number: PRJEB48005; The methylated sequencing data in this paper have been deposited in the OMIX, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (https://ngdc.cncb.ac.cn/omix), accession number: OMIX694.

ETHICS STATEMENT

The animal study was reviewed and approved by Animal Ethics Committee of Shenzhen Center for Disease Control and Prevention (CDC).

AUTHOR CONTRIBUTIONS

XC and XW contributed to devise the project, the main conceptual ideas and proof outline. JHe, JHa and JL worked out almost all of the technical details, and performed the suggested experiment. RY and JW analysed the sequencing data.

FUNDING

This work was supported by grants from Shenzhen Science and Technology Innovation Committee (No. JCYJ20200109142444449), Medical Research Fund of Guangdong Province (No. A2020322), Special fund of Foshan Summit plan (Nos. 2020B019, 2020A015, 2019A006, 2019C002, 2019D008), Special Fund for Science and Technology Innovation Strategy of Guangdong Province (No. 2020A1515011402), Regional Joint Fund of Natural Science Foundation of Guangdong Province (No. 2020A1515110432).

REFERENCES

Balkwill, F., and Mantovani, A. (2001). Inflammation and Cancer: Back to Virchow? *The Lancet* 357 (9255), 539–545. doi:10.1016/s0140-6736(00)04046-0

Bartsch, H., and Nair, J. (2002). Potential Role of Lipid Peroxidation Derived DNA Damage in Human colon Carcinogenesis: Studies on Exocyclic Base Adducts as Stable Oxidative Stress Markers. *Cancer Detect. Prev.* 26 (4), 308–312. doi:10.1016/s0361-090x(02)00093-4

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

Calvo, J. A., Meira, L. B., Lee, C.-Y. I., Moroski-Erkul, C. A., Abolhassani, N., Taghizadeh, K., et al. (2012). DNA Repair Is Indispensable for Survival after Acute Inflammation. *J. Clin. Invest.* 122 (7), 2680–2689. doi:10.1172/jci63338

Clevers, H. (2006). Wnt/β-Catenin Signaling in Development and Disease. *Cell* 127, 469–480. doi:10.1016/j.cell.2006.10.018

Coussens, L. M., and Werb, Z. (2002). Inflammation and Cancer. *Nature* 420 (6917), 860–867. doi:10.1038/nature01322

de Ligt, J., Willemsen, M. H., van Bon, B. W. M., Kleefstra, T., Yntema, H. G., Kroes, T., et al. (2012). Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N. Engl. J. Med.* 367 (20), 1921–1929. doi:10.1056/nejmoa1206524

Dhir, M., Montgomery, E. A., Glöckner, S. C., Schuebel, K. E., Hooker, C. M., Herman, J. G., et al. (2008). Epigenetic Regulation of WNT Signaling Pathway Genes in Inflammatory Bowel Disease (IBD) Associated Neoplasia. *J. Gastrointest. Surg.* 12 (10), 1745–1753. doi:10.1007/s11605-008-0633-5

Ehrlich, M. (2009). DNA Hypomethylation in Cancer Cells. *Epigenomics* 1 (2), 239–259. doi:10.2217/epi.09.33

Fattahi, S., Amjadi-Moheb, F., Tabaripour, R., Ashrafi, G. H., and Akhavan-Niaki, H. (2020). PI3K/AKT/mTOR Signaling in Gastric Cancer: Epigenetics and beyond. *Life Sci.* 262, 118513. doi:10.1016/j.lfs.2020.118513

Gillen, C. D., Walmsley, R. S., Prior, P., Andrews, H. A., and Allan, R. N. (1994). Ulcerative Colitis and Crohn's Disease: a Comparison of the Colorectal Cancer Risk in Extensive Colitis. *Gut* 35 (11), 1590–1592. doi:10.1136/gut.35.11.1590

Hartnett, L., and Egan, L. J. (2012). Inflammation, DNA Methylation and Colitis-Associated Cancer. *Carcinogenesis* 33 (4), 723–731. doi:10.1093/carcin/bgs006

Helicobacter and Cancer Collaborative GroupCancer Collaborative G (2001). Gastric Cancer and *Helicobacter pylori*: a Combined Analysis of 12 Case Control Studies Nested within Prospective Cohorts. *Gut* 49 (3), 347–353. doi:10.1136/gut.49.3.347

Hoeijmakers, J. H. J. (2009). DNA Damage, Aging, and Cancer. *N. Engl. J. Med.* 361 (15), 1475–1485. doi:10.1056/nejmra0804615

Jones, P. A., and Baylin, S. B. (2007). The Epigenomics of Cancer. *Cell* 128, 683–692. doi:10.1016/j.cell.2007.01.029

Kaname, T., Yanagi, K., and Naritomi, K. (2014). A Commentary on the Promise of Whole-Exome Sequencing in Medical Genetics. *J. Hum. Genet.* 59 (3), 117–118. doi:10.1038/jhg.2014.7

Lonkar, P., and Dedon, P. C. (2011). Reactive Species and DNA Damage in Chronic Inflammation: Reconciling Chemical Mechanisms and Biological Fates. *Int. J. Cancer* 128 (9), 1999–2009. doi:10.1002/ijc.25815

Maisonneuve, P., and Lowenfels, A. B. (2002). Chronic Pancreatitis and Pancreatic Cancer. *Dig. Dis.* 20 (1), 32–37. doi:10.1159/000063165

Maynard, S., Schurman, S. H., Harboe, C., de Souza-Pinto, N. C., and Bohr, V. A. (2009). Base Excision Repair of Oxidative DNA Damage and Association with Cancer and Aging. *Carcinogenesis* 30 (1), 2–10. doi:10.1093/carcin/bgn250

