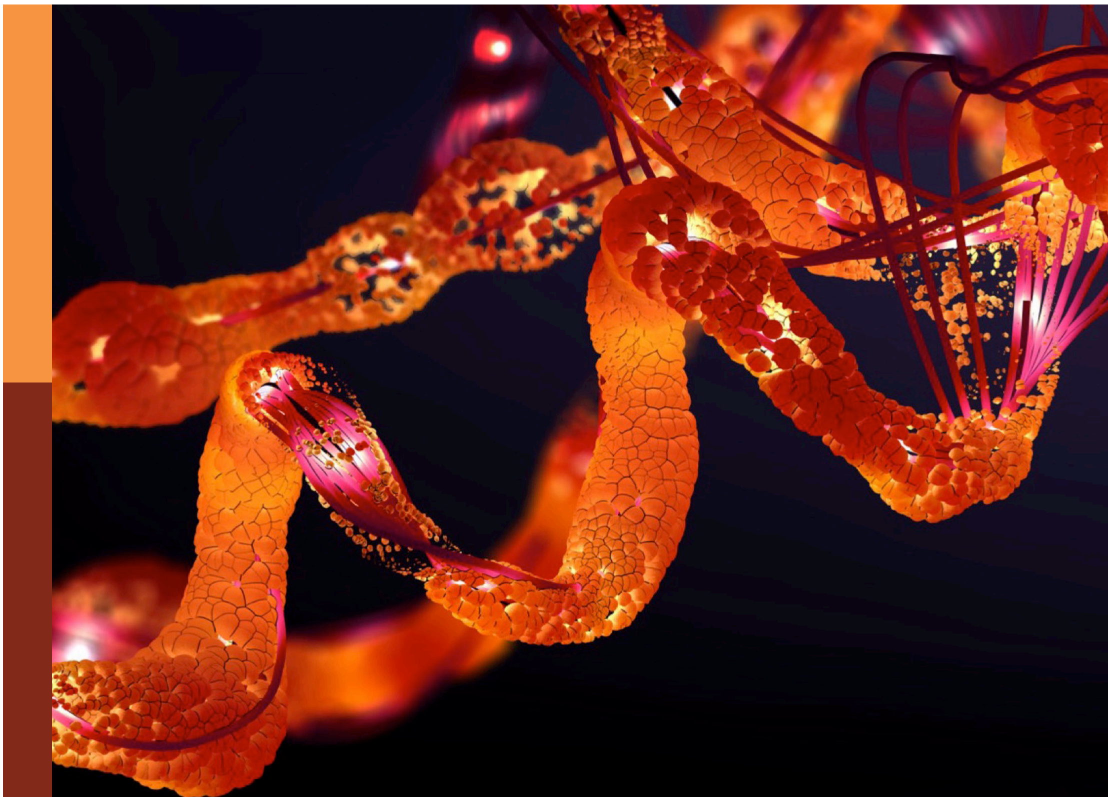
frontiers | Research Topics

Leveraging machine learning for omics-driven biomarker discovery

Edited by
Sheng Li, Charles Hsu, Tianyi Zhao and Liangcan He

Published in
Frontiers in Molecular Biosciences



ISSN 1664-8714
ISBN 978-2-83251-000-1
DOI 10.3389/978-2-83251-000-1

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

Leveraging machine learning for omics-driven biomarker discovery

Topic editors

Sheng Li — Wuhan University, China
Charles Hsu — Qatar University, Qatar
Tianyi Zhao — Harbin Institute of Technology, China
Liangcan He — Harbin Institute of Technology, China

Citation

Li, S., Hsu, C., Zhao, T., He, L., eds. (2023). *Leveraging machine learning for omics-driven biomarker discovery*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83251-000-1

Table of contents

05 Editorial: Leveraging machine learning for omics-driven biomarker discovery
Sheng Li, Charles Hsu, Tianyi Zhao and Liangcan He

08 Identification of a Novel Defined Immune-Autophagy-Related Gene Signature Associated With Clinical and Prognostic Features of Kidney Renal Clear Cell Carcinoma
Guangyuan Zhang, Lei Zhang, Si Sun and Ming Chen

26 Pan-Cancer Integrated Analysis of HSF2 Expression, Prognostic Value and Potential Implications for Cancer Immunity
Fei Chen, Yumei Fan, Xiaopeng Liu, Jianhua Zhang, Yanan Shang, Bo Zhang, Bing Liu, Jiajie Hou, Pengxiu Cao and Ke Tan

46 Large-Scale Gastric Cancer Susceptibility Gene Identification Based on Gradient Boosting Decision Tree
Qing Chen, Ji Zhang, Banghe Bao, Fan Zhang and Jie Zhou

53 Metastatic Immune-Related Genes for Affecting Prognosis and Immune Response in Renal Clear Cell Carcinoma
Si Sun, Weipu Mao, Lilin Wan, Kehao Pan, Liting Deng, Lei Zhang, Guangyuan Zhang and Ming Chen

69 Inferring Retinal Degeneration-Related Genes Based on Xgboost
Yujie Xia, Xiaojie Li, Xinlin Chen, Changjin Lu and Xiaoyi Yu

76 CRISPR Detection and Research on Screening Mutant Gene *TTN* of Moyamoya Disease Family Based on Whole Exome Sequencing
Yilei Xiao, Weidong Liu, Jiheng Hao, Qunlong Jiang, Xingbang Wang, Donghu Yu, Liyong Zhang, Zhaogang Dong and Jiyue Wang

88 Identification of a 3-Gene Prognostic Index for Papillary Thyroid Carcinoma
Lin-Kun Zhong, Xing-Yan Deng, Fei Shen, Wen-Song Cai, Jian-Hua Feng, Xiao-Xiong Gan, Shan Jiang, Chi-Zhuai Liu, Ming-Guang Zhang, Jiang-Wei Deng, Bing-Xing Zheng, Xiao-Zhang Xie, Li-Qing Ning, Hui Huang, Shan-Shan Chen, Jian-Hang Miao and Bo Xu

97 Network Pharmacology and Experimental Validation to Reveal the Pharmacological Mechanisms of Chongcaoyishen Decoction Against Chronic Kidney Disease
Zhenliang Fan, Jingjing Chen, Qiaorui Yang and Jiabei He

110 The Value of H2BC12 for Predicting Poor Survival Outcomes in Patients With WHO Grade II and III Gliomas
Jie Zhou, Zhaoquan Xing, Yilei Xiao, Mengyou Li, Xin Li, Ding Wang and Zhaogang Dong

123 Identification of Four Novel Prognostic Biomarkers and Construction of Two Nomograms in Adrenocortical Carcinoma: A Multi-Omics Data Study *via* Bioinformatics and Machine Learning Methods
Xiaochun Yi, Yueming Wan, Weiwei Cao, Keliang Peng, Xin Li and Wangchun Liao

140 Identifying Topics and Evolutionary Trends of Literature on Brain Metastases Using Latent Dirichlet Allocation
Jiarong Chen, Matt Williams, Yanming Huang and Shijing Si

149 Machine learning prediction of postoperative unplanned 30-day hospital readmission in older adult
Linji Li, Linna Wang, Li Lu and Tao Zhu

frontiers | Frontiers in Molecular Biosciences

TYPE Editorial
PUBLISHED 09 January 2023
DOI 10.3389/fmolb.2022.1119644

Check for updates

OPEN ACCESS

EDITED BY
William C. Cho,
QEH, Hong Kong SAR, China

REVIEWED BY
Sandip Kuamr Patel,
Buck Institute for Research on Aging,
United States

*CORRESPONDENCE
Sheng Li,
✉ lisheng-znyy@whu.edu.cn
Charles Hsu,
✉ xuchang2016@runbox.com
Tianyi Zhao,
✉ zty2009@hit.edu.cn
Liangcan He,
✉ liangcanhe@hit.edu.cn

SPECIALTY SECTION
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 09 December 2022
ACCEPTED 22 December 2022
PUBLISHED 09 January 2023

CITATION
Li S, Hsu C, Zhao T and He L (2023), Editorial: Leveraging machine learning for omics-driven biomarker discovery. *Front. Mol. Biosci.* 9:1119644. doi: 10.3389/fmolb.2022.1119644

Editorial: Leveraging machine learning for omics-driven biomarker discovery

Sheng Li[1]*, Charles Hsu[2]*, Tianyi Zhao[3]* and Liangcan He[3]*

[1]Zhongnan Hospital, Wuhan University, Wuhan, China, [2]Department of Population Medicine, College of Medicine, Qatar University, Doha, Qatar, [3]Harbin Institute of Technology, Harbin, China

KEYWORDS
biomarker, machine learning, cancer, omics data analysis, computational resource

Editorial on the Research Topic
Leveraging machine learning for omics-driven biomarker discovery

Here, we organized a Research Topic on “Leveraging Machine Learning for Omics-driven Biomarker Discovery.” In total, about 12 outstanding works were presented in this thematic issue, and they have been highlighted as follows.

- Chen et al. comprehensively investigated the expression dysregulation and prognostic significance of HSF2, and the relationship with clinicopathological parameters and immune infiltration across cancers. Their study revealed the varied expression of HSF2 in different types and stages of cancers, which suggests that the effects of HSF2 on oncogenesis may vary across different cancer types. A significant correlation between HSF2 expression and patients’ prognosis was observed. HSF2 expression was strongly related to immune cell infiltration, immune checkpoints, TMB, and MSI. They integrated existing data to explore the potential function of HSF2 in cancers and provides insights for targeting HSF2 to improve the therapeutic efficacy of immunotherapy.
- Zhang et al. found that CANX, BID, NAMPT, and BIRC5 were immune-autophagy-related genes with independent prognostic value, and the risk prognostic model based on them was theyll constructed. Through GSE168845, immune-related genes, autophagy-related genes, and immune-autophagy-related differentially expressed genes (IAR-DEGs) were identified. Then, the lasso Cox regression model was established to evaluate the correlation of IAR-DEGs with the immune score, immune checkpoints, methylation, and one-class logistic regression (OCLR) score. Further analysis showed that CANX, BID, NAMPT, and BIRC5 were potential targets and effective prognostic biomarkers for immunotherapy combined with autophagy in kidney renal clear cell carcinoma.
- Sun et al. analyzed the correlation of hub mIR-DEGs with clinicopathological factors, immune invasion, and immune checkpoints, and re-evaluated the expression of hub mIR-DEGs and their effect on the tumor by OCLR scores in KIRC. Co-expressed metastatic immune-related differentially expressed genes (mIR-DEGs) were screened out, and the mIR-DEGs-based prognostic model that had good predictive potential was established. In addition, targeted small-molecule drugs were predicted for mIR-DEGs. This study preliminarily confirmed that FGF17, PRKCG, SSTR1, and SCTR were targeted genes that can be used as potential therapeutic targets and prognostic biomarkers for renal cancer. Preliminary validation found that PRKCG and SSTR1 were consistent with predictions.

- Zhong et al. objectives are to screen for characteristic genes specific to PTC and establish an accurate model for diagnosis and prognostic evaluation of PTC. They screened differentially expressed genes in TCGA database and discovered a three-gene signature (GJB4, RIPPLY3, ADRA1B) that was statistically significant and externally validated. For experimental validation, immunohistochemistry in tissue microarrays showed that thyroid samples' proteins expressed by this three-gene were differentially expressed. The protocol discovered a robust three-gene signature that can distinguish prognosis, which will have daily clinical application.
- Chen et al. proposed a method based on Gradient Boosting Decision Tree (GBDT) to identify the susceptible genes of gastric cancer through a gene interaction network. Based on the known genes related to gastric cancer, they collected more genes that can interact with them and constructed a gene interaction network. Random Walk was used to extract the network association of each gene and they used GBDT to identify the gastric cancer-related genes. To verify the AUC and AUPR of their algorithm, they implemented 10-fold cross-validation. GBDT achieved AUC of .89 and AUPR of .81. This work selected ftheir other methods to compare with GBDT and found GBDT performed best.
- Zhou et al. aimed to satisfy the increasing demand for novel sensitive biomarkers and potential therapeutic targets in the treatment of GII and GIII gliomas. Their study revealed the multi-omics landscape of H2BC12 in gliomas through bioinformatics approaches. They identified the differentially up-regulated expression of H2BC12 in GII and GIII glioma tissue and proved its significant ability in predicting the adverse overall survival of GII and GIII gliomas patients. They verified that H2BC12 was a promising biomarker for the diagnosis and prognosis of patients with WHO grade II and III gliomas In a forward-looking way.
- Xia et al. purposed Xgboost to identify RP-related genes. Xgboost adds a regular term to control the complexity of the model, hence using Xgboost to find out true RD-related genes from complex and massive genes is suitable. The problem of overfitting can be avoided to some extent. To verify the potheyr of Xgboost to identify RD-related genes, they did 10-cross validation and compared it with three traditional methods: Random Forest, Back Propagation network, and Support Vector Machine. The accuracy of Xgboost is 99.13% and AUC is much higher than the other three methods. Therefore, this article can provide technical support for the efficient identification of RD-related genes and help researchers have a deeper understanding of the genetic characteristics of RD.
- Xiao et al. identified familial cohorts showing MMD susceptibility and performed THEYS on five affected individuals to identify susceptibility loci, which identified point mutation sites in the titin (TTN) gene. Moreover, TTN mutations were not found in a cohort of 50 sporadic MMD cases. They also analyzed mutation frequencies and used bioinformatic predictions to reveal mutation harmfulness, functions, and probabilities of disease correlation. rs771533925 and rs72677250 were likely harmful mutations with the involvement of TTN in MMD etiology-related pathways. CRISPR-Cas12a assays designed to detect TTN mutations provided results consistent with THEYS analysis, which was further confirmed by Sanger sequencing. This study recognized TTN as a new familial gene marker for moyamoya disease and demonstrated that CRISPR-Cas12a has the advantages of rapid detection, low cost, and simple operation, and has broad prospects in the practical application of rapid detection of MMD mutation sites.
- Fan et al. explored the pharmacological mechanisms of Chongcaoyishen decoction (CCYSD) against chronic kidney disease (CKD) *via* network pharmacology analysis combined with experimental validation. The bioactive components and potential regulatory targets of CCYSD were extracted from the TCMSP database, and the putative CKD-related target proteins were collected from the GeneCards and OMIM database. 114 kinds of cellular functional activities and 112 related cellular signaling pathways were involved in this network pharmacological analysis. Except for the autophagy and oxidative stress injury, the mechanism of CCYSD against CKD may also relate to inflammatory injury, cell cycle regulation, apoptosis, and other mechanisms. Their work provided an integrative network pharmacology approach combined with *in vivo* experiments to explore underlying mechanisms governing the CCYSD, promoting the explanation and understanding of CCYSD in CKD's treatment.
- Chen et al. aimed to illustrate what topics the research focused on and how they varied in different periods of all the studies on brain metastases with topic modeling. They used the latent Dirichlet allocation model to analyze the titles and abstracts of 50,176 articles on brain metastases retrieved from web of Science, Embase, and MEDLINE. The work further stratified the articles to find out the topic trends of different periods. The study identified that a rising number of studies on brain metastases were published in recent decades at a higher rate than all cancer articles. Overall, the major themes focused on treatment and histopathology. Radiotherapy took over the first and third places in the top 20 topics. Since the 2010s, increasing attention concerned with gene mutations. Targeted therapy was a popular topic of brain metastases research after 2020.
- Yi et al. found candidate prognostic biomarkers and provided clinicians with an accurate method for survival prediction of ACC *via* bioinformatics methods. Linear discriminant analysis, K-nearest neighbor, support vector machine, and time-dependent ROC were performed to identify meaningful prognostic biomarkers (MPBs). Four MPBs (ASPM, BIRC5, CCNB2, and CDK1) with high accuracy of survival prediction were screened out, and their mutations and copy number variants were associated with the overall survival of ACC patients. They established two nomograms which provided clinicians with an accurate, quick, and visualized method for survival prediction, which might constitute a breakthrough in the treatment and prognosis prediction of patients with ACC.
- Li et al. aimed to investigate if machine learning approaches can be used to predict postoperative unplanned 30-day hospital readmission in old surgical patients. They extracted demographic, comorbidity, laboratory, surgical, and medication data of elderly patients older than 65 who underwent surgeries under general anesthesia in west China Hospital, Sichuan University from July 2019 to February 2021. Different machine learning approaches were performed to evaluate whether unplanned 30-day hospital readmission can

be predicted. Model performance was assessed using the following metrics: AUC, accuracy, precision, recall, and F1 score; and RF + XGBoost showed the best prediction capability. The most five important features of RF + XGBoost were operation duration, white blood cell count, BMI, total bilirubin concentration, and blood glucose concentration. Machine learning algorithms can accurately predict postoperative unplanned 30-day readmission in elderly surgical patients.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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: 20 December 2021
doi: 10.3389/fmolb.2021.790804

Identification of a Novel Defined Immune-Autophagy-Related Gene Signature Associated With Clinical and Prognostic Features of Kidney Renal Clear Cell Carcinoma

Guangyuan Zhang[1,2], *Lei Zhang*[1,2], *Si Sun*[1]* and *Ming Chen*[1,2,3]*

[1]Department of Urology, Zhongda Hospital, Southeast University, Nanjing, China, [2]Surgical Research Center, Institute of Urology, Southeast University Medical School, Nanjing, China, [3]Department of Urology, Nanjing Lishui District People's Hospital, Zhongda Hospital Lishui Branch, Southeast University, Nanjing, China

OPEN ACCESS

Edited by:
Chang Xu,
Qatar University, Qatar

Reviewed by:
Xiaolong Wang,
Temple University, United States
Rui Cao,
Capital Medical University, China

Correspondence:
Ming Chen
mingchenseu@126.com
Si Sun
1375828752@qq.com

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *07 October 2021*
Accepted: *22 November 2021*
Published: *20 December 2021*

Citation:
Zhang G, Zhang L, Sun S and Chen M (2021) Identification of a Novel Defined Immune-Autophagy-Related Gene Signature Associated With Clinical and Prognostic Features of Kidney Renal Clear Cell Carcinoma. Front. Mol. Biosci. 8:790804. doi: 10.3389/fmolb.2021.790804

Background: As a common cancer of the urinary system in adults, renal clear cell carcinoma is metastatic in 30% of patients, and 1–2 years after diagnosis, 60% of patients die. At present, the rapid development of tumor immunology and autophagy had brought new directions to the treatment of renal cancer. Therefore, it was extremely urgent to find potential targets and prognostic biomarkers for immunotherapy combined with autophagy.

Methods: Through GSE168845, immune-related genes, autophagy-related genes, and immune-autophagy-related differentially expressed genes (IAR-DEGs) were identified. Independent prognostic value of IAR-DEGs was determined by differential expression analysis, prognostic analysis, and univariate and multivariate Cox regression analyses. Then, the lasso Cox regression model was established to evaluate the correlation of IAR-DEGs with the immune score, immune checkpoint, iron death, methylation, and one-class logistic regression (OCLR) score.

Results: In this study, it was found that CANX, BID, NAMPT, and BIRC5 were immune-autophagy-related genes with independent prognostic value, and the risk prognostic model based on them was well constructed. Further analysis showed that CANX, BID, NAMPT, and BIRC5 were significantly correlated with the immune score, immune checkpoint, iron death, methylation, and OCLR score. Further experimental results were consistent with the bioinformatics analysis.

Conclusion: CANX, BID, NAMPT, and BIRC5 were potential targets and effective prognostic biomarkers for immunotherapy combined with autophagy in kidney renal clear cell carcinoma.

Keywords: immune-autophagy, kidney renal clear cell carcinoma, prognosis, biomarkers, autophagy

INSTRUCTION

Renal cell carcinoma (RCC), which originates from renal tubular epithelial cells, has always been one of the most common malignant tumors, second only to bladder cancer in adult urinary system malignancies (Siegel et al., 2020). Among them, kidney renal clear cell carcinoma (KIRC) is the most common subtype (accounting for 70–80% of all RCC cases), and it is also one of the most aggressive subtypes with the worst prognosis (Linehan, 2012; Vuong et al., 2019). These tumors are asymptomatic in the early stages of the disease and are usually diagnosed by complications of distant metastasis in the later stages (Hsieh et al., 2017; Tito et al., 2021), 60% of patients with renal clear cell carcinoma die within 1–2 years after diagnosis, and 30% of patients have distant metastases at the time of diagnosis (Casuscelli et al., 2019). Treatment of KIRC can be partial or radical nephrectomy, ablation therapy, and active monitoring of KIRC, while metastatic tumors are treated with therapeutic action, but the overall prognosis is still limited, and immune-related adverse events still need to be improved (Choueiri and Motzer, 2017; Loo et al., 2019; Hofmann et al., 2020; Rizzo et al., 2021). Due to the complex etiology of KIRC and the high heterogeneity of tumor tissues, the treatment and diagnosis of patients are still not ideal. Therefore, it is urgent to find new markers to guide the clinical treatment and diagnosis of KIRC.

Previous studies have confirmed that KIRC is closely related to von Hippel-Lindau (VHL) gene changes (Zhang et al., 2018; Zhang et al., 2020). In addition, ferroptosis-related genes, some miRNAs, and pathways also participate in regulating the process of KIRC regulation (Lu et al., 2021; Zhang et al., 2021). Autophagy plays a vital role in cell physiology, including adaptation to metabolic stress, removal of dangerous substances, renewal during differentiation and development, and prevention of genome damage (Levine and Kroemer, 2008; Levine and Kroemer, 2019). Enormous studies have shown that autophagy is a double-edged sword in the occurrence and treatment of tumors. On the one hand, autophagy can degrade damaged organelles before cell canceration to maintain cell homeostasis and exert a tumor suppressor effect; on the other hand, autophagy can promote the circulation of cell metabolites and meet the nutritional needs of cells. Therefore, in the advanced stage of tumor development, autophagy can provide energy and nutrition for tumor cell proliferation and invasion and can improve tumor cell tolerance to radiotherapy and chemotherapy (Galluzzi et al., 2015; Galluzzi and Green, 2019; Kocaturk et al., 2019).

Since the relationship between iron death and tumors is regulated by many autophagy-related genes, the expression of autophagy-related genes in tumor tissues can be used to assess the prognosis of patients. This study obtained KIRC gene expression information by analyzing The Cancer Genome Atlas (TCGA) database and then analyzed the differential expression of immune autophagy-related genes in the sample, so as to construct a model containing multiple genes to effectively predict the survival of KIRC patients, analyze the risk scoring model correlation with immune status, explore potential mechanisms, provide diagnosis and treatment basis for clinical treatment, and find new therapeutic targets.

MATERIALS AND METHODS

Microarray Data Analysis and Screening of Immune-Autophagy-Related Differentially Expressed Genes

To compare immune-autophagy-related differentially expressed genes (IAR-DEGs) in KIRC, the Gene GEO database was used. The GSE186645 dataset was selected for subsequent analyses. A total of 1,793 human immune-related genes (IRGs) were downloaded from ImmPort database (https://www.immport.org./home), and a total of 223 human autophagy-related genes were downloaded from the Human Autophagy Database (HADb) (http://autophagy.lu/clustering/index.html). The cutoff conditions were set to an adjusted p-value <0.05, and the absolute value of log-fold change | log2FC| ≥ 1 was statistically significant for the DEGs. ImageGP was used to create volcano maps and venn maps online.

Functional Enrichment Analysis of Immune-Autophagy-Related Differentially Expressed Genes in Kidney Renal Clear Cell Carcinoma

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were performed by ClusterProfiler software package to explore functional annotation and enrichment pathways, with $p < 0.05$ representing statistically significant differences.

Survival Analysis and Verification

In order to further evaluate the expression and prognostic value of IAR-DEGs in KIRC, differential analysis and prognostic analysis through “survival” package were conducted. Based on the Cox proportional hazards model and Kaplan–Meier model, the hazard ratio (HR) was calculated, with $p < 0.05$ representing statistically significant differences.

Construction and Validation of the Immune-Autophagy-Related Differential Expressed Gene-Related Prognostic Model

According to the preliminary screening of IAR-DEGs with differentially expressed and prognostic significance, univariate Cox analysis of overall survival (OS) was performed to identify the survival-related IAR-DEGs with a significant prognosis value ($p < 0.05$). Then, multivariate Cox regression analysis was performed to construct a prediction model based on IAR-DEGs, and the IAR-DEGs were independent prognostic factors. Signatures were established based on the coefficients corresponding to independent prognostic genes. Patients from TCGA-KIRC dataset were divided into low- and high-risk groups weighted by the risk score obtained from the multivariate Cox regression. t-Distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA) were used to explore the distribution characteristics of different groups by R packages. Finally, the effectiveness of prognostic indicators was evaluated by the area under the curve (AUC) of "time receiver operating characteristic curve (ROC)."

Construction of Clinicopathological Correlation Analysis and the Nomogram

Based on "survival" package in R software, combined with the clinicopathological characteristics, the correlation between IAR-DEGs and clinicopathological characteristics was analyzed. Through R package "rms," the nomogram and calibration curve were obtained. Risk scores associated with prognostic models were used as prognostic factors to evaluate 1-, 3-, and 5-year OS.

Relationship Between Immune-Autophagy-Related Differentially Expressed Genes and Immune Microenvironment

The relationship between IAR-DEGs expression levels and immune cells was analyzed using the xCell algorithm in the "immunedeconv" R package. The immune score and the effects of gene expression levels on eight immune checkpoint-related genes were also analyzed using the "ggplot2" R package. Finally, TIDE algorithm was used to evaluate two different mechanisms of tumor immune escape using IAR-DEG markers.

Relationship Between Methylation and Ferroptosis With Immune-Autophagy-Related Differentially Expressed Genes

The third-order RNA sequencing data of genes were obtained based on TCGA dataset, and the association with ferroptosis-related genes and m^6A-related genes in "ggplot2" R package was analyzed.

One-Class Logistic Regression Scores of Immune-Autophagy-Related Differentially Expressed Genes in Kidney Renal Clear Cell Carcinoma

Tumor-associated RNA-seq data were obtained from TCGA-KIRC, mRNAsi was calculated by one-class logistic regression (OCLR) algorithm, and the dryness index was obtained.

Cell Lines, Patient Samples, RNA Extraction, and Quantitative Real-Time PCR

Human kidney cell line HK-2 and human KIRC cell lines, 786-O and caki-1, were originally purchased from the cell repository of Shanghai Institute of Life Sciences. The cells were cultured in 1640 Medium (Gibco, Grand Island, NY, USA), containing 10% fetal bovine serum (FBS) (Gibco), penicillin (25 U/ml), and streptomycin (25 mg/ml), with 5% CO_2 environment.

In this study, 19 fresh samples, including tumor tissues and adjacent normal kidney tissues, were collected from patients who underwent laparoscopic radical nephrectomy for KIRC from 2019 to 2020 in the Department of Urology, Zhongda Hospital, and stored at 80°C. All patients were diagnosed with KIRC and did not receive any antitumor therapy preoperatively, and none of them had a history of long-term drug use. The clinical characteristics of 19 clear cell RCC (ccRCC) patients are listed in **Table 1**. The methodology of this study followed the criteria outlined in the Declaration of Helsinki (revised in 2013), and ethical approval was obtained from the Ethics Committee and Institutional Review Board for Clinical Research of Zhongda Hospital (ZDKYSB077). All patients or their relatives who participated were informed and signed an informed consent form.

Total RNA was isolated with Total RNA Kit (OMEGAbiotec, Guangzhou, China) according to the manufacturer's instructions. Complementary DNA was synthesized using the HiScript II Q RT SuperMix (R223-01) reagent kit (Vazyme Biotech Co., ltd., Nanjing, China). The qRT-PCR was performed using the SYBR green PCR mix (vazyme). The specific primers set for mIR-DEGs and GAPDH are listed in **Supplementary Table S1**. Data were normalized to GAPDH expression levels using the $2^{-\Delta\Delta Ct}$ method.

Tissue Microarray Construction and Immunohistochemistry

All specimens were fixed in 10% neutral formaldehyde solution and embedded in paraffin. Envision two-step dyeing and DAB color development were used. Primary antibodies BID (ab32060, Abcam, Cambridge, UK), NAMPT (ab236874, Abcam), and BIRC5 (ab76424, Abcam) were used in this study.

Western Blotting Analysis

Total proteins from HK-2 and human KIRC cells lysed in radioimmunoprecipitation assay (RIPA) (KeyGen, Nanjing, China) buffer were extracted and quantified by bicinchoninic acid (BCA) assay (KeyGen, China). Proteins were analyzed by 10% sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE), and the gels were transferred onto polyvinylidene fluoride (PVDF) membranes. Then, bovine serum albumin (BSA)-blocked PVDF membranes were incubated with specific primary antibodies BID (1:1,000; ab32060), NAMPT (1:1000; ab236874), BIRC5 (1:5,000; ab76424), and CANX (1:2000; ab133615) overnight at 4°C, followed by incubation of secondary antibodies for 1 h. Finally, bands were visualized using an enhanced chemiluminescence (ECL) kit (vazyme, China).

Statistical Analysis

The statistical analysis was carried out by R software (version 4.0.2). The Perl programming language (version 5.30.2) was used for data processing. Multivariate Cox regression analyses were used to evaluate prognostic significance. When $p < 0.05$ or log-rank $p < 0.05$, the difference was statistically significant.

TABLE 1 | Clinical characteristics of 19 ccRCC patients.

Sample number	Age	Gender	AJCC	T	N	M	Fuhrman	Tumor size (cm)	Chemotherapy	Radiotherapy
1	59	Female	I	T1	N0	M0	I	3	No	No
2	74	Male	III	T3	N0	M0	II	3.1	No	No
3	52	Male	I	T1	N0	M0	II	6	No	No
4	78	Female	III	T3	N0	M0	III	8.5	No	No
5	82	Female	III	T3	N0	M0	II	4	No	No
6	54	Male	III	T1	N1	M0	I	2.5	No	No
7	46	Male	IV	T3	N0	M1	IV	16	No	No
8	64	Male	I	T1	N0	M0	III	3	No	No
9	23	Male	I	T1	N0	M0	II	2	No	No
10	82	Female	I	T1	N0	M0	III	3.3	No	No
11	77	Male	I	T1	N0	M0	II	3.4	No	No
12	68	Male	I	T1	N0	M0	II	0.8	No	No
13	43	Male	IV	T4	N0	M0	II	9.5	No	No
14	65	Female	III	T3	N1	M0	IV	10	No	No
15	70	Female	II	T2	N0	M0	II	9	No	No
16	58	Male	I	T1	N0	M0	II	4.3	No	No
17	74	Female	I	T1	N0	M0	II	2	No	No
18	40	Male	IV	T3	N0	M1	III	10.7	No	No
19	44	Male	I	T1	N0	M0	I	1.8	No	No

Note. ccRCC, clear cell renal cell carcinoma; AJCC, American Joint Committee on Cancer.

RESULTS

Identification of Immune-Autophagy-Related Differentially Expressed Genes in Kidney Renal Clear Cell Carcinoma Compared With Normal Renal Tissues

The volcano map shows 1,826 upregulated DEGs and 1,809 downregulated DEGs that we screened in GSE168845 (**Figure 1A**). Then, 1,793 human IRGs from ImmPort database and 223 human autophagy-related genes from HADb were analyzed by Venn diagram, and five co-expressed genes were obtained: CANX, MAPK1, BIRC5, NAMPT, and BID (**Figure 1B**). In the GO/KEGG pathway enrichment analyses, we found five co-expressed differential genes enriched in "aging" in biological process (BP); "dendrite cytoplasm," "neuron projection cytoplasm," and "plasma membrane projection cytoplasm" in cellular component (CC); and molecular function (MF) enriched in "MAP kinase activity," "MAP kinase activity," and "death receptor binding." Importantly, the five co-expressed genes in KEGG were mainly enriched in "Platinum drug resistance," "Apoptosis," and "Apoptosis-multiple species" (**Figure 1C**).

Differential Expression Analysis and Survival Analysis of Immune-Autophagy-Related Differentially Expressed Genes in Kidney Renal Clear Cell Carcinoma

Through a screening in TCGA-KIRC database, we compared the expression levels of CANX, MAPK1, BIRC5, NAMPT, and BID in normal kidney tissues and renal clear cell tumor tissues, and we found that their expression levels in tumor tissues were upregulated (**Figure 2A**). And Kaplan–Meier model analysis shows that the expression levels of the above five DEGs are significantly related to the prognosis, the high expression of CANX and MAPK1 is associated with a good prognosis (**Figures 2C,F**), and the high expression of BID, BIRC5, and MAPK1 is associated with a poor prognosis (**Figures 2B,D,E**). Univariate Cox regression analysis (**Figure 3A**) and multivariate Cox regression analysis (**Figure 3B**) were used to further explore the correlation between the five DEGs and prognosis, showing that CANX, BIRC5, NAMPT, and BID are independent prognostic factors for KIRC.

Construction and Validation of the Immune-Autophagy-Related Differentially Expressed Gene Prognostic Risk Model

We used lasso Cox regression to construct a prognostic model of DEG-related risks, Risk Score = (−0.4879) * CANX + (0.3075) * NAMPT + (−0.3041) * BIRC5 + (0.694) * BID (**Figure 4A, Figure 4B**). According to the median risk score (50%), patients were divided into high-risk and low-risk groups. It can be seen in the t-SNE and PCA heat maps that BID, BIRC5, and NAMPT are highly expressed in the high-risk group, and CANX is low in the high-risk group (**Figure 4C**). If HR = 2.333, the prognosis model can be considered as a risk factor model. The median survival time of the high-risk group was significantly lower than that of the low-risk group (**Figure 4C**). We used ROC to evaluate the prognostic prediction efficiency of the model, and the results showed that the AUC was 0.73 (1-year OS), 0.685 (3-year OS), and 0.697 (5-year OS) (**Figure 4C**).

Relationship Between Immune-Autophagy-Related Differentially Expressed Genes and Clinicopathological Factors and the Construction Nomogram

Regarding the correlation between CANX, BID, NAMPT, BIRC5, and clinicopathological characteristics in the risk prognosis

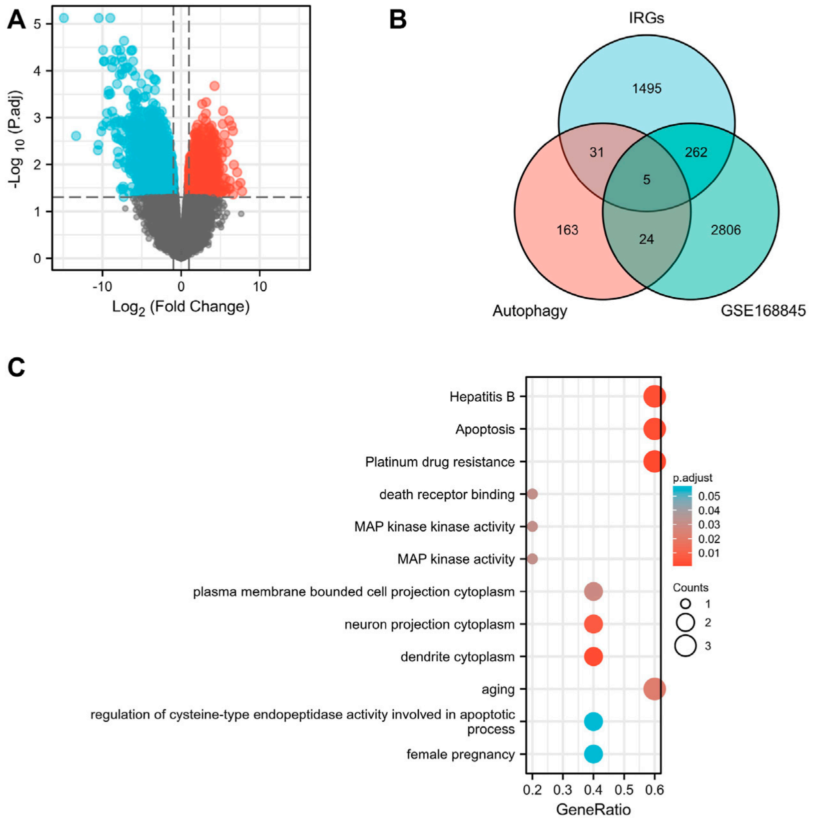


FIGURE 1 | Screening of differentially expressed genes. Volcano plots of differentially expressed genes (DEGs) between normal renal tissues and renal cancer in GSE168845 samples **(A)**. Adjusted *p*-value < 0.05 and log2-fold change (absolute) > 1.3; 635 DEGs were screened with 1,826 upregulated genes and 1,809 downregulated genes. Red represents upregulated genes, and blue indicates downregulated genes. A total of 1,793 human immune-related genes (IRGs) were downloaded from ImmPort database (https://www.immport.org./home), and a total of 223 human autophagy-related genes were downloaded from the Human Autophagy Database (HADb) (http://autophagy.lu/clustering/index.html). Venn diagram showing the five immune-autophagy genes according to the three datasets **(B)**. Graph showing the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of the five immune-autophagy genes **(C)**. The five immune-autophagy genes were CANX, MAPK1, BIRC5, NAMPT, and BID.

model, our results show that the immune-autophagy-related DEGs associated with T stage, N stage, M stage, and pathological stage are BIRC5 and BID (**Figures 5A–D**), there is no age-related gene, and the gene related to the patient's gender is BIRC5 (**Figures 5E,F**). We used the nomogram to predict 1-, 3-, and 5-year OS in the entire TCGA cohort (**Figure 5G**). We also found that the 1-, 3-, and 5-year OS on the nomogram is consistent with the calibration curve of predicted probability, and the 1-year OS is the highest (**Figure 5H**).

Gene Set Enrichment Analysis of Immune-Autophagy-Related Differentially Expressed Genes

We used Gene Set Enrichment Analysis (GSEA) to analyze the KIRC patient data in TCGA-KIRC database. The results showed that both BIRC5 and BID mediate ion channel transport (**Figures 6A,B**), and both NAMPT and CANX mediate the channel of NABA secretion (**Figures 6C,D**).

Correlation Between the Expression of Immune Infiltrating Cells in Kidney Renal Clear Cell Carcinoma Tissues and Immune-Autophagy-Related Differentially Expressed Genes

The KIRC population in TCGA-KIRC database was divided into immune-autophagy-related DEG low-expression group (G1) and immune-autophagy-related DEG high-expression group (G2), and the correlation between the expression of immune-infiltrating cells and immune-infiltrating cells was analyzed. The results show that CANX, NAMPT, BIRC5, and BID are highly correlated with the expression levels of a variety of immune infiltrating cells, and the expression levels of monocytes, myeloid dendritic cells, and $CD8^+$ effector memory T cells are significantly correlated with CANX, NAMPT, BIRC5, and BID. It suggests that these cells may be related to the progression of KIRC (**Figure 7**).

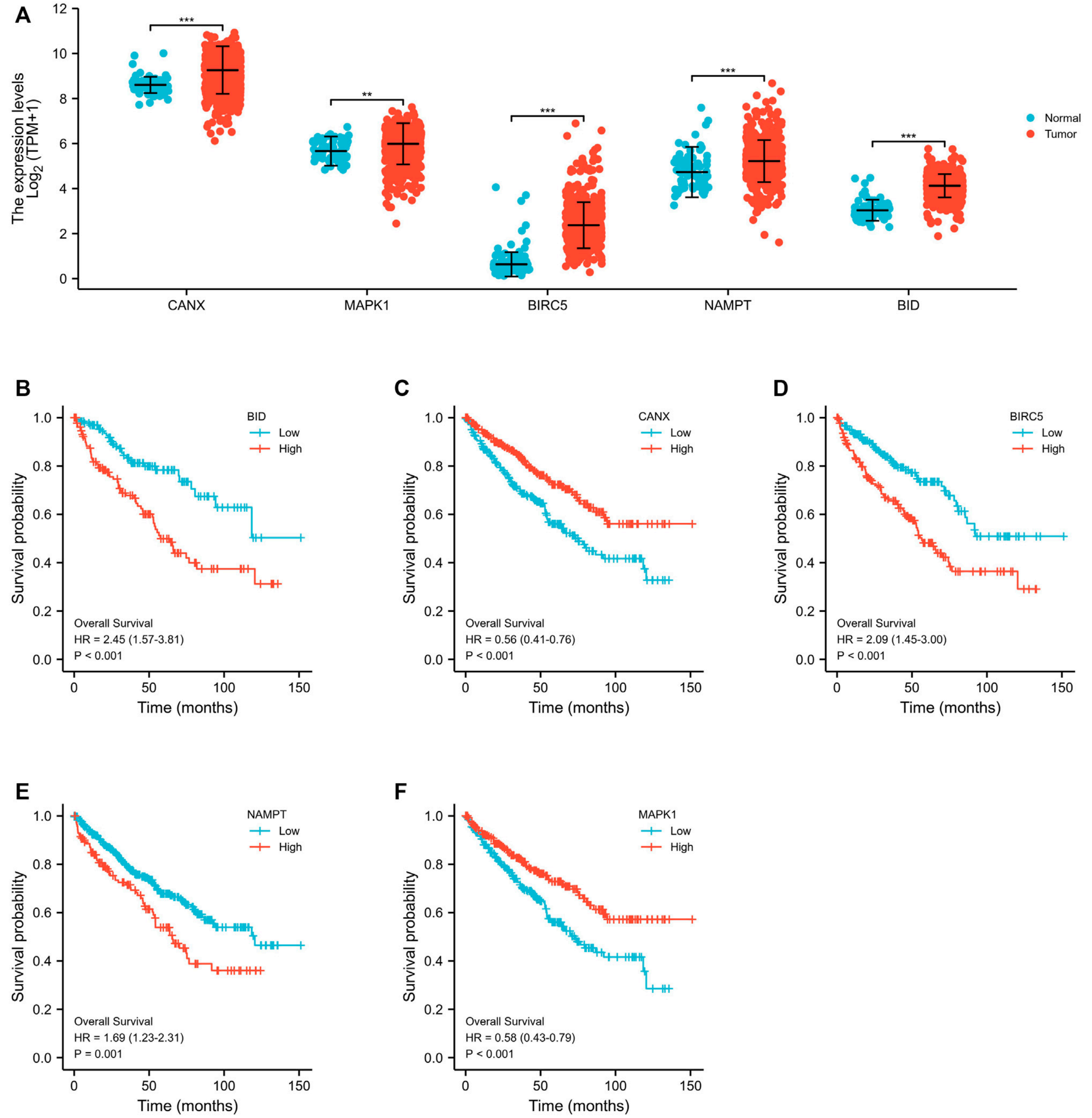


FIGURE 2 | Differential expression and survival analyses of immune-autophagy genes in kidney renal clear cell carcinoma (KIRC). Expression profile of the five immune-autophagy genes in KIRC samples compared with normal tissues **(A)**. Kaplan–Meier plots showing CANX, MAPK1, BIRC5, NAMPT, and BID with prognostic value **(B–F)**.