McLarnon, A. (2011). Risk of Colorectal Cancer in African Americans with Ulcerative Colitis. *Nat. Rev. Gastroenterol. Hepatol.* 8 (11), 598. doi:10.1038/nrgastro.2011.168

Meira, L. B., Bugni, J. M., Green, S. L., Lee, C. W., Pang, B., Borenshtein, D., et al. (2008). DNA Damage Induced by Chronic Inflammation Contributes to colon Carcinogenesis in Mice. *J. Clin. Invest.* 118 (7), 2516–2525. doi:10.1172/JCI35073

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in Understanding Cancer Genomes through Second-Generation Sequencing. *Nat. Rev. Genet.* 11 (10), 685–696. doi:10.1038/nrg2841

Ndlovu, M. N., Denis, H., and Fuks, F. (2011). Exposing the DNA Methylome Iceberg. *Trends Biochem. Sci.* 36 (7), 381–387. doi:10.1016/j.tibs.2011.03.002

Pandurangan, A. K., Divya, T., Kumar, K., Dineshbabu, V., Velavan, B., and Sudhandiran, G. (2018). Colorectal Carcinogenesis: Insights into the Cell Death and Signal Transduction Pathways: a Review. *World J. Gastrointest. Oncol.* 10, 244–259. doi:10.4251/wjgo.v10.i9.244

Philip, M., Rowley, D. A., and Schreiber, H. (2004). Inflammation as a Tumor Promoter in Cancer Induction. *Semin. Cancer Biol.* 14 (6), 433–439. doi:10.1016/j.semcancer.2004.06.006

Qadri, Q., Rasool, R., Gulzar, G. M., Naqash, S., and Shah, Z. A. (2014). H. pylori Infection, Inflammation and Gastric Cancer. *J. Gastrointest. Canc* 45 (2), 126–132. doi:10.1007/s12029-014-9583-1

Qi, L., Sun, B., Liu, Z., Cheng, R., Li, Y., and Zhao, X. (2014). Wnt3a Expression Is Associated with Epithelial-Mesenchymal Transition and Promotes colon Cancer Progression. *J. Exp. Clin. Cancer Res.* 33, 107. doi:10.1186/s13046-014-0107-4

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The Promise of Whole-Exome Sequencing in Medical Genetics. *J. Hum. Genet.* 59 (1), 5–15. doi:10.1038/jhg.2013.114

Raimondi, S., Lowenfels, A. B., Morselli-Labate, A. M., Maisonneuve, P., and Pezzilli, R. (2010). Pancreatic Cancer in Chronic Pancreatitis; Aetiology, Incidence, and Early Detection. *Best Pract. Res. Clin. Gastroenterol.* 24 (3), 349–358. doi:10.1016/j.bpg.2010.02.007

Risques, R. A., Lai, L. A., Himmetoglu, C., Ebaee, A., Li, L., Feng, Z., et al. (2011). Ulcerative Colitis-Associated Colorectal Cancer Arises in a Field of Short Telomeres, Senescence, and Inflammation. *Cancer Res.* 71 (5), 1669–1679. doi:10.1158/0008-5472.can-10-1966

Senol, K., Ozkan, M. B., Vural, S., and Tez, M. (2014). The Role of Inflammation in Gastric Cancer. *Adv. Exp. Med. Biol.* 816, 235–257. doi:10.1007/978-3-0348-0837-8_10

Tominaga, K., Fujii, S., Mukawa, K., Fujita, M., Ichikawa, K., Tomita, S., et al. (2005). Prediction of Colorectal Neoplasia by Quantitative Methylation Analysis of Estrogen Receptor Gene in Nonneoplastic Epithelium from Patients with Ulcerative Colitis. *Clin. Cancer Res.* 11, 8880–8885. doi:10.1158/1078-0432.ccr-05-1309

Wang, F. Y., Arisawa, T., Tahara, T., Takahama, K., Watanabe, M., Hirata, I., et al. (2008). Aberrant DNA Methylation in Ulcerative Colitis without Neoplasia. *Hepatogastroenterology* 55, 62–65.

Wang, R. Q., Lan, Y. L., Lou, J. C., Lyu, Y. Z., Hao, Y. C., Su, Q. F., et al. (2019). Expression and Methylation Status of LAMA2 Are Associated with the Invasiveness of Nonfunctioning PitNET. *Ther. Adv. Endocrinol. Metab.* 10, 1–11. doi:10.1177/2042018818821296

Wirtz, S., Neufert, C., Weigmann, B., and Neurath, M. F. (2007). Chemically Induced Mouse Models of Intestinal Inflammation. *Nat. Protoc.* 2 (3), 541–546. doi:10.1038/nprot.2007.41

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., et al. (2013). Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N. Engl. J. Med.* 369 (16), 1502–1511. doi:10.1056/nejmoa1306555

Ying, Y., and Tao, Q. (2009). Epigenetic Disruption of the WNT/ß-catenin Signaling Pathway in Human Cancers. *Epigenetics* 4, 307–312. doi:10.4161/epi.4.5.9371

Yu, D.-H., Waterland, R. A., Zhang, P., Schady, D., Chen, M.-H., Guan, Y., et al. (2014). Targeted p16Ink4a Epimutation Causes Tumorigenesis and Reduces Survival in Mice. *J. Clin. Invest.* 124 (9), 3708–3712. doi:10.1172/jci76507

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

The reviewer JC declared a shared affiliation with the authors JH, JH, JL, XW to the handling editor at time of review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.