Correlation Between the Expression of Immune Checkpoint in Kidney Renal Clear Cell Carcinoma Tissues and Immune-Autophagy-Related Differentially Expressed Genes

Based on the original intention of this study to have a positive effect on the targeted drug therapy of KIRC, we also statistically analyzed the correlation between the expression level of immune checkpoints in KIRC tissues and the expression of immune-autophagy-related DEGs. The results showed that CD274, HAVCR2, LAG3, and PDCDILG2 were significantly correlated with BID (**Figure 8A**); CD274 and PDCDILG2 were significantly correlated with BIRC5 (**Figure 8C**); CTLA4, LAG3, PDCD1, PDCDILG2, TIGIT, and SIGLEC15 were significantly correlated with NAMPT (**Figure 8E**); and CD274, CTLA4, LAG3, PDCD1,

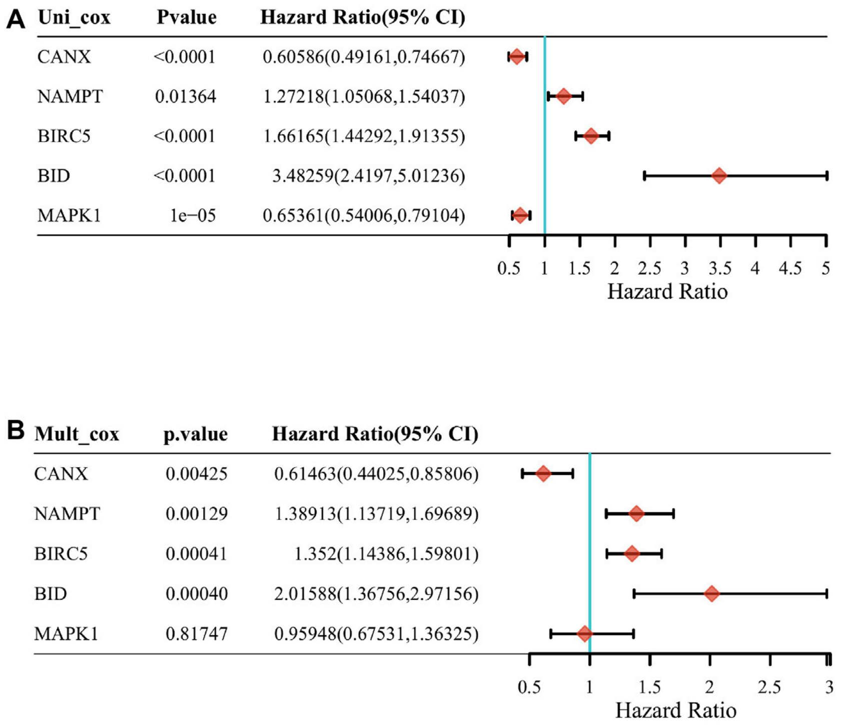


FIGURE 3 | The correlation between the five differentially expressed genes (DEGs) and prognosis. The forest plot shows the results of the univariate Cox regression analyses of the five immune-autophagy genes in The Cancer Genome Atlas–kidney renal clear cell carcinoma (TCGA-KIRC) **(A)**. The forest plot shows the results of the multivariate Cox regression analyses of the five immune-autophagy genes in TCGA-KIRC **(B)**. And CANX, BIRC5, NAMPT, and BID were significant.

and TIGIT were significantly correlated with CANX (**Figure 8G**). It can be seen that both CD274 and PDCDILG2 have appeared three times. It is speculated that they are sensitive immune checkpoints for KIRC treatment and diagnosis.

In addition, the response of CANX, BID, NAMPT, and BIRC5 with different expression levels to immune checkpoint inhibitors was predicted based on Tumor Immune Dysfunction and Exclusion (TIDE) algorithm (**Figures 8B–H**). The results indicated that the *p*-values of all immune-autophagy-related genes except NAMPT were <0.05, which indicated that the immune checkpoint inhibitors were effective against KIRC with high expression of CANX, BID, and BIRC5, and the survival period was prolonged after immune checkpoint inhibitor treatment.

Relationship Between Methylation, Ferroptosis, and Expression of Immune-Autophagy-Related Differentially Expressed Genes

Following the same analysis method, the results in **Figure 9** show that the expression of immune-autophagy-related DEGs is correlated with the expression levels of multiple ferroptosis-related genes, and NCOA4, EMC2, NFE2L2, HSPB1, SAT1, and DPP4 are significantly correlated with CANX, NAMPT, BIRC5, and BID.

In addition, we analyzed the correlation between m^{6}A methylation-related genes and immune-autophagy-related DEGs by the same method and found that CANX, NAMPT, BIRC5, and BID were significantly correlated with multiple methylated genes (**Figure 10**). We further verified that m^{6}A-related genes were differentially expressed in kidney cancer and normal tissues and were statistically significantly associated with patient prognosis (**Supplementary Figures S1, S2**). In particular, METTL14, VIRMA, ZC3H13, TYHDC2, YTHDF3, YTFDF2, IGF2BP2, and RBMX were significantly associated with four immune-autophagy-related DEGs.

Assessment of the One-Class Logistic Regression Scores of Immune-Autophagy-Related Differentially Expressed Genes in Kidney Renal Clear Cell Carcinoma

By OCLR scores, we found that, except for BID, the expression levels of CANX, NAMPT, and BIRC5 were significantly different from the dryness degree of KIRC (**Figure 11**). These results suggested that CANX, NAMPT, and BIRC5 may influence the degree of similarity between KIRC cells and stem cells and thus affect the BP and degree of dedifferentiation of tumors.

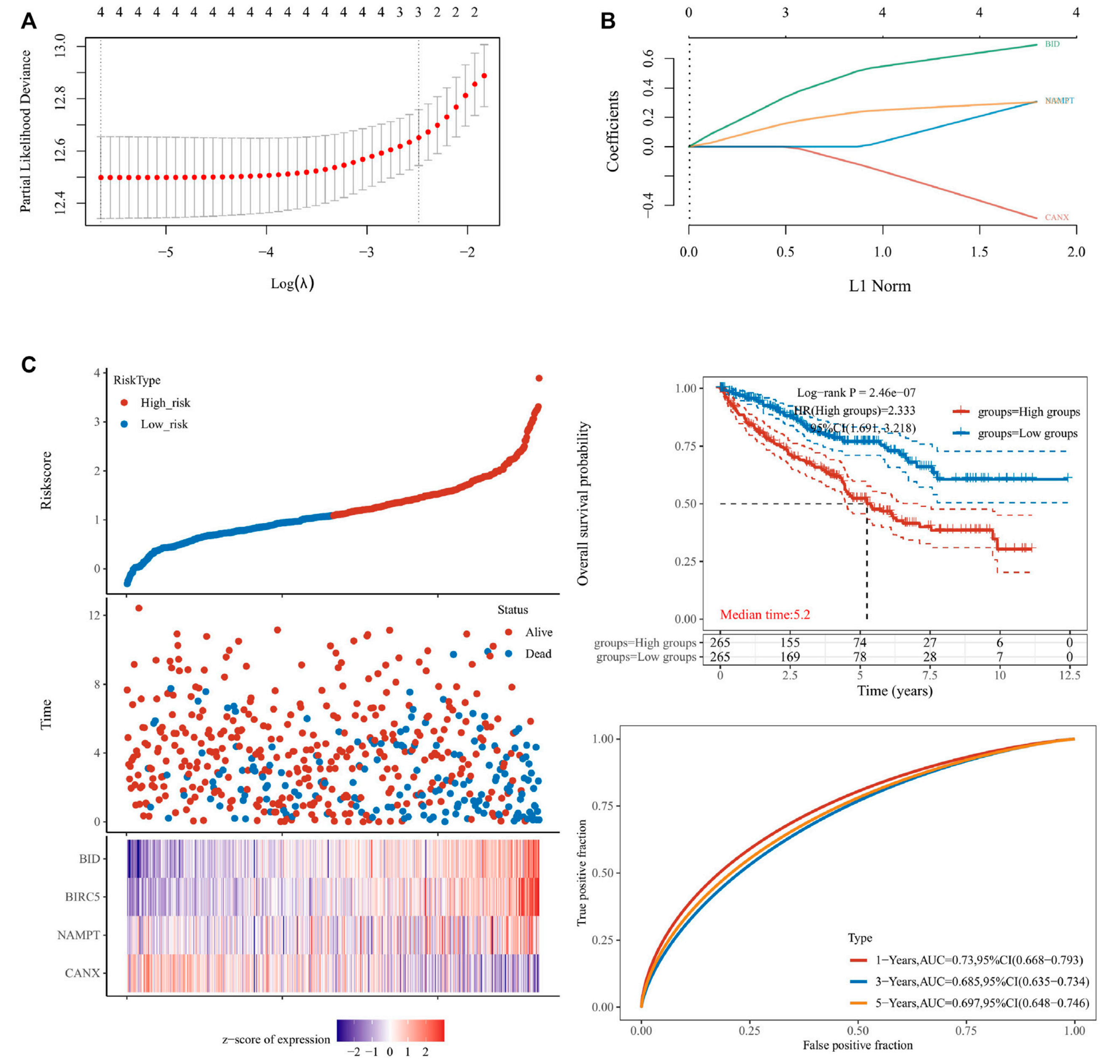


FIGURE 4 | Construction of a prognostic model for the risks associated with differentially expressed genes (DEGs). The calculations for the model according to the multivariate Cox regression analyses **(A,B)**. The prognostic model was analyzed by survival time, survival status, target gene expression heat map, and 1/3/5-year overall survival **(C)**. lambda.min = 0.0035. Riskscore = (−0.4879) * CANX + (0.3075) * NAMPT + (−0.3041) * BIRC5 + (0.694) * BID.

Validation of the Expression of Differentially Expressed Genes in Clinical Tissue Samples

To detect the expression of four genes (CANX, BID, NAMPT, and BIRC5) in KIRC, we performed the qRT-PCR in KIRC cells and clinical tissue samples. We verified the expression levels of four genes in normal kidney cell lines (HK-2 cells) and two KIRC cell lines (786-O and caki-1). The results showed that the expression levels of four genes were significantly increased in KIRC cells compared with normal kidney cells (**Figures 12A–D**). In addition, Western blotting results showed that protein levels of NAMPT and BIRC5 were expressed at increased levels in RCC cell lines 786 and caki-1, but there was no significant difference in protein levels of BID and CANX (**Figure 12J**). BID, NAMPT, and BIRC5 were detected with the same results in tumor tissues and with adjacent normal kidney tissues, while CANX was not significantly different (**Figures 12E–H**). Then we detected

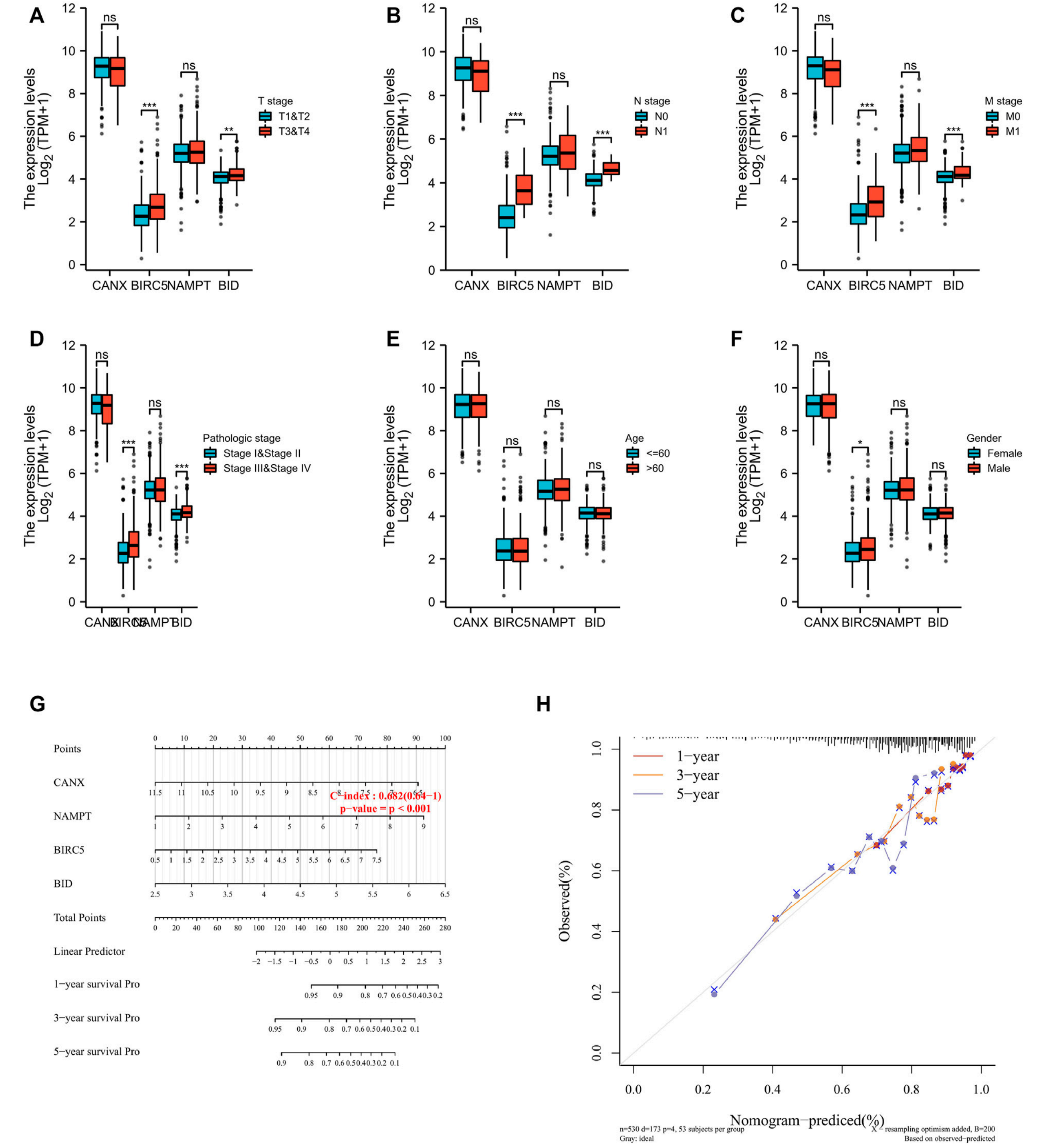


FIGURE 5 | The four immune-autophagy genes significantly correlate with multiple clinicopathological factors in kidney renal clear cell carcinoma (KIRC) patients. The relationships between CANX, BIRC5, NAMPT, and BID and clinicopathological factors in the entire The Cancer Genome Atlas (TCGA) cohort **(A–F)**. Nomogram for predicting 1-, 3-, and 5-year overall survival (OS) in the entire TCGA cohort **(G)**. Calibration curves of nomogram on consistency between predicted and observed 1-, 3-, and 5-year survival in entire TCGA cohort **(F)**. Dashed line at 45° indicates a perfect prediction.

the protein expression of BID, NAMPT, and BIRC5 in the tissues by immunohistochemistry (IHC). Results demonstrated that NAMPT and BIRC5 were significantly increased in KIRC tissues compared with adjacent normal kidney tissues. However, BID was negative in most tissues (**Figure 12I**).

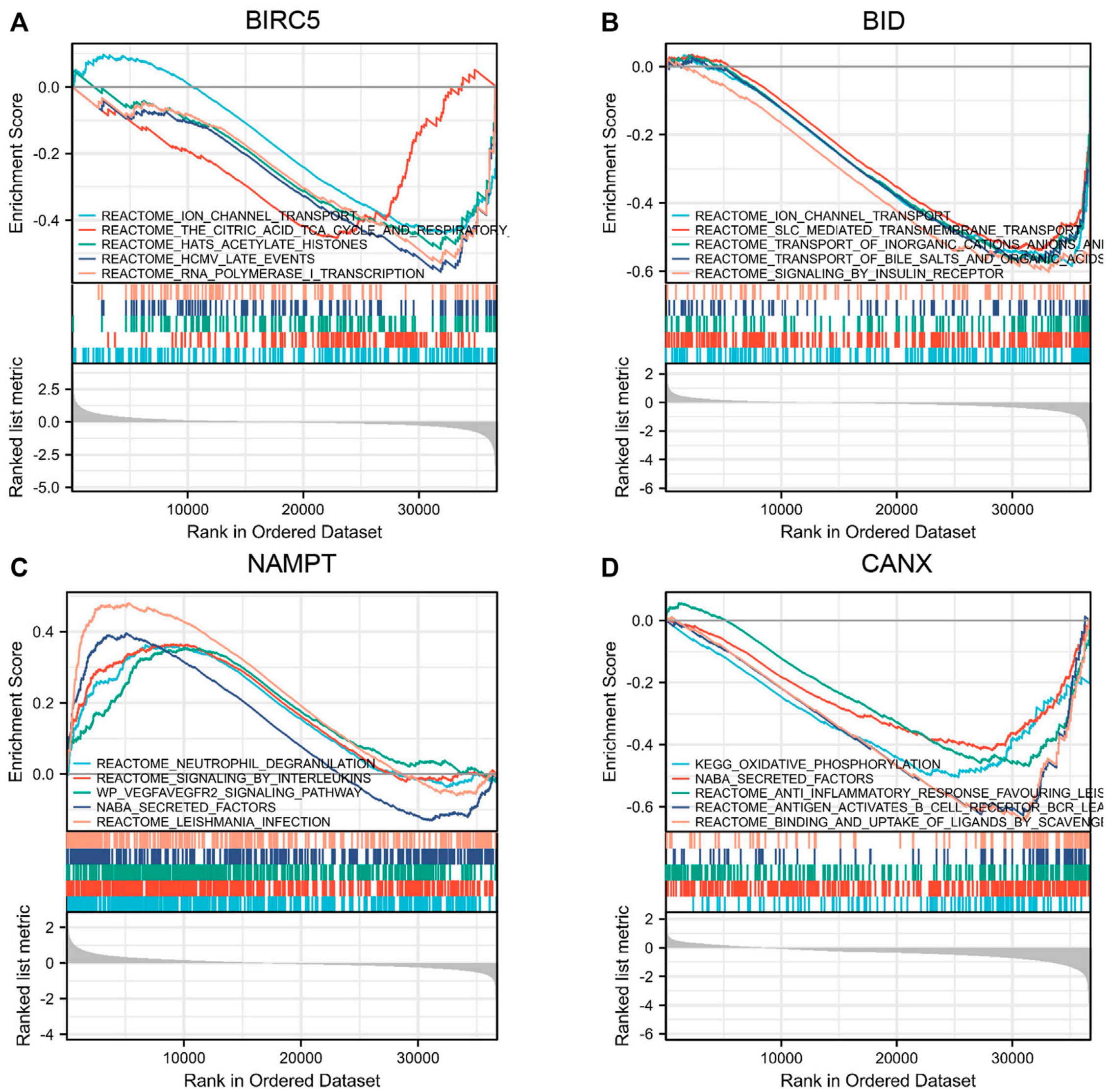


FIGURE 6 | Gene Set Enrichment Analysis (GSEA) of immune-autophagy-related differentially expressed genes (DEGs). Single gene enrichment analysis of BIRC5 **(A)**, BID **(B)**, NAMPT **(C)**, and CANX **(D)**.

DISCUSSION

Early symptoms of clear cell RCC are insidious, and patients often have metastases at the time of diagnosis. Because of its complex biological characteristics, surgical resection is not easy, more than one-tenth of patients will have a fatal relapse within 5 years after traditional partial or radical nephrectomy, and it is not sensitive to radiotherapy and chemotherapy (Fu et al., 2016; Pandey et al., 2020). In recent years, targeted therapies against vascular endothelial growth factor (VEGF) and immunotherapy have gradually replaced nonspecific immune methods as the primary medical treatment for patients with KIRC (Şenbabaoğlu et al., 2016; Barata and Rini, 2017; Smith et al., 2018; Dizman et al., 2020). Even though researchers have made some progress in this area, the selection of biomarkers, the combined use of drugs, and the ambiguity of immune checkpoints are still crucial issues that cannot be ignored (Tang et al., 2013; Ghatalia et al., 2017; Mao et al., 2021). Therefore, studying the mechanism of the occurrence and development of clear cell RCC has become a clinically urgent need to solve the problem. We understand that autophagy-related genes are closely related to cancer, and their expression levels differ at different cancer stages. Few studies are linking the prognosis and treatment of KIRC with autophagy-related genes. We hope to illustrate this kind of relevance through some analyses.

In this study, we first conducted Venn diagram analysis from the genes in the GSE168845, ImmPort database, and HADb to obtain five co-expressed immune-autophagy-related DEGs, and we discarded MAPK1 after performing multivariate Cox

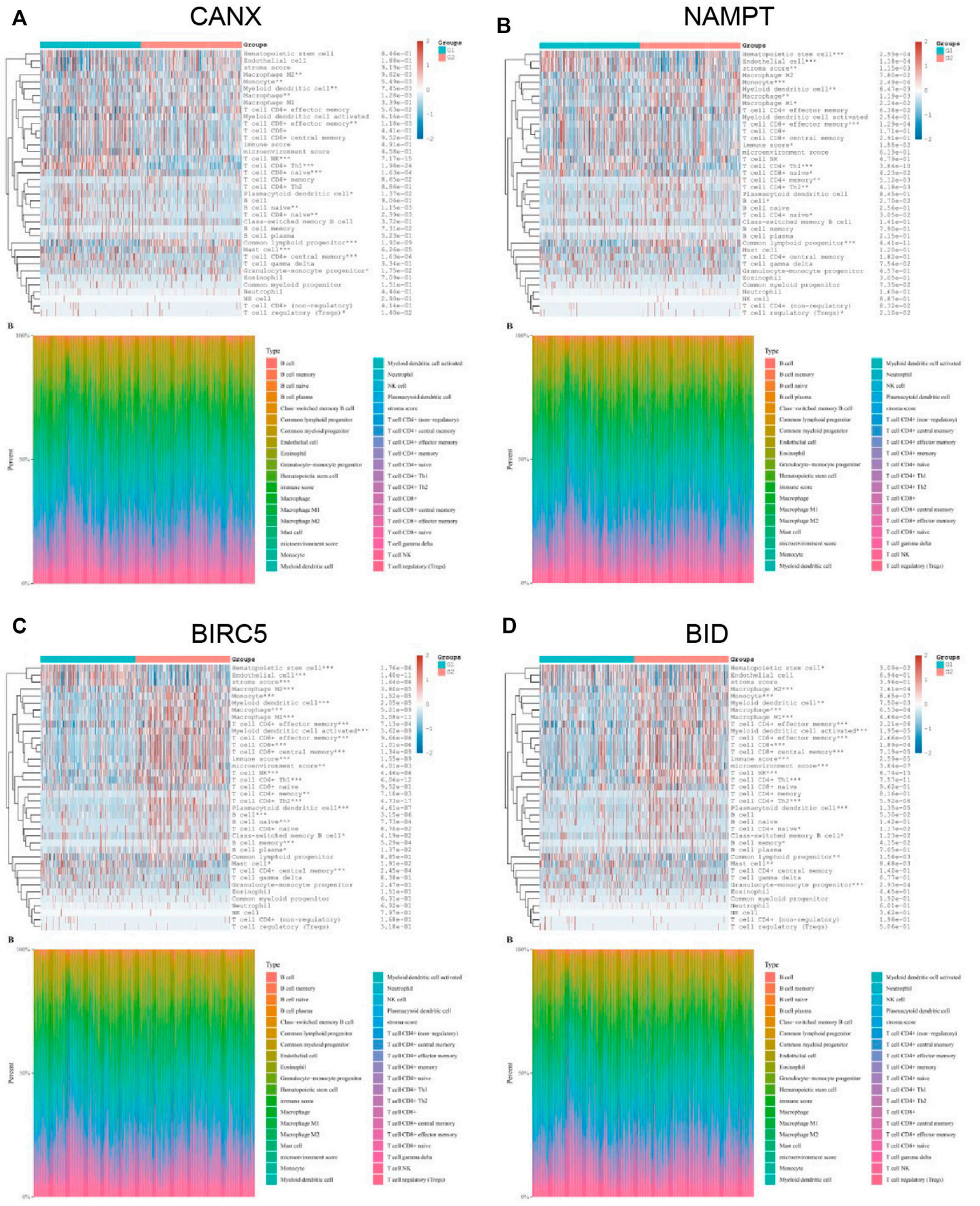


FIGURE 7 | Correlation between the expression of immune infiltrating cells in kidney renal clear cell carcinoma (KIRC) tissues and immune-autophagy-related differentially expressed genes (DEGs). The difference of expression of immune infiltration cells in KIRC tissues with high and low CANX **(A)**, NAMPT **(B)**, BIRC5 **(C)**, and BID **(D)** gene expression. G1 is a low-expression group, and G2 is a high-expression group.

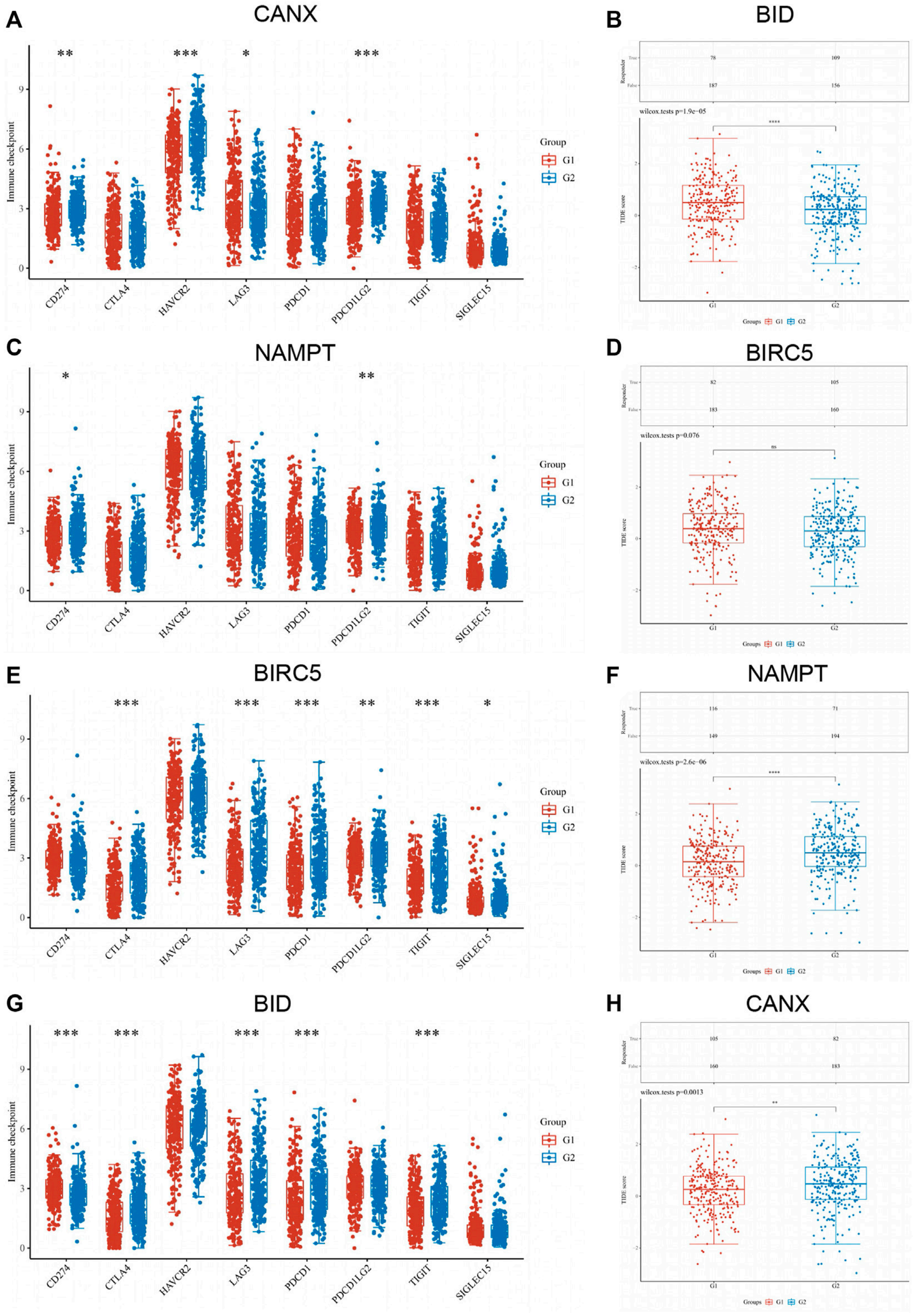


FIGURE 8 | Correlation between the expression of immune checkpoint in kidney renal clear cell carcinoma (KIRC) tissues and immune-autophagy-related differentially expressed genes (DEGs). The difference of expression of immune checkpoint in KIRC tissues with high and low CANX **(A)**, NAMPT **(C)**, BIRC5 **(E)**, and BID **(G)** gene expression. The difference of expression of ICB response in KIRC tissues with high and low BID **(B)**, BIRC5 **(D)**, NAMPT **(F)**, and CANX **(H)** gene expression. G1 is a low-expression group, and G2 is a high-expression group.

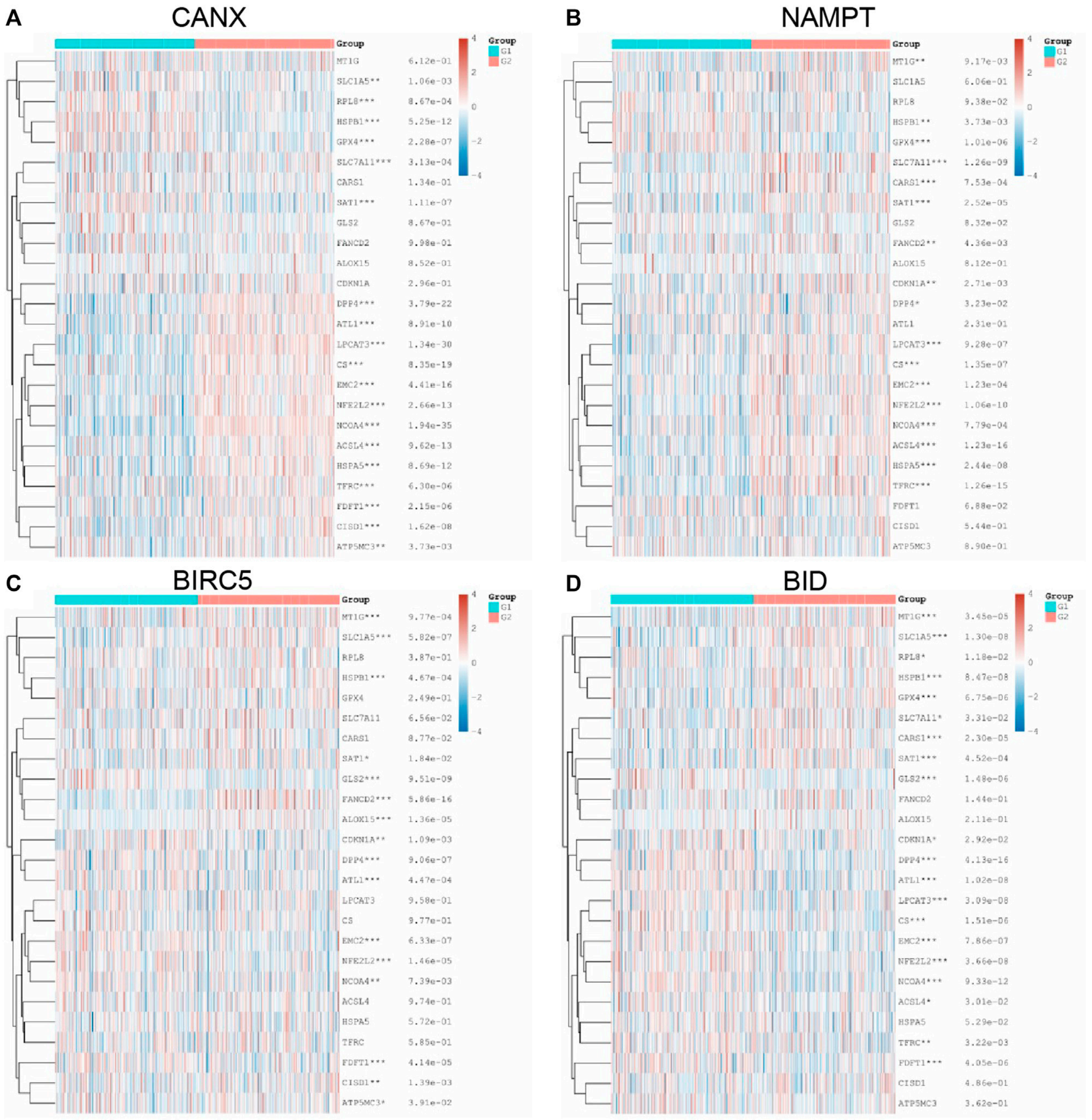


FIGURE 9 | Relationship between methylation, ferroptosis, and expression of immune-autophagy-related differentially expressed genes (DEGs). The difference of expression of ferroptosis-related genes in kidney renal clear cell carcinoma (KIRC) tissues with high and low CANX **(A)**, NAMPT **(B)**, BIRC5 **(C)**, and BID **(D)** gene expression. G1 is a low-expression group, and G2 is a high-expression group.

regression analysis. We found that the expression levels of CANX, BIRC5, BID, and NAMPT in tumor tissues were significantly higher than their expression levels in normal tissues, indicating that they are all significantly related to tumor occurrence and development. The Kaplan–Meier model we established shows that patients with high expression of BID and BIRC5 have a worse prognosis. By contrast, patients with high expression of CANX have a better prognosis, which is consistent with the results of our DEG-related risk prognosis model constructed by lasso Cox regression. In order to better understand the correlation between these four immune-autophagy-related DEGs and tumors, we also statistically analyzed their correlation with tumor stage, histopathological morphology, patient age, patient gender, and other clinicopathological characteristics. In addition,

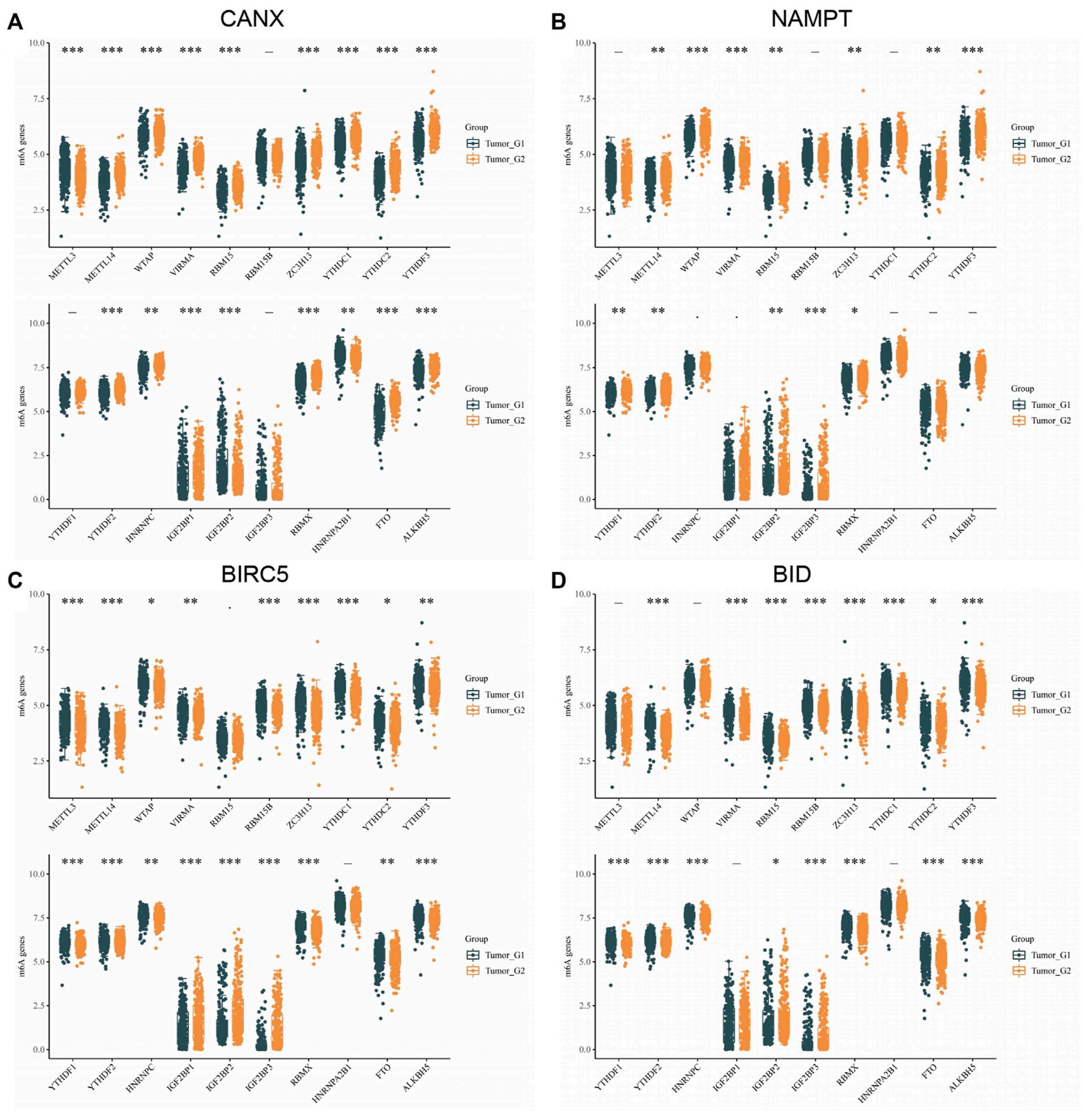


FIGURE 10 | Correlation between m^{6}A methylation-related genes and immune-autophagy-related differentially expressed genes (DEGs). The difference of expression of methylation of m^{6}A related genes in kidney renal clear cell carcinoma (KIRC) tissues with high and low CANX **(A)**, NAMPT **(B)**, BIRC5 **(C)**, and BID **(D)** gene expression. G1 is a low-expression group, and G2 is a high-expression group.

the calibration curves and nomogram showed a good prediction effect. The expression level of immune-autophagy-related DEGs is also significantly correlated with immune infiltration, immune checkpoints, methylation, and iron death. Among these results, the performance of BIRC5 and BID is particularly outstanding; immune infiltrating cells such as monocytes, myeloid dendritic cells, CD8$^+$ effector memory T cells, and immune checkpoint CD274 deserve special attention. Furthermore, we performed the qRT-PCR analysis and IHC in clinical samples and found that the expression of NAMPT and BIRC5 was significantly higher in ccRCC tissues when compared with that in adjacent normal tissues. More *in vivo* and *in vitro* experiments are needed to authenticate these findings.

Baculoviral IAP repeat containing 5 (BIRC5) has been broadly studied among cancer therapeutic targets, and its main function is to suppress cell death (Li et al., 2019). Numerous researches have

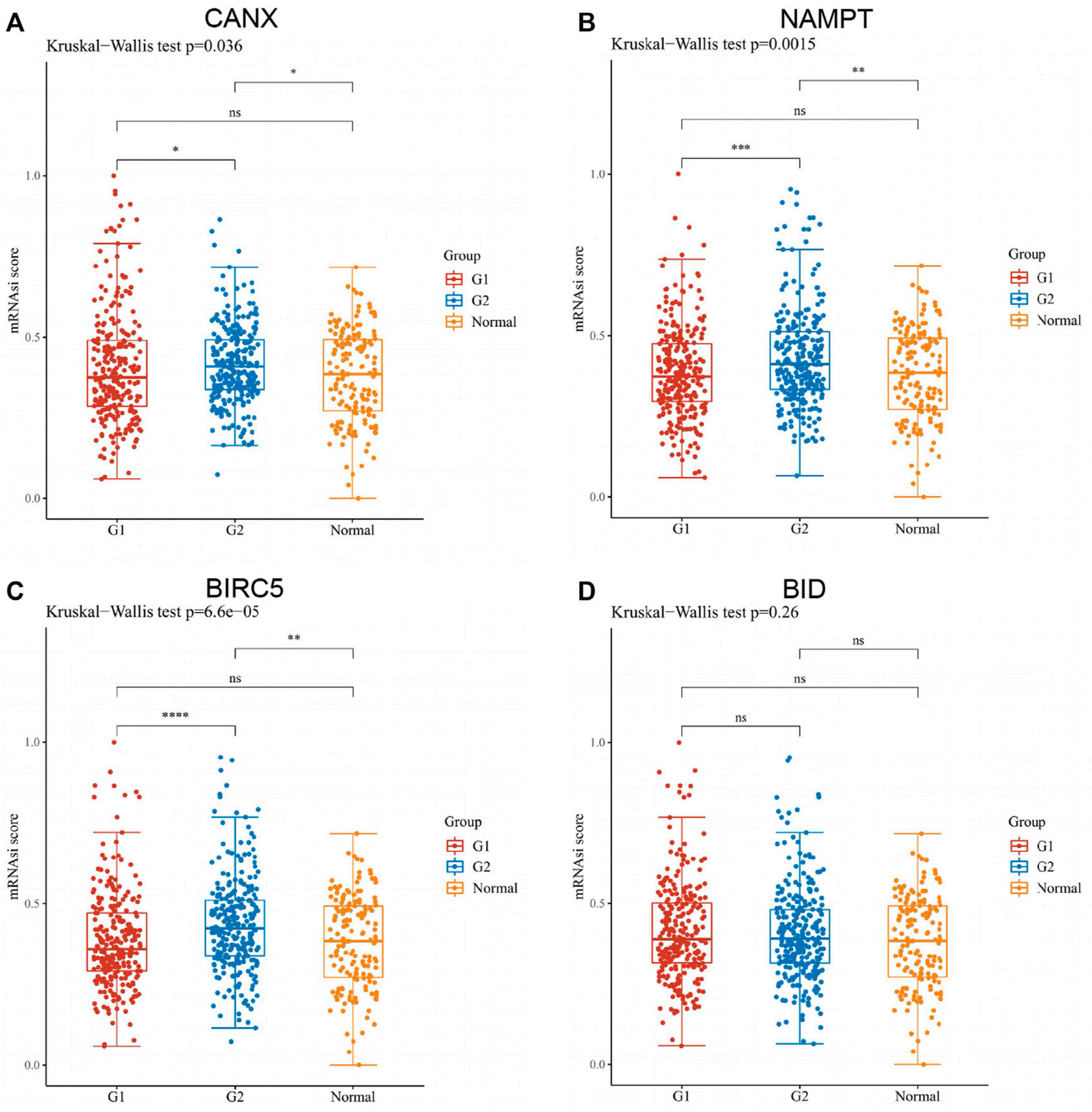


FIGURE 11 | Assessment of the one-class logistic regression (OCLR) scores of immune-autophagy-related differentially expressed genes (DEGs) in kidney renal clear cell carcinoma (KIRC). Scatter diagram illustrating the relationship between CANX **(A)**, NAMPT **(B)**, BIRC5 **(C)**, and BID **(D)** and OCLR score in KIRC. The horizontal axis in the figure represents the gene expression distribution, and the vertical axis is the OCLR score distribution. G1 is a low-expression group, and G2 is a high-expression group.

shown that BIRC5 contributes to tumor cell immune escape by inhibiting apoptosis and confirmed that its expression is strongly correlated with prognostic status and OS in various cancers (e.g., lung, colorectal, prostate, and ovarian cancers) (Cao et al., 2019; Filipchiuk et al., 2020; Wang et al., 2021). However, there are no relevant studies to explore the therapeutic effects of BIRC5 small-molecule inhibitors in tumors (Li et al., 2019). BH3-Interacting Domain Death Agonist (BID), as the activator and integrator, is involved in apoptosis-related pathways (Billen et al., 2008; Gryko et al., 2014). Lee found that BID proteins are involved in mediating DNA damage responses and promoting normal cell apoptosis (Lee et al., 2004). Regrettably, there are no relevant studies that explored the specific action mechanism and related functions of BID in tumors. Nonetheless, studies have suggested that TAT-BID + DOX may be a potentially effective combination for the treatment of cancers, but no final conclusions can be drawn due to the absence of protein and cytokine pathways (Zhang et al., 2004; Goncharenko-Khaider et al., 2010; Orzechowska et al., 2015). Additionally, nicotinamide phosphoribosyltransferase (NAMPT) is an important cofactor involved in various biochemical reactions (Travelli et al., 2018). It is now generally believed that NAMPT is highly expressed in cells with active proliferation, especially tumor cells (Garten et al., 2015), which implicates NAMPT-targeted small-molecule

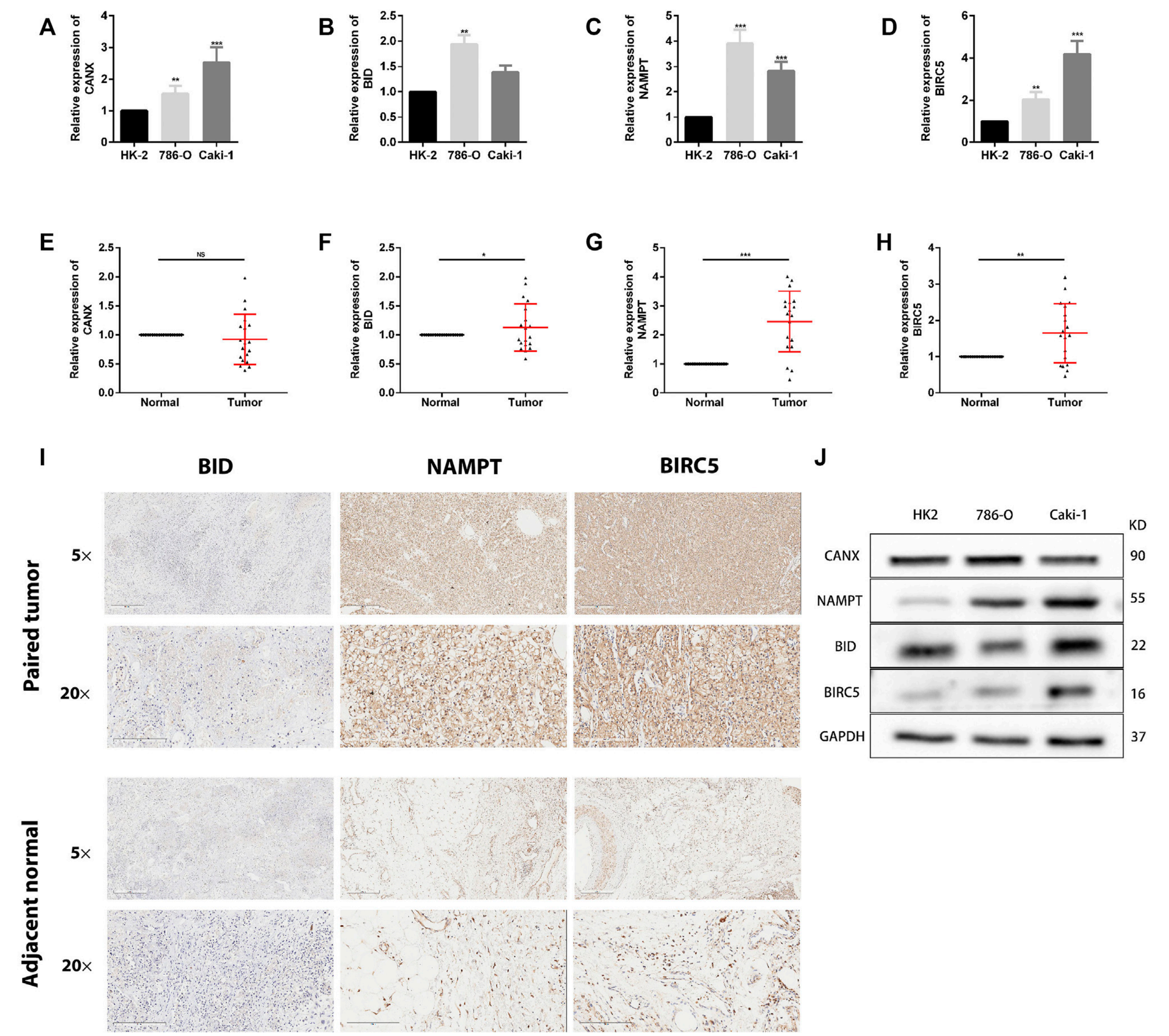


FIGURE 12 | The expression of these genes in human kidney renal clear cell carcinoma (KIRC) specimens, adjacent normal tissues, and cell lines. **(A–D)** qRT-PCR analysis of CANX **(A)**, BID **(B)**, NAMPT **(C)**, and BIRC5 **(D)** in KIRC cell lines. GAPDH was used as a loading control. **(E–H)** qRT-PCR analysis of CANX **(E)**, BID **(F)**, NAMPT **(G)**, and BIRC5 **(H)** in paired KIRC tissues ($n = 19$). **(I)** Representative images of BID, NAMPT, and BIRC5 protein immunochemistry in KIRC tissues compared with adjacent normal kidney tissues. Magnification, ×5 and ×20. **(J)** Western blotting analysis of related differentially expressed genes (DEGs) expression levels in normal kidney cell line (HK-2 cells) and two KIRC cell lines (786-O and caki-1). $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

inhibitors as potential tumor therapeutic agents. Existing related studies have identified NAMPT inhibitors and their vectors as important directions for anticancer therapy (Garten et al., 2009; Bi and Che, 2010; Lucena-Cacace et al., 2018; Zhang et al., 2019; Galli et al., 2020). Ultimately, calnexin (CANX), an endoplasmic reticulum lectin chaperone protein (Ellgaard et al., 2016; Kozlov and Gehring, 2020), has been confirmed to be upregulated in tumors including lung cancer and oral squamous carcinomas, and its ability to inhibit the proliferation of $CD4^{+}$ T and $CD8^{+}$ T cells in tumor tissues (Kobayashi et al., 2015; Alam et al., 2019; Chen et al., 2019), as well as the release of cytokines (PD-1, IFN-γ, and TNF), which eventually promotes tumor growth. Unfortunately, there is no clear mechanism for the regulation of CANX in tumors (Li et al., 2001; Kobayashi et al., 2015; Ellgaard et al., 2016; Alam et al., 2019; Chen et al., 2019; Kozlov and Gehring, 2020).

In summary, BIRC5, BID, NAMPT, and CANX, which were finally screened by bioinformatics analysis of autophagy-immune-related genes, are important in tumorigenesis, progression, and apoptosis. Regrettably, there are no relevant studies to explore their specific mechanisms and functions in KIRC and the potential efficacy of relevant targeted small-molecule inhibitors. We believe that this will be an important concept and direction for the academic community to investigate the mechanism and function of autophagy-immunity in renal cancer afterward.

Our study still had some limitations. The dataset we used to construct and validate the IAR-DEG prognostic signature was obtained from ImmPort database. We failed to locate suitable data from other immunological databases to verify the reliability of the screened genes. We only performed preliminary expression studies on these four IAR-DEGs in the signature. However, further functional analysis and mechanistic studies were not carried out.

CONCLUSION

In this study, we obtained immune-autophagy-related genes with independent prognostic value through comprehensive bioinformatics analysis. We established the prognostics risk model. A significant correlation was found among immune-autophagy-related genes and the immune score, immune checkpoint, methylation, ferroptosis, and OCLR score. As a result, CANX, BID, NAMPT, and BIRC5 were potential targets and effective prognostic biomarkers for immunotherapy combined with autophagy.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The 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

MC and SS were responsible for study design, data acquisition and analysis, and manuscript writing. GZ and LZ performed bioinformatics and statistical analyses. GZ and SS prepared the figures and tables for the manuscript. SS and MC were responsible for the integrity of the entire study and manuscript review. All authors read and approved the final manuscript.

FUNDING

This study was funded by Natural Science Foundation of China (82170703); Natural Science Foundation of China (82100732); Natural Science Foundation of Jiangsu Province (BK20200360); Excellent Youth Development Fund of Zhongda Hospital, SEU (2021ZDYYYQPY04).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.790804/full#supplementary-material

REFERENCES

Alam, A., Taye, N., Patel, S., Thube, M., Mullick, J., Shah, V. K., et al. (2019). SMAR1 Favors Immunosurveillance of Cancer Cells by Modulating Calnexin and MHC I Expression. *Neoplasia* 21, 945–962. doi:10.1016/j.neo.2019.07.002

Barata, P. C., and Rini, B. I. (2017). Treatment of Renal Cell Carcinoma: Current Status and Future Directions. *CA: a Cancer J. clinicians* 67, 507–524. doi:10.3322/caac.21411

Bi, T.-q., and Che, X.-m. (2010). Nampt/PBEF/visfatin and Cancer. *Cancer Biol. Ther.* 10, 119–125. doi:10.4161/cbt.10.2.12581

Billen, L. P., Shamas-Din, A., and Andrews, D. W. (2008). Bid: a Bax-like BH3 Protein. *Oncogene* 27, S93–S104. doi:10.1038/onc.2009.47

Cao, Y., Zhu, W., Chen, W., Wu, J., Hou, G., and Li, Y. (2019). Prognostic Value of BIRC5 in Lung Adenocarcinoma Lacking EGFR, KRAS, and ALK Mutations by Integrated Bioinformatics Analysis. *Dis. markers* 2019, 1–12. doi:10.1155/2019/5451290

Casuscelli, J., Becerra, M. F., Manley, B. J., Zabor, E. C., Reznik, E., Redzematovic, A., et al. (2019). Characterization and Impact of TERT Promoter Region Mutations on Clinical Outcome in Renal Cell Carcinoma. *Eur. Urol. focus* 5, 642–649. doi:10.1016/j.euf.2017.09.008

Chen, Y., Ma, D., Wang, X., Fang, J., Liu, X., Song, J., et al. (2019). Calnexin Impairs the Antitumor Immunity of CD4+ and CD8+ T Cells. *Cancer Immunol. Res.* 7, 123–135. doi:10.1158/2326-6066.Cir-18-0124

Choueiri, T. K., and Motzer, R. J. (2017). Systemic Therapy for Metastatic Renal-Cell Carcinoma. *N. Engl. J. Med.* 376, 354–366. doi:10.1056/NEJMra1601333

Dizman, N., Arslan, Z. E., Feng, M., and Pal, S. K. (2020). Sequencing Therapies for Metastatic Renal Cell Carcinoma. *The Urol. Clin. North America* 47, 305–318. doi:10.1016/j.ucl.2020.04.008

Ellgaard, L., McCaul, N., Chatsisvili, A., and Braakman, I. (2016). Co- and Post-Translational Protein Folding in the ER. *Traffic* 17, 615–638. doi:10.1111/tra.12392

Filipchiuk, C., Laganà, A. S., Beteli, R., Ponce, T. G., Christofolini, D. M., Martins Trevisan, C., et al. (2020). BIRC5/Survivin Expression as a Non-invasive Biomarker of Endometriosis. *Diagnostics* 10, 533. doi:10.3390/diagnostics10080533

Fu, Q., Chang, Y., Zhou, L., An, H., Zhu, Y., Xu, L., et al. (2016). Positive Intratumoral Chemokine (C-C Motif) Receptor 8 Expression Predicts High Recurrence Risk of post-operation clear-cell Renal Cell Carcinoma Patients. *Oncotarget* 7, 8413–8421. doi:10.18632/oncotarget.6761

Galli, U., Colombo, G., Travelli, C., Tron, G. C., Genazzani, A. A., and Grolla, A. A. (2020). Recent Advances in NAMPT Inhibitors: A Novel Immunotherapic Strategy. *Front. Pharmacol.* 11, 656. doi:10.3389/fphar.2020.00656

Galluzzi, L., and Green, D. R. (2019). Autophagy-Independent Functions of the Autophagy Machinery. *Cell* 177, 1682–1699. doi:10.1016/j.cell.2019.05.026

Galluzzi, L., Pietrocola, F., Bravo-San Pedro, J. M., Amaravadi, R. K., Baehrecke, E. H., Cecconi, F., et al. (2015). Autophagy in Malignant Transformation and Cancer Progression. *Embo J.* 34, 856–880. doi:10.15252/embj.201490784

Garten, A., Petzold, S., Körner, A., Imai, S.-i., and Kiess, W. (2009). Nampt: Linking NAD Biology, Metabolism and Cancer. *Trends Endocrinol. Metab.* 20, 130–138. doi:10.1016/j.tem.2008.10.004

Garten, A., Schuster, S., Penke, M., Gorski, T., de Giorgis, T., and Kiess, W. (2015). Physiological and Pathophysiological Roles of NAMPT and NAD Metabolism. *Nat. Rev. Endocrinol.* 11, 535–546. doi:10.1038/nrendo.2015.117

Ghatalia, P., Zibelman, M., Geynisman, D. M., and Plimack, E. R. (2017). Checkpoint Inhibitors for the Treatment of Renal Cell Carcinoma. *Curr. Treat. Options. Oncol.* 18, 7. doi:10.1007/s11864-017-0458-0

Goncharenko-Khaider, N., Lane, D., Matte, I., Rancourt, C., and Piché, A. (2010). The Inhibition of Bid Expression by Akt Leads to Resistance to TRAIL-Induced Apoptosis in Ovarian Cancer Cells. *Oncogene* 29, 5523–5536. doi:10.1038/onc.2010.288

Gryko, M., Pryczynicz, A., Zareba, K., Kędra, B., Kemona, A., and Guzińska-Ustymowicz, K. (2014). The Expression of Bcl-2 and BID in Gastric Cancer Cells. *J. Immunol. Res.* 2014, 1–5. doi:10.1155/2014/953203

Hofmann, F., Hwang, E. C., Lam, T. B., Bex, A., Yuan, Y., Marconi, L. S., et al. (2020). Targeted Therapy for Metastatic Renal Cell Carcinoma. *Cochrane database Syst. Rev.* 2020, Cd012796. doi:10.1002/14651858.CD012796.pub2

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

Kobayashi, M., Nagashio, R., Jiang, S.-X., Saito, K., Tsuchiya, B., Ryuge, S., et al. (2015). Calnexin Is a Novel Sero-Diagnostic Marker for Lung Cancer. *Lung Cancer* 90, 342–345. doi:10.1016/j.lungcan.2015.08.015

Kocaturk, N. M., Akkoc, Y., Kig, C., Bayraktar, O., Gozuacik, D., and Kutlu, O. (2019). Autophagy as a Molecular Target for Cancer Treatment. *Eur. J. Pharm. Sci.* 134, 116–137. doi:10.1016/j.ejps.2019.04.011

Kozlov, G., and Gehring, K. (2020). Calnexin Cycle - Structural Features of the ER Chaperone System. *Febs J.* 287, 4322–4340. doi:10.1111/febs.15330

Lee, J. H., Soung, Y. H., Lee, J. W., Park, W. S., Kim, S. Y., Cho, Y. G., et al. (2004). Inactivating Mutation of the Pro-apoptotic geneBID in Gastric Cancer. *J. Pathol.* 202, 439–445. doi:10.1002/path.1532

Levine, B., and Kroemer, G. (2008). Autophagy in the Pathogenesis of Disease. *Cell* 132, 27–42. doi:10.1016/j.cell.2007.12.018

Levine, B., and Kroemer, G. (2019). Biological Functions of Autophagy Genes: A Disease Perspective. *Cell* 176, 11–42. doi:10.1016/j.cell.2018.09.048

Li, F., Aljahdali, I., and Ling, X. (2019). Cancer Therapeutics Using Survivin BIRC5 as a Target: what Can We Do after over Two Decades of Study? *J. Exp. Clin. Cancer Res.* 38, 368. doi:10.1186/s13046-019-1362-1

Li, F., Mandal, M., Barnes, C. J., Vadlamudi, R. K., and Kumar, R. (2001). Growth Factor Regulation of the Molecular Chaperone Calnexin. *Biochem. Biophysical Res. Commun.* 289, 725–732. doi:10.1006/bbrc.2001.6001

Linehan, W. M. (2012). Genetic Basis of Kidney Cancer: Role of Genomics for the Development of Disease-Based Therapeutics. *Genome Res.* 22, 2089–2100. doi:10.1101/gr.131110.111

Loo, V., Salgia, M., Bergerot, P., Philip, E. J., and Pal, S. K. (2019). First-Line Systemic Therapy for Metastatic Clear-Cell Renal Cell Carcinoma: Critical Appraisal of Emerging Options. *Targ Oncol.* 14, 639–645. doi:10.1007/s11523-019-00676-y

Lu, T., Xu, R., Li, Q., Zhao, J. Y., Peng, B., Zhang, H., et al. (2021). Systematic Profiling of Ferroptosis Gene Signatures Predicts Prognostic Factors in Esophageal Squamous Cell Carcinoma. *Mol. Ther. Oncolytics* 21, 134–143. doi:10.1016/j.omto.2021.02.011

Lucena-Cacace, A., Otero-Albiol, D., Jiménez-García, M. P., Muñoz-Galvan, S., and Carnero, A. (2018). NAMPT Is a Potent Oncogene in Colon Cancer Progression that Modulates Cancer Stem Cell Properties and Resistance to Therapy through Sirt1 and PARP. *Clin. Cancer Res.* 24, 1202–1215. doi:10.1158/1078-0432.Ccr-17-2575

Mao, W., Wang, K., Xu, B., Zhang, H., Sun, S., Hu, Q., et al. (2021). ciRS-7 Is a Prognostic Biomarker and Potential Gene Therapy Target for Renal Cell Carcinoma. *Mol. Cancer* 20, 142. doi:10.1186/s12943-021-01443-2

Orzechowska, E., Girstun, A., Staron, K., and Trzcinska-Danielewicz, J. (2015). Synergy of BID with Doxorubicin in the Killing of Cancer Cells. *Oncol. Rep.* 33, 2143–2150. doi:10.3892/or.2015.3841

Pandey, N., Lanke, V., and Vinod, P. K. (2020). Network-based Metabolic Characterization of Renal Cell Carcinoma. *Sci. Rep.* 10, 5955. doi:10.1038/s41598-020-62853-8

Rizzo, A., Mollica, V., Santoni, M., Ricci, A. D., Rosellini, M., Marchetti, A., et al. (2021). Impact of Clinicopathological Features on Survival in Patients Treated with First-Line Immune Checkpoint Inhibitors Plus Tyrosine Kinase Inhibitors for Renal Cell Carcinoma: A Meta-Analysis of Randomized Clinical Trials. *Eur. Urol. focus* [Epub ahead of print]. doi:10.1016/j.euf.2021.03.001

Şenbabaoğlu, Y., Gejman, R. S., Winer, A. G., Liu, M., Van Allen, E. M., de Velasco, G., et al. (2016). Tumor Immune Microenvironment Characterization in clear Cell Renal Cell Carcinoma Identifies Prognostic and Immunotherapeutically Relevant Messenger RNA Signatures. *Genome Biol.* 17, 231. doi:10.1186/s13059-016-1092-z

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

Smith, C. C., Beckermann, K. E., Bortone, D. S., De Cubas, A. A., Bixby, L. M., Lee, S. J., et al. (2018). Endogenous Retroviral Signatures Predict Immunotherapy Response in clear Cell Renal Cell Carcinoma. *J. Clin. Invest.* 128, 4804–4820. doi:10.1172/jci121476

Tang, X., Liu, T., Zang, X., Liu, H., Wang, D., Chen, H., et al. (2013). Adoptive Cellular Immunotherapy in Metastatic Renal Cell Carcinoma: A Systematic Review and Meta-Analysis. *Plos One* 8, e62847. doi:10.1371/journal.pone.0062847

Tito, C., De Falco, E., Rosa, P., Iaiza, A., Fazi, F., Petrozza, V., et al. (2021). Circulating microRNAs from the Molecular Mechanisms to Clinical Biomarkers: A Focus on the Clear Cell Renal Cell Carcinoma. *Genes* 12, 1154. doi:10.3390/genes12081154

Travelli, C., Colombo, G., Mola, S., Genazzani, A. A., and Porta, C. (2018). NAMPT: A Pleiotropic Modulator of Monocytes and Macrophages. *Pharmacol. Res.* 135, 25–36. doi:10.1016/j.phrs.2018.06.022

Vuong, L., Kotecha, R. R., Voss, M. H., and Hakimi, A. A. (2019). Tumor Microenvironment Dynamics in Clear-Cell Renal Cell Carcinoma. *Cancer Discov.* 9, 1349–1357. doi:10.1158/2159-8290.Cd-19-0499

Wang, J., Chen, M., Dang, C., Zhang, H., Wang, X., Yin, J., et al. (2021). The Early Diagnostic and Prognostic Value of BIRC5 in Clear-Cell Renal Cell Carcinoma Based on the Cancer Genome Atlas Data. *Urol. Int.*, 1–8. [Epub ahead of print]. doi:10.1159/000517310

Zhang, H., Fang, D.-C., Wang, R.-Q., Yang, S.-M., Liu, H.-F., and Luo, Y.-H. (2004). Effect ofHelicobacter Pyloriinfection on Expressions of Bcl-2 Family Members in Gastric Adenocarcinoma. *Wjg* 10, 227–230. doi:10.3748/wjg.v10.i2.227

Zhang, H., Zhang, N., Liu, Y., Su, P., Liang, Y., Li, Y., et al. (2019). Epigenetic Regulation of NAMPT by NAMPT-AS Drives Metastatic Progression in Triple-Negative Breast Cancer. *Cancer Res.* 79, 3347–3359. doi:10.1158/0008-5472.Can-18-3418

Zhang, J., Wu, T., Simon, J., Takada, M., Saito, R., Fan, C., et al. (2018). VHL Substrate Transcription Factor ZHX2 as an Oncogenic Driver in clear Cell Renal Cell Carcinoma. *Science* 361, 290–295. doi:10.1126/science.aap8411

Zhang, J., Yan, A., Cao, W., Shi, H., Cao, K., and Liu, X. (2020). Development and Validation of a VHL-Associated Immune Prognostic Signature for clear Cell Renal Cell Carcinoma. *Cancer Cel Int* 20, 584. doi:10.1186/s12935-020-01670-5

Zhang, L., Tao, H., Li, J., Zhang, E., Liang, H., and Zhang, B. (2021). Comprehensive Analysis of the Competing Endogenous circRNA-lncRNA-miRNA-mRNA Network and Identification of a Novel Potential Biomarker for Hepatocellular Carcinoma. *Aging* 13, 15990–16008. doi:10.18632/aging.203056

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: 11 January 2022
doi: 10.3389/fmolb.2021.789703

Pan-Cancer Integrated Analysis of HSF2 Expression, Prognostic Value and Potential Implications for Cancer Immunity

Fei Chen[1†], *Yumei Fan*[1†], *Xiaopeng Liu*[1,2†], *Jianhua Zhang*[1†], *Yanan Shang*[1], *Bo Zhang*[1], *Bing Liu*[1], *Jiajie Hou*[1], *Pengxiu Cao*[1] *and Ke Tan*[1]*

[1]*Ministry of Education Key Laboratory of Molecular and Cellular Biology, Key Laboratory of Animal Physiology, Biochemistry and Molecular Biology of Hebei Province, College of Life Sciences, Hebei Normal University, Shijiazhuang, China,* [2]*Department of Neurosurgery, The Second Hospital of Hebei Medical University, Shijiazhuang, China*

OPEN ACCESS

Edited by:
Liangcan He,
Harbin Institute of Technology, China

Reviewed by:
Shanshan Yang,
Harbin Medical University Cancer Hospital, China
Qunying Li,
Zhejiang University, China

***Correspondence:**
Ke Tan
tanke@hebtu.edu.cn

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics,
a section of the journal Frontiers in Molecular Biosciences

Received: *05 October 2021*
Accepted: *20 December 2021*
Published: *11 January 2022*

Citation:
Chen F, Fan Y, Liu X, Zhang J, Shang Y, Zhang B, Liu B, Hou J, Cao P and Tan K (2022) Pan-Cancer Integrated Analysis of HSF2 Expression, Prognostic Value and Potential Implications for Cancer Immunity.
Front. Mol. Biosci. 8:789703.
doi: 10.3389/fmolb.2021.789703

Heat shock factor 2 (HSF2), a transcription factor, plays significant roles in corticogenesis and spermatogenesis by regulating various target genes and signaling pathways. However, its expression, clinical significance and correlation with tumor-infiltrating immune cells across cancers have rarely been explored. In the present study, we comprehensively investigated the expression dysregulation and prognostic significance of HSF2, and the relationship with clinicopathological parameters and immune infiltration across cancers. The mRNA expression status of HSF2 was analyzed by TCGA, GTEx, and CCLE. Kaplan-Meier analysis and Cox regression were applied to explore the prognostic significance of HSF2 in different cancers. The relationship between HSF2 expression and DNA methylation, immune infiltration of different immune cells, immune checkpoints, tumor mutation burden (TMB), and microsatellite instability (MSI) were analyzed using data directly from the TCGA database. HSF2 expression was dysregulated in the human pan-cancer dataset. High expression of HSF2 was associated with poor overall survival (OS) in BRCA, KIRP, LIHC, and MESO but correlated with favorable OS in LAML, KIRC, and PAAD. The results of Cox regression and nomogram analyses revealed that HSF2 was an independent factor for KIRP, ACC, and LIHC prognosis. GO, KEGG, and GSEA results indicated that HSF2 was involved in various oncogenesis- and immunity-related signaling pathways. HSF2 expression was associated with TMB in 9 cancer types and associated with MSI in 5 cancer types, while there was a correlation between HSF2 expression and DNA methylation in 27 types of cancer. Additionally, HSF2 expression was correlated with immune cell infiltration, immune checkpoint genes, and the tumor immune microenvironment in various cancers, indicating that HSF2 could be a potential therapeutic target for immunotherapy. Our findings revealed the important roles of HSF2 across different cancer types.

Keywords: HSF2, pan-cancer, prognosis, immune infiltration, immune checkpoint genes, multi-omics

INTRODUCTION

The incidence and mortality of cancer are increasing rapidly every year worldwide, posing a serious threat to public health (Sung et al., 2021). Among the most common cancers, breast, lung, and liver are the main causes of high mortality worldwide. Although scientists have made considerable efforts to improve the diagnosis and treatment of cancer, the 5-years survival rate for cancer patients remains disappointing (Ferlay et al., 2021; Sung et al., 2021). Concurrently, the economic burden of cancer on countries worldwide is gradually increasing (Ferlay et al., 2021; Sung et al., 2021). Therefore, there is an urgent need to find diagnostic biomarkers and new treatments for cancer.

Cancer cells face multiple internal and external stresses that are distinct from those faced by normal cells (Jeggo et al., 2016). These stimuli cause dysfunction of proteostasis as a result of protein misfolding, gene mutation, oncogene activation, inhibition of tumor suppressors, chromosomal rearrangement, oxidative stress, hypoxia, and impaired degradation of proteins (Hanahan and Weinberg, 2011; Jeggo et al., 2016). Upon exposure to these various stimulators, heat shock factor (HSF), which is the original regulator of the heat shock response (HSR), controls the rapid and dynamic expression of heat shock proteins (HSPs) (Akerfelt et al., 2010; Fujimoto and Nakai, 2010; Gomez-Pastor et al., 2018). HSPs, acting as molecular chaperones, are involved in various physiological and pathological processes, such as the folding and assembly of nascent polypeptides and the intracellular transport of proteins, and to exhibit cytoprotective effects. The HSF family contains five members, including HSF1, HSF2, HSF4, HSF5, and HSFY (Akerfelt et al., 2010; Fujimoto and Nakai, 2010; Gomez-Pastor et al., 2018). Numerous studies have demonstrated that HSF1 is associated with DNA damage repair, reprogrammed metabolism oncogenesis, and metastasis (Mendillo et al., 2012; Dai and Sampson, 2016; Dai, 2018; Wang G. et al., 2020; Puustinen and Sistonen, 2020). Thus, HSF1 is believed to be a potential therapeutic target for anticancer therapy (Zhang B. et al., 2021; Chen et al., 2021).

In contrast to HSF1, HSF2 has been shown to play an important role in mediating organ development, differentiation, and the ubiquitin proteasome pathway (Rallu et al., 1997; Mathew et al., 1998; Widlak and Vydra, 2017). HSF2 is necessary for embryogenesis and spermatogenesis as evidenced by knocking out the *HSF2* gene in mice (Sarge et al., 1994; Rallu et al., 1997; Björk et al., 2010). Apoptosis of spermatocytes is remarkably increased, and the maturation of male germ cells is impaired in *HSF2*-null mice (Sarge et al., 1994; Rallu et al., 1997; Björk et al., 2010). A recent study revealed that HSF2 promoted spermatogenesis by regulating the expression of HSP and Y chromosomal multicopy genes, including SLX, SLY, and SSTY2 (Akerfelt et al., 2008). HSF2 is also associated with brain development, as evidenced by HSF2-null mice exhibiting enlarged ventricles, a small hippocampus, and neurons mispositioning (Kallio et al., 2002; Wang et al., 2003; Chang et al., 2006). As a member of the HSF family, previous studies have suggested that HSF2 could form heterotrimers with HSF1 to promote the transcription of HSP and some other genes (Sistonen et al., 1994; Ostling et al., 2007; Sandqvist et al., 2009). However, the precise function and molecular mechanisms of HSF2 in tumorigenesis still need to be explored.

Although increasing evidence indicates that HSF2 may play a vital role in the tumorigenesis of some specific types of cancers, a systematic pan-cancer analysis of HSF2 has not yet been conducted. Therefore, the aims of this study were to explore the expression profile, prognostic value, methylation level of HSF2, and potential relationship between HSF2 expression and immunological functions in 33 different types of cancer.

MATERIALS AND METHODS

Heat Shock Factor 2 Expression in Pan-Cancer

The Cancer Genome Atlas (TCGA, https://www.cancer.gov/) database, which is widely used for comprehensive analyses of human cancers, was employed to investigate the differential expression of HSF2 across different cancer types. RNA sequencing data and clinical follow-up information for patients with 33 types of cancers were downloaded from the TCGA database. Because the normal tissues sequencing data included in the TCGA are very limited and many patients lack transcriptome sequencing results for their normal tissues, we obtained data for normal tissues from the Genotype-Tissue Expression (GTEx) database. The cell line expression matrix of HSF2 in pan-cancer was obtained from the CCLE dataset (https://portals.broadinstitute.org/ccle/about). The above analyses were constructed using the R (v4.0.3) software package ggplot2 (v3.3.3). R software v4.0.3 and ggplot2 (v3.3.3) were used for visualization. R software v4.0.3 was used for statistical analysis.

Heat Shock Factor 2 Expression and its Clinical Correlation in Pan-Cancer

The correlations of HSF2 expression with tumor stage and DNA methylation were investigated using the UALCAN database (http://ualcan.path.uab.edu/).

Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Set Enrichment Analysis (GSEA)

GO, KEGG, and GSEA were conducted to examine the biological and molecular functions of HSF2 across different cancer types using a total of 300 genes that were positively correlated with HSF2. GO analysis was applied to investigate the BP, CC, and MF associated with HSF2 in different cancers. All three analyses were performed using the R package Cluster Profiler.

cBioPortal Database

The genetic alterations of HSF2 in different cancer types were obtained using the cBioPortal database.

The Prognostic Potential of Heat Shock Factor 2 in Pan-Cancer

The survival data from 33 types of cancer were obtained from the TCGA database for further overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI) analyses. Univariate Cox regression analysis was used to analyze HSF2-related survival with the R package limma, survival, and forestplot to show the *p* value, HR, and 95% CI. The Kaplan-Meier (KM) method was used to investigate the prognostic value of HSF2 in human cancers using the R packages limma, survival, and survminer. R software v4.0.3 was used for statistical analysis.

Univariate and Multivariate Cox Regression Analyses and Construction of a Nomogram

Cox regression analysis, including univariate, and multivariate analyses, was used to examine the prognostic value of HSF2 in KIRP, ACC, and LIHC. The forest plot was constructed using the R package "forest plot" to exhibit the hazard ratio (HR), 95% CI, and *p*-value. The nomogram was constructed using the R package "rms".

Correlation of Heat Shock Factor 2 Expression With Tumor Cell Infiltration and Immune Modulator Genes in Pan-Cancer

We obtained the data for 33 types of human cancer in TCGA from the GDC data portal website. For reliable immune score evaluation, we used the R software package "Immuneeconv" to integrate the two latest algorithms, including TIMER, and xCell. Heatmaps of the immune infiltration scores or immune modulator genes and HSF2 expression in different cancer types were generated with Spearman correlation analysis. The horizontal axis in the heatmaps shows the type of cancer, the vertical axis shows different immune cell infiltration scores, and the color shows the correlation coefficients. Additionally, R software v4.0.3 was used for statistical analysis.

Relationships Between Heat Shock Factor 2 Expression and TMB or MSI in Pan-Cancer

We obtained the data for 33 types of human cancer in TCGA from the GDC data portal website. For pan-cancer analysis, the horizontal axis shows the correlation coefficient between HSF2 expression and TMB/MSI, the ordinate is the type of cancer, the size of the dots in the figure shows the degree of the correlation coefficient, and the different colors represent the significance of the *p* value. Correlation analysis between HSF2 and TMB/MSI was performed using Spearman's method and R software v4.0.3 was used for statistical analysis. A *p*-value less than 0.05 was considered statistically significant.

RESULTS

Heat Shock Factor 2 is Abnormally Expressed in Human Pan-Cancer

Based on the results from The Cancer Genome Atlas (TCGA) data alone, HSF2 expression was increased in CHOL, COAD, ESCA, HNSC, LIHC, LUSC, and STAD, but decreased in BRCA, KICH, KIRC, LUAD, PRAD, THCA, and UCEC tissues compared with adjacent normal tissues (**Figure 1A**). We also estimated HSF2 expression in paired cancer tissues and adjacent normal tissues in pan-cancer using TCGA datasets. HSF2 expression was significantly higher in CHOL, COAD, ESCA, HNSC, LIHC, and LUSC, but remarkably lower in BRCA, KICH, KIRC, PRAD, and THCA than in paired adjacent normal tissues (**Figure 1B**). Because several cancers lack corresponding normal tissue controls, we therefore combined the data from the TCGA, and Genotype Tissue-Expression (GTEx) (**Figure 1C**). After combining the data from TCGA and GTEx, the expression difference of HSF2 achieved significance in 25 out of 33 cancer types. HSF2 expression was higher in CHOL, DLBC, GBM, HNSC, LGG, LIHC, PAAD, and THYM but lower in ACC, BLCA, BRCA, CESC, COAD, KICH, KIRC, LAML, LUAD, OV, PRAD, READ, SKCM, TGCT, THCA, UCEC, and UCS (**Figure 1C**). Moreover, we also investigated the expression of HSF2 in different cancer cell lines according to the Cancer Cell Line Encyclopedia (CCLE) database (**Figure 1D**).

Association of Heat Shock Factor 2 Expression With Clinicopathological Features in Different Cancer Types

The relationship between HSF2 expression and the clinicopathological characteristics of patients with different cancers was investigated based on individual cancer stages, including stages 1, 2, 3, and 4. HSF2 expression was generally increased in CHOL, COAD, ESCA, KIRP, LIHC, LUSC, STAD, and UCS (**Figure 2**). In contrast, HSF2 expression was dramatically decreased in BRCA, KIRC, KICH, LUAD, SKCM, THCA, UCEC, and UVM (**Figure 2**). Moreover, HSF2 expression was stable in some cancers, including ACC, BLCA, CESC, DLBC, HNSC, MESO, OV, PAAD, READ, and TGCT (**Supplementary Figure S1**).

Prognostic Values of Heat Shock Factor 2 in Human Pan-Cancer

Next, we investigated the interrelationship between HSF2 expression and the prognosis of pan-cancer patients, including overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI). Regarding the OS analysis, Cox regression results from 33 types of cancer suggested that HSF2 expression was markedly related to OS in 6 types of cancer, including ACC, GBM, KICH, KIRP, LIHC, and PAAD (**Figure 3A**). The results from the Kaplan-Meier (KM) survival curves demonstrated that higher HSF2 expression was correlated with worse OS in BRCA, LIHC, KIRP, and MESO, but with better OS in LAML, KIRC, and PAAD (**Figure 3B**). Moreover, we explored the relationship between HSF2 expression and DSS in cancer patients. As shown in **Supplementary Figure S2A**, HSF2 expression was associated with poor DSS in three types of cancer, including KIRP, LIHC, and UCEC. KM of DSS analysis indicated that upregulated HSF2 expression corresponded with poor DSS in patients with KIRP, LIHC, and KICH but with favorable DSS in patients with KIRC (**Supplementary Figure S2B**). Moreover, Cox regression analysis of PFI demonstrated that upregulated HSF2 expression was a risk factor in ACC, KICH, KIRP, and LIHC and was a protective

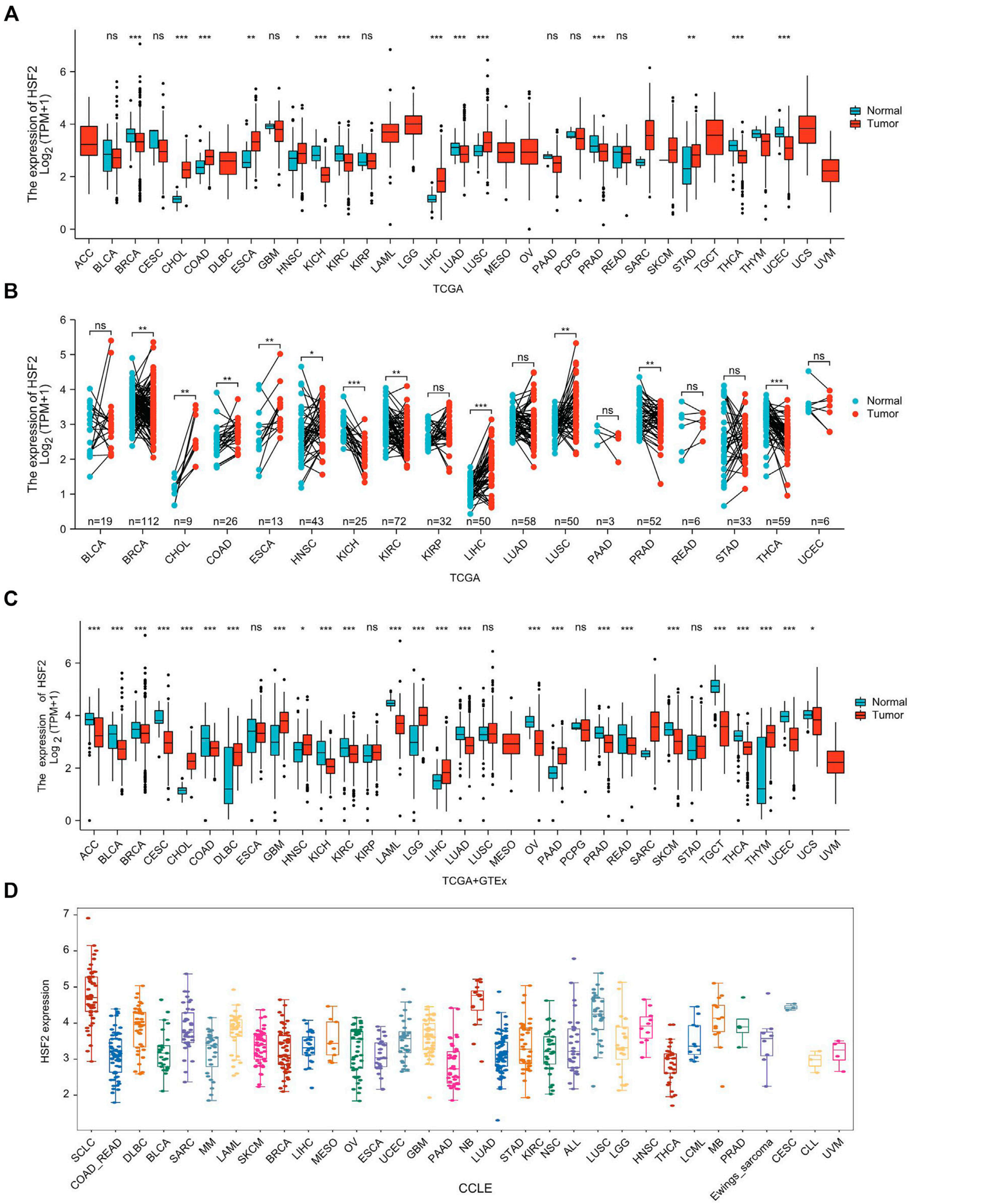


FIGURE 1 | HSF2 expression levels in pan-cancer. **(A)** Upregulated or downregulated expression of HSF2 in various human cancers from TCGA datasets. **(B)** Increased or decreased expression of HSF2 in paired cancer tissues and adjacent normal tissues from TCGA datasets. **(C)** HSF2 differential expression across different cancer types in the TCGA and GTEx databases. **(D)** The mRNA level of HSF2 in different cancer cells according to the CCLE database. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

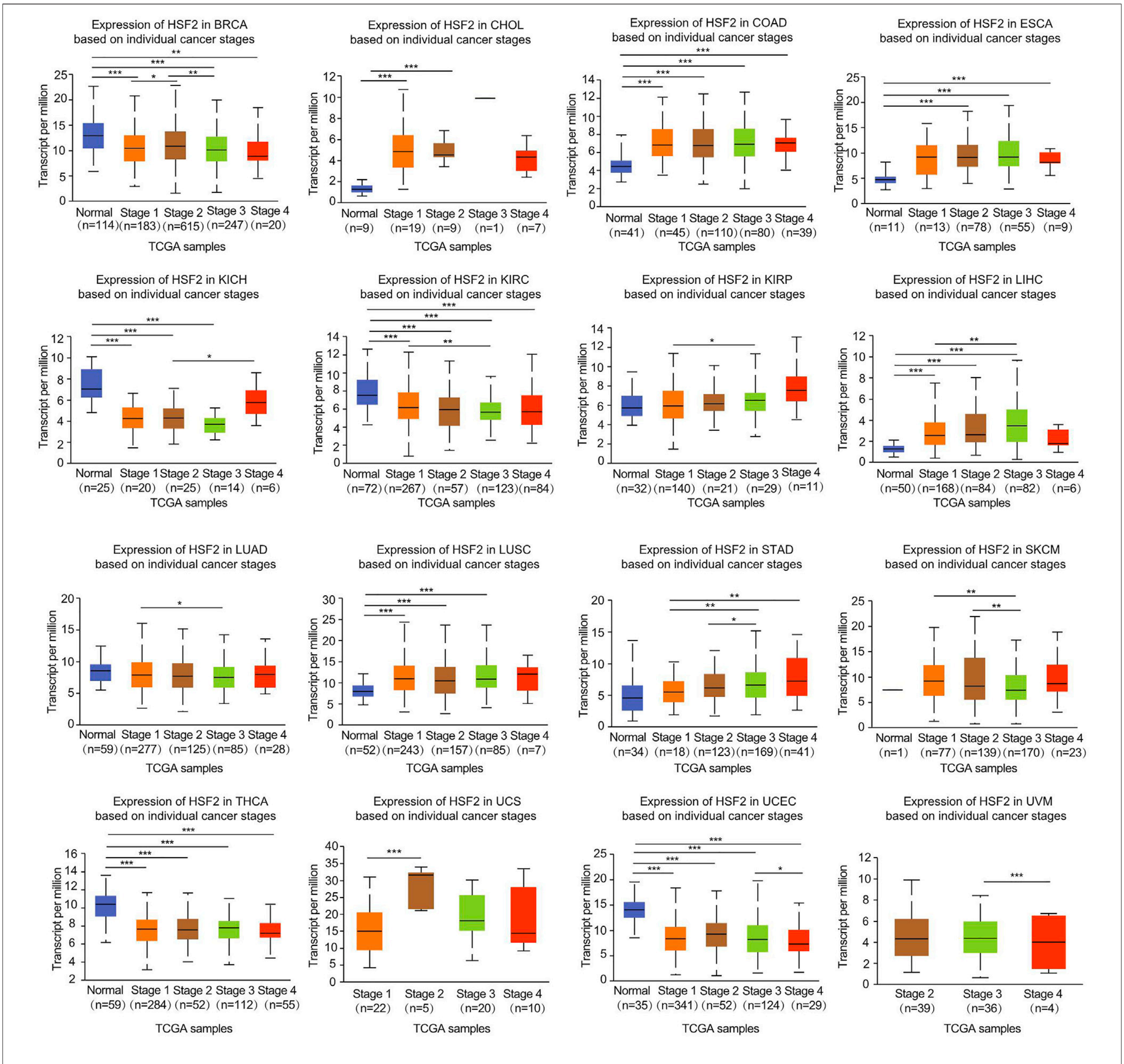


FIGURE 2 | Correlation of HSF2 expression and clinicopathological parameters across different cancer types. The clinical correlations between HSF2 expression levels and tumor stage in different cancer types were examined using the UALCAN database. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

factor in PAAD (**Supplementary Figure S3A**). Results from the KM of PFI analysis suggested that increased HSF2 expression was associated with a poor PFI in ACC and LIHC but with a favorable PFI in CHOL, KIRC, and LGG (**Supplementary Figure S3B**). Subsequently, we also assessed the association between HSF2 expression and DFI and identified that dysregulated HSF2 expression influenced DFI in patients with KIRP, LIHC, and UCEC (**Supplementary Figure S4A**). KM DFI analysis revealed that increased HSF2 mRNA expression was correlated with an unfavorable DFI in BLCA, CESC, and LIHC (**Supplementary Figure S4B**).

Heat Shock Factor 2 is an Independent Prognostic Factor in KIPR, ACC, and LIHC

To further confirm whether HSF2 was an independent prognostic factor in cancers, univariate and multivariate Cox regression analyses were performed based on various clinicopathological characteristics, such as age, T stage, N stage, M stage, TNM stage, and grade. Univariate Cox regression analysis demonstrated that HSF2 expression ($p < 0.001$), T stage ($p < 0.001$), N stage ($p < 0.001$), M stage ($p < 0.001$), and TNM stage ($p < 0.001$) were significantly correlated with OS in KIPR (**Figure 4A**); HSF2 expression ($p < 0.05$), T stage ($p < 0.001$), M stage ($p <$

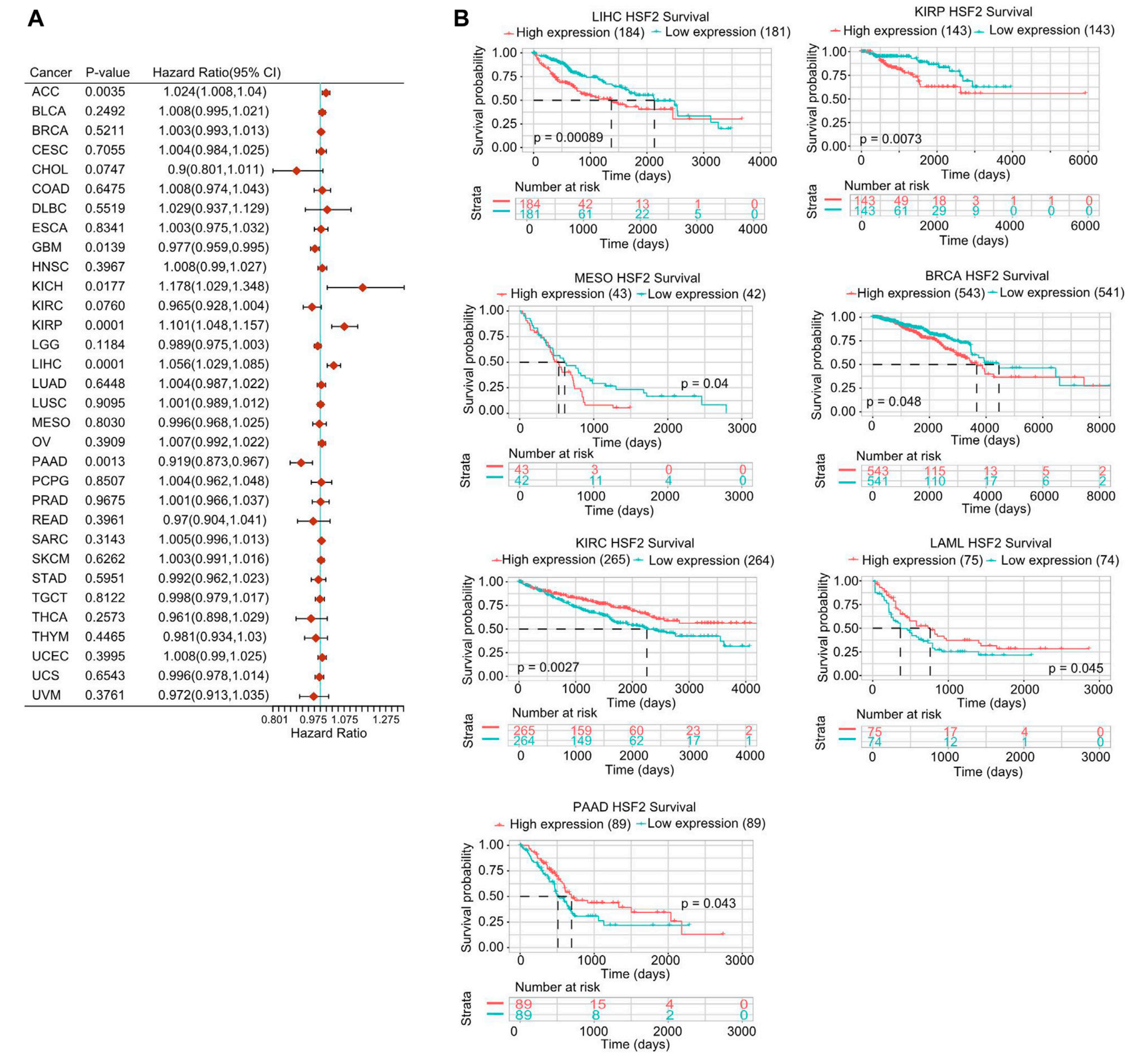
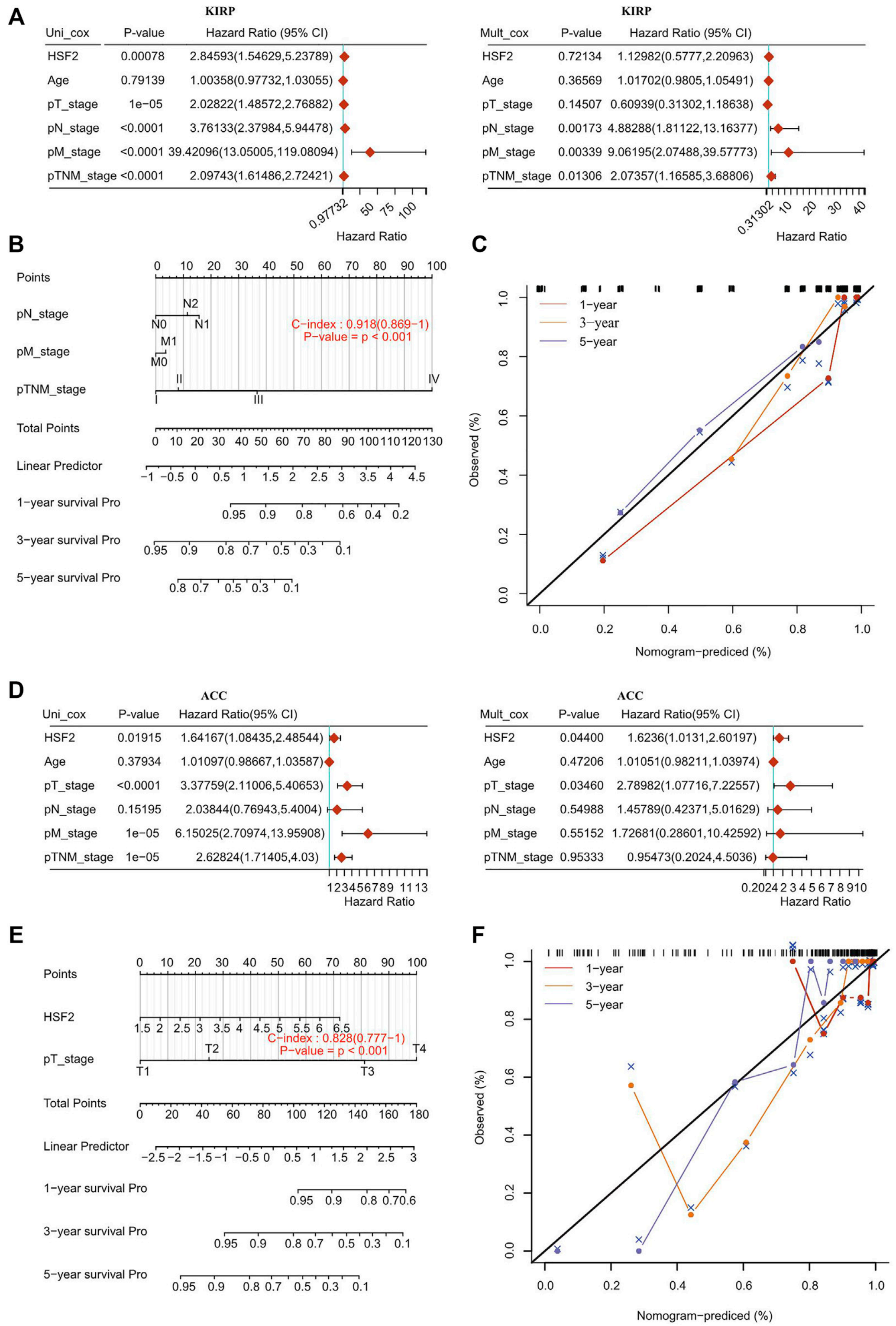


FIGURE 3 | Prognostic potential of HSF2 in pan-cancer. **(A)** Correlation analysis of HSF2 expression with OS by the Cox regression model in various cancers. **(B)** OS curves comparing high and low expression of HSF2 in multiple cancer types using Kaplan-Meier methodology.

0.001), and TNM stage ($p < 0.001$) were obviously correlated with OS in ACC (**Figure 4D**); HSF2 expression ($p < 0.001$), M stage ($p < 0.05$) and TNM stage ($p < 0.001$) were strongly correlated with OS in LIHC (**Supplementary Figure S5A**). Multivariate analysis indicated that N stage ($p < 0.01$), M stage ($p < 0.01$), and TNM stage ($p < 0.05$) were significantly correlated with OS in KIPR (**Figure 4A**); HSF2 expression ($p < 0.05$) and T stage ($p < 0.05$) were obviously correlated with OS in ACC (**Figure 4D**); HSF2 expression ($p < 0.001$) and TNM stage ($p < 0.001$) were markedly correlated with OS in LIHC (**Supplementary Figure S5A**). In addition, a nomogram was constructed based on multivariate analysis (**Figures 4B,E**; **Supplementary Figure S5B**). The C-index and calibration curve confirmed the accuracy in predicting the 1-, 3-, and 5-years survival rates of cancer patients. The C-index of the prognostic nomogram was 0.918, 0.828, and 0.696 in KIPR, ACC, and LIHC, respectively (**Figures 4C,F**; **Supplementary Figure S5C**).

DNA Methylation and Genetic Alteration Analysis of Heat Shock Factor 2 in Pan-Cancer

A growing body of evidence suggests that DNA methylation is an epigenetic molecular mechanism for gene expression and that DNA



Uni_cox	P-value	Hazard Ratio (95% CI)
HSF2	0.00078	2.84593(1.54629,5.23789)
Age	0.79139	1.00358(0.97732,1.03055)
pT_stage	1e−05	2.02822(1.48572,2.76882)
pN_stage	<0.0001	3.76133(2.37984,5.94478)
pM_stage	<0.0001	39.42096(13.05005,119.08094)
pTNM_stage	<0.0001	2.09743(1.61486,2.72421)

Mult_cox	P-value	Hazard Ratio (95% CI)
HSF2	0.72134	1.12982(0.5777,2.20963)
Age	0.36569	1.01702(0.9805,1.05491)
pT_stage	0.14507	0.60939(0.31302,1.18638)
pN_stage	0.00173	4.88288(1.81122,13.16377)
pM_stage	0.00339	9.06195(2.07488,39.57773)
pTNM_stage	0.01306	2.07357(1.16585,3.68806)

Uni_cox	P-value	Hazard Ratio(95% CI)
HSF2	0.01915	1.64167(1.08435,2.48544)
Age	0.37934	1.01097(0.98667,1.03587)
pT_stage	<0.0001	3.37759(2.11006,5.40653)
pN_stage	0.15195	2.03844(0.76943,5.4004)
pM_stage	1e−05	6.15025(2.70974,13.95908)
pTNM_stage	1e−05	2.62824(1.71405,4.03)

Mult_cox	P-value	Hazard Ratio(95% CI)
HSF2	0.04400	1.6236(1.0131,2.60197)
Age	0.47206	1.01051(0.98211,1.03974)
pT_stage	0.03460	2.78982(1.07716,7.22557)
pN_stage	0.54988	1.45789(0.42371,5.01629)
pM_stage	0.55152	1.72681(0.28601,10.42592)
pTNM_stage	0.95333	0.95473(0.2024,4.5036)

FIGURE 4 | Internal validation of HSF2 as an independent prognostic factor for KIRP patients and ACC patients. **(A,D)** Univariate and multivariate Cox regression analyses were performed to determine HSF2 as an independent prognostic factor. **(B,E)** A prognostic nomogram integrating HSF2 expression and clinicopathologic variables was constructed to estimate OS. **(C,F)** Calibration plots to predict the OS of KIRP and ACC at 1, 3, and 5 years.

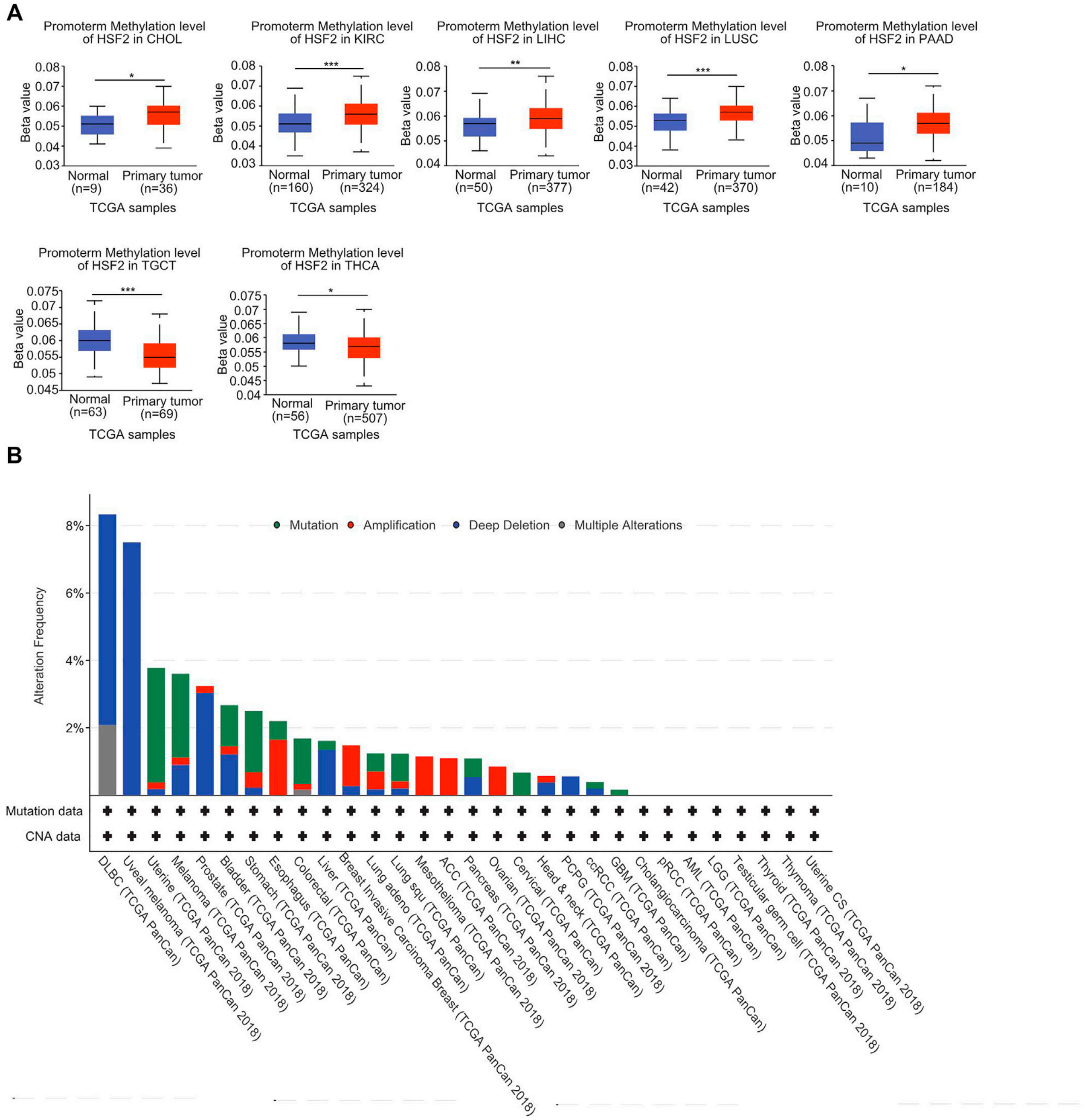


FIGURE 5 | DNA methylation and mutation profile of HSF2 in pan-cancer. **(A)** The promoter methylation level of HSF2 in across different cancer types was investigated according to the UALCAN database. **(B)** The alteration frequency of HSF2 with different mutation types was obtained from the cBioPortal database. *p < 0.05; **p < 0.01; ***p < 0.001.

hypermethylation leads to the inactivation of a broad range of tumor suppressor genes. Therefore, we investigated the potential link between DNA methylation and HSF2 expression. With respect to the TCGA database, we observed that the DNA methylation level of HSF2 was obviously increased in CHOL, KIRC, LIHC, LUSC, and PAAD but decreased in TGCT and THCA based on the UALCAN database (**Figure 5A**). Moreover, we observed that DNA methylation was negatively correlated with HSF2 expression in many types of cancer, including ACC, BLCA, BRCA, CESC, CHOL, DLBC, ESCA, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, PAAD, PCPG, PRAD, SARC, SKCM, TGCT, THYM, UCEC, USC, and UVM (**Supplementary Figure S6**). In contrast, DNA methylation was positively associated with HSF2 expression in OV (**Supplementary Figure S6**).

In addition, we investigated the alteration frequency of HSF2 in different cancer types according to the cBioPortal database.

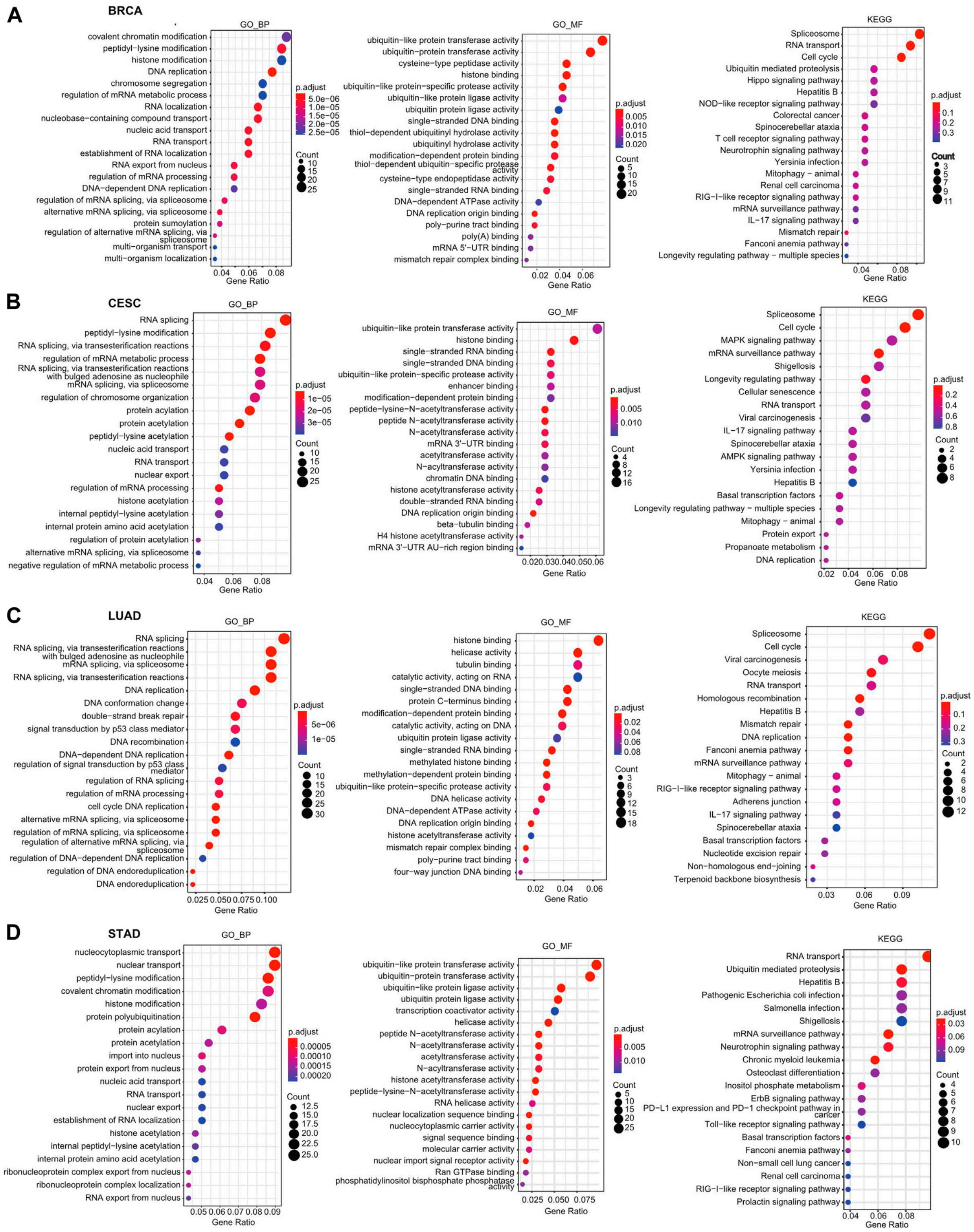


FIGURE 6 | GO and KEGG enrichment analyses for HSF2 in cancers. Top 20 pathways enriched in the BP, MF, and KEGG analyses in **(A)** BRCA, **(B)** CESC, **(C)** LUAD, and **(D)** STAD.

The highest incidence rate of genetic variations of HSF2 was observed in DLBC, and deep depletion was the primary type (**Figure 5B**).

GO and KEGG Analyses of Heat Shock Factor 2 in Pan-Cancer

First, we identified genes with positive or negative coexpression with HSF2 using the TCGA database (**Supplementary Table S1**), and the top 50 genes that were positively and negatively associated with HSF2 in different cancers are shown (**Supplementary Figure S7**). To explore the molecular mechanisms by which HSF2 regulates oncogenesis, we performed GO and KEGG analyses using the 300 genes that were positively related to HSF2 in several cancers (**Figure 6**). The top 5 enriched BP GO terms were covalent chromatin modification, peptidyl-lysine modification, histone modification, DNA replication, and chromosome segregation in BRCA; RNA splicing, peptidyl-lysine modification, regulation of mRNA metabolic process, regulation of chromosome organization, and protein acylation in CESC; RNA splicing, mRNA splicing, DNA replication, DNA conformation change, and double-strand break repair in LUAD; and nucleocytoplasmic transport, nuclear transport, peptidyl-lysine modification, covalent chromatin modification, and histone modification in STAD (**Figures 6A–D**). The top 3 enriched MF terms of GO were ubiquitin-like protein transferase activity, ubiquitin-protein transferase activity, and cysteine-type peptidase activity in BRCA; ubiquitin-like protein transferase activity, histone binding, and single-stranded DNA binding in CESC; ubiquitin-like protein transferase activity, ubiquitin-protein transferase activity, and ubiquitin-like protein ligase activity in LUAD; and histone binding, helicase activity, and tubulin binding in STAD (**Figures 6A–D**). The top 3 enriched CC terms of GO were nuclear speck, chromosomal region, and nuclear envelope in BRCA; nuclear chromatin, nuclear speck, and chromosomal region in CESC; chromosomal region, nuclear speck, and spindle in LUAD; and nuclear speck, nuclear envelope, and chromosomal region in STAD (**Supplementary Figures S8A–D**).

Moreover, KEGG pathway analysis suggested that HSF2 was associated with signaling pathways related to the spliceosome, RNA transport, cell cycle, ubiquitin-mediated proteolysis, and Hippo signaling pathway in BRCA; spliceosome, cell cycle, MAPK signaling pathway, mRNA surveillance pathway, and shigellosis in CESC; spliceosome, cell cycle, viral carcinogenesis, oocyte meiosis, and RNA transport in LUAD; and RNA transport, ubiquitin-mediated proteolysis, hepatitis B infection, pathogenic *Escherichia coli* infection, and Salmonella infection in STAD (**Figures 6A–D**).

Heat Shock Factor 2-Related Signaling Pathways in Cancers Identified by GSEA

GSEA was further performed to explore the signaling pathways and molecular mechanisms that were differentially affected by HSF2 in human cancers. Regarding the GO terms, the top 3 pathways influenced by HSF2 were the histone acetyltransferase complex, alternative RNA spicing via the spliceosome, and mitotic sister chromatid segregation in BRCA, CESC, LUAD, and STAD (**Figures 7A–D**). Among the KEGG terms, the top 3 pathways affected by HSF2 were ubiquitin-mediated proteolysis, herpes simplex virus 1 infection, and the mRNA surveillance pathway in BRCA; basal transcription factor, inositol phosphate metabolism, and cell cycle in CESC; DNA replication, mismatch repair and cell cycle in LUAD; and ubiquitin-mediated proteolysis, RNA degradation, and spliceosome in STAD (**Figures 7A–D**). More importantly, regarding the Reactome terms, the outcome of GSEA indicated that in addition to the cell response to stress, different immunity-related pathways were associated with HSF2, including the adaptive immune response, TRIF (TICAM1)-mediated TLR4 signaling, MyD88-dependent TLR4 cascade, TLR3 cascade, TLR4 cascade, and various bacterial or viral infections. Taken together, these findings imply that there is a close relationship among HSF2, the inflammatory response, and the tumor microenvironment (TME) (**Figures 7A–D**).

Association of Heat Shock Factor 2 Expression and Immune Cell Infiltration in Pan-Cancer

Because immune-infiltrating cells play an important role in cancer initiation and development, we then estimated the association between HSF2 expression and the infiltration levels of six major immune cells in 32 types of cancers. Using the data obtained from the TIMER database, the correlation between HSF2 expression, and the infiltration levels of these immune cells was investigated separately. The results implied that HSF2 expression was markedly correlated with the infiltrating level of B cells in 16 types of cancer, $CD4^+$ T cells in 12 types of cancer, $CD8^+$ T cells in 16 types of cancer, macrophages in 17 types of cancer, neutrophils in 18 types of cancer, and DCs in 18 types of cancer (**Figure 8A**). In addition, HSF2 positively correlated with these six types of immune cells in KIRC, LIHC, PAAD, and PRAD but negatively correlated with these immune cells in SARC (**Figure 8A**). To further confirm the relationship between HSF2 expression and infiltration of 38 subtypes of immune cell subtypes, we utilized the xCell database. HSF2 expression was negatively related to the infiltration levels of most immune cells in LUSC, SARC, and UCEC (**Figure 8B**).

Relationships Between Heat Shock Factor 2 Expression and Immune Checkpoint Genes, Chemokines, Immunostimulators, and MHC-Related Genes in Pan-Cancer

Because immune checkpoint genes play an important role in tumor immunotherapy, the correlations between HSF2 and immune checkpoint genes, immunoinhibitors, and immunostimulators were subsequently analyzed. Notably, we observed that HSF2 was significantly correlated with most immune checkpoint genes, including PD-1, PD-L1,

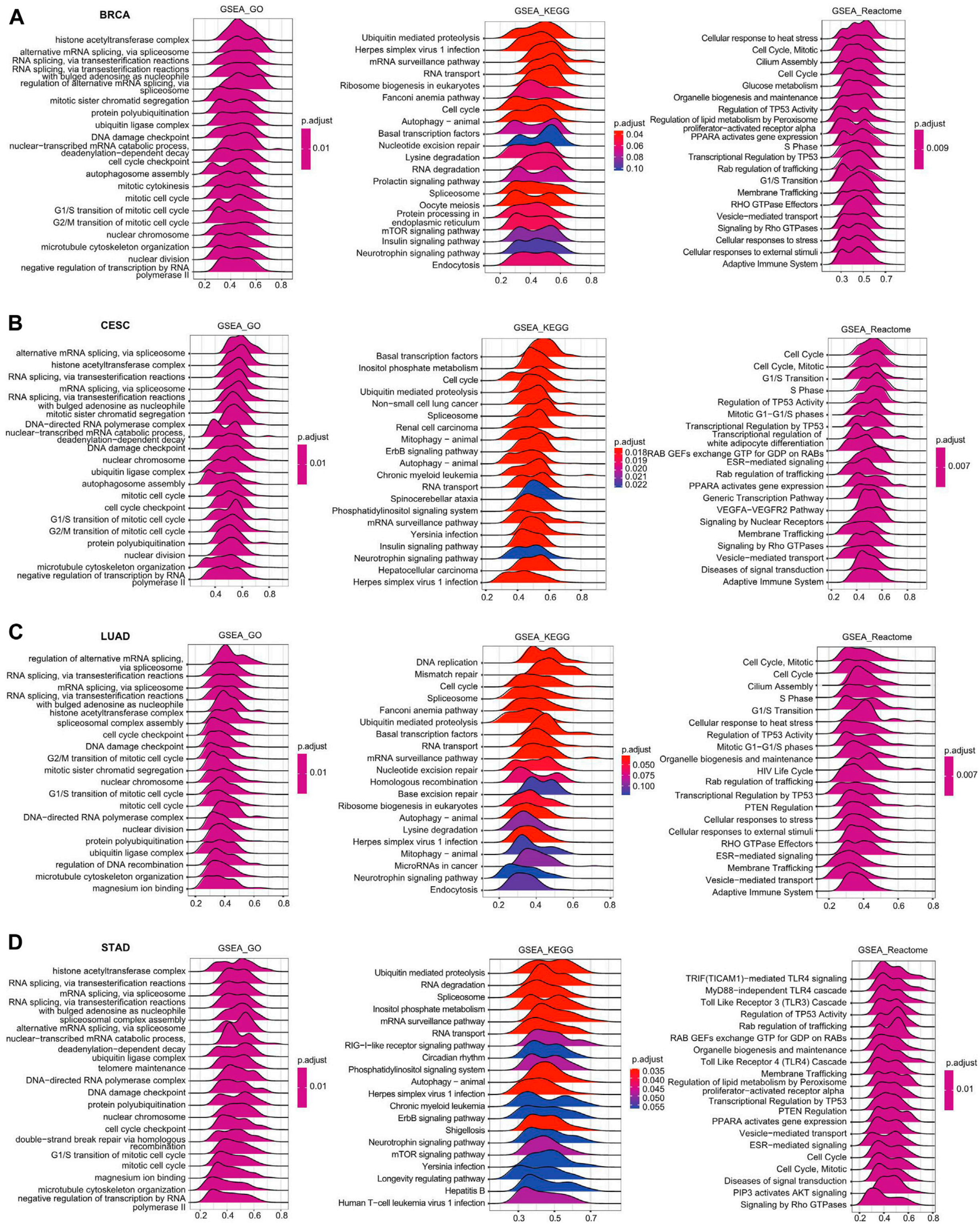


FIGURE 7 | Merged enrichment plots for HSF2 according to GSEA in cancers. Merged plots of GSEA indicating the signaling pathways correlated with HSF2 based on the GO, KEGG, and Reactome analyses in **(A)** BRCA, **(B)** CESC, **(C)** LUAD, and **(D)** STAD.

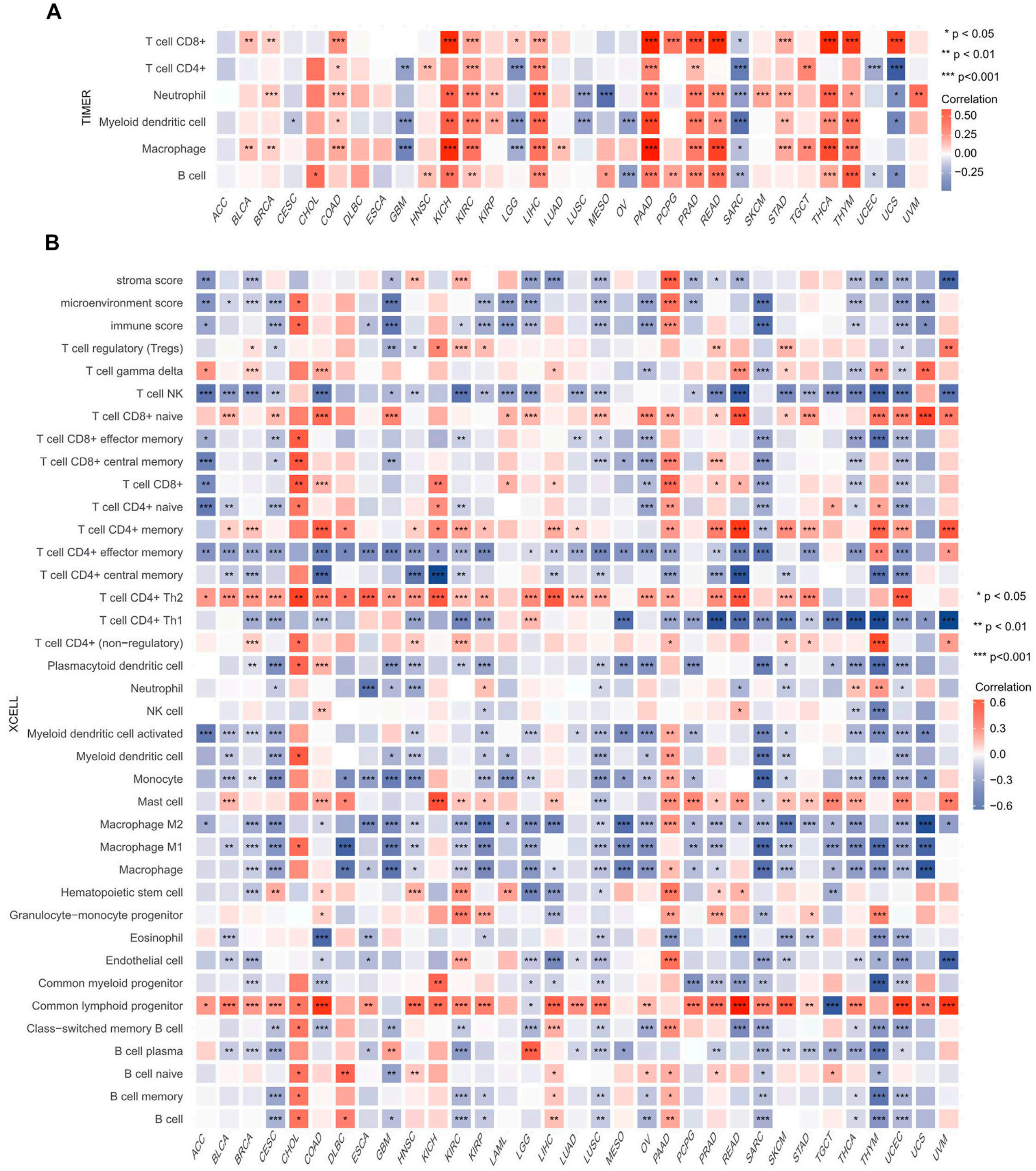


FIGURE 8 | Correlations of HSF2 expression with the infiltration level of immune cells across different cancer types. **(A)** Heatmap of correlations between the expression of HSF2 and the level of immune infiltration in 32 types of human cancer using TIMER. **(B)** Heatmap of correlations between the expression of HSF2 and the level of immune infiltration in 33 types of human cancer using xCell. **p* < 0.05; ***p* < 0.01; ****p* < 0.001.

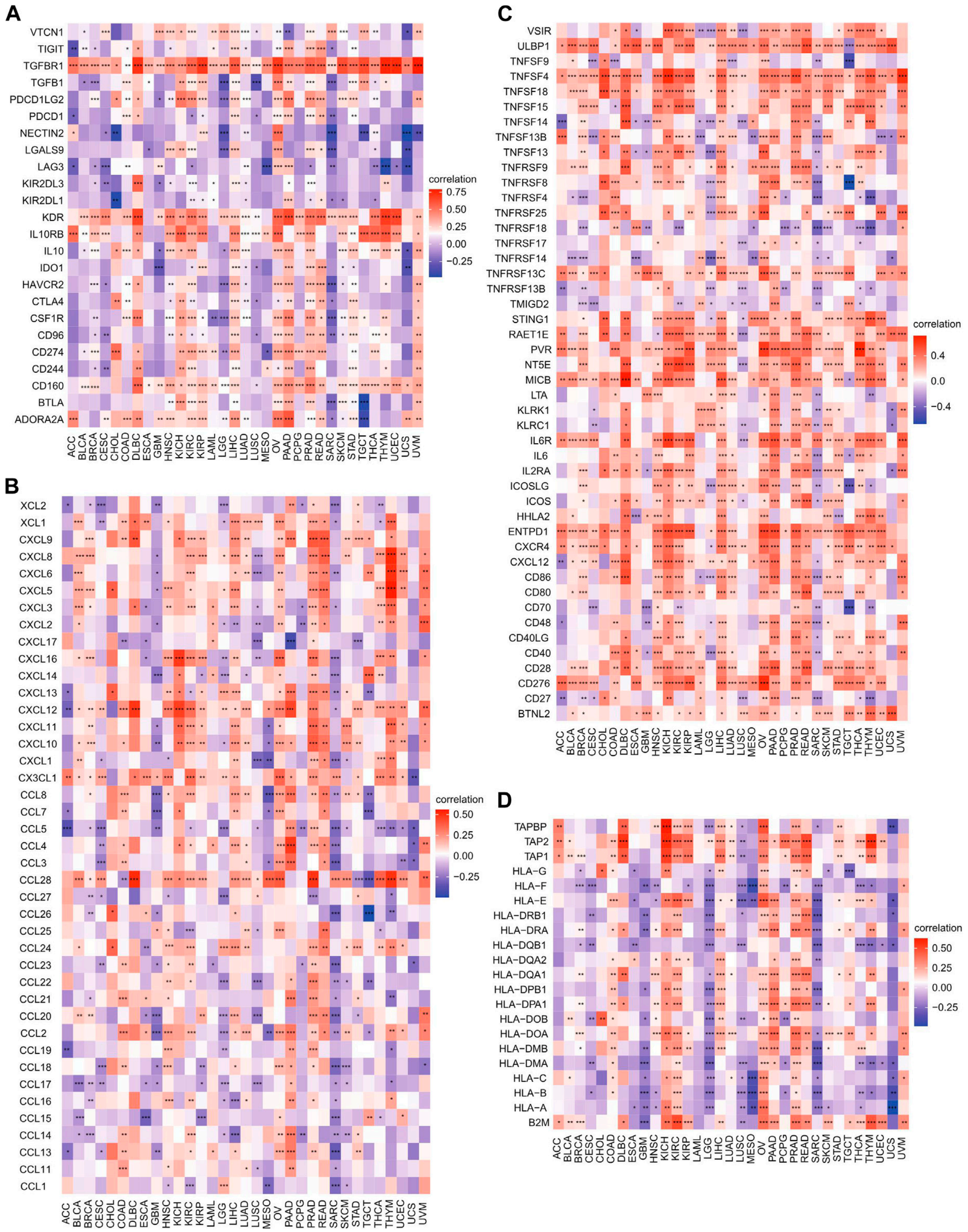


FIGURE 9 | Correlations of HSF2 expression with immune checkpoint genes, chemokines, immunostimulators, and MHC-related genes across different cancer types. **(A)** Heatmap of correlations between HSF2 expression and immune checkpoint genes. **(B)** Heatmap of correlation between HSF2 expression and chemokines. **(C)** Heatmap of the correlation between HSF2 expression and immunostimulators. **(D)** Heatmap of the correlation between HSF2 expression and MHC-related genes. **p* < 0.05; ***p* < 0.01; ****p* < 0.001.

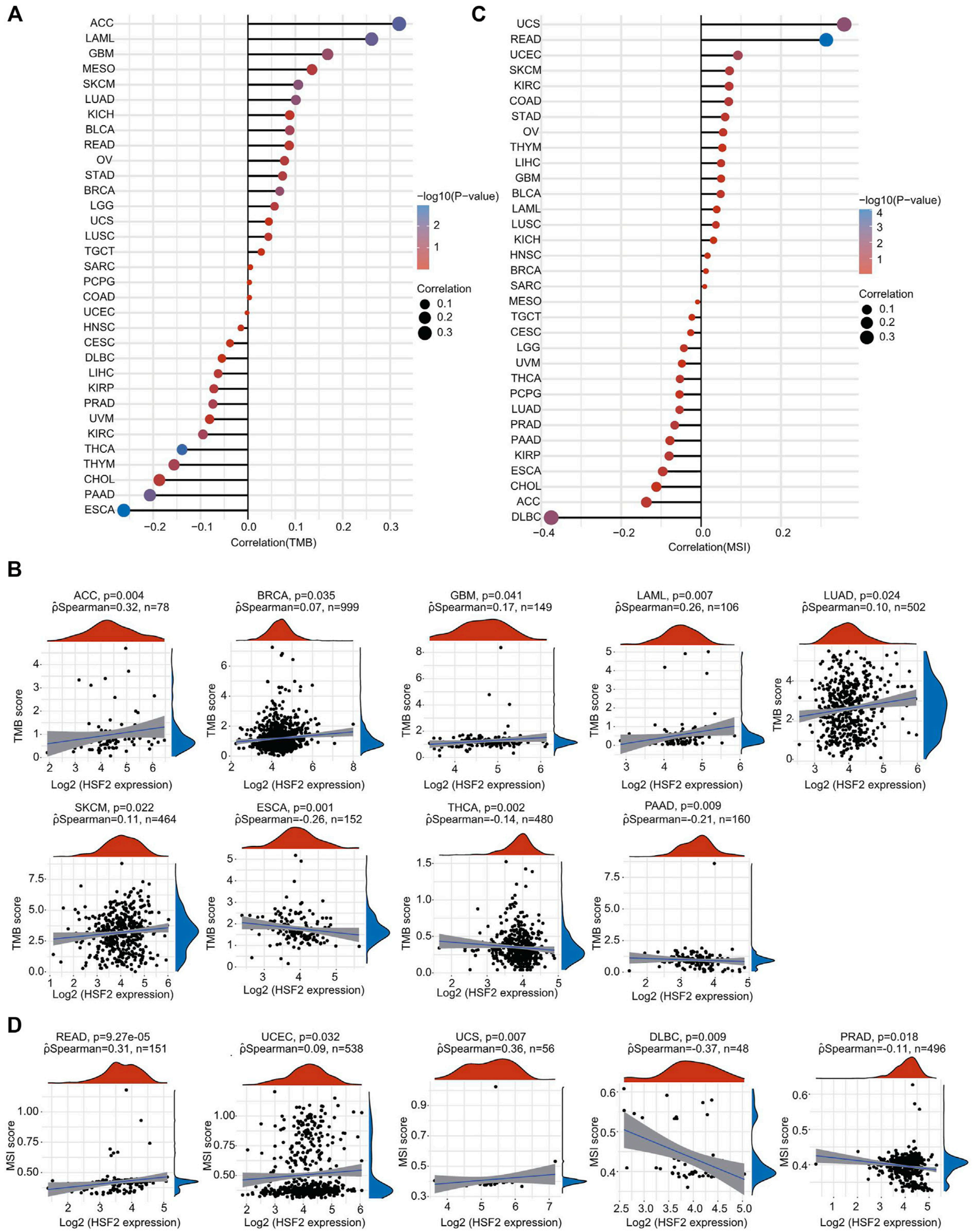


FIGURE 10 | Correlations of HSF2 expression and TMB and MSI in pan-cancer. **(A)** The stick chart shows the associations between HSF2 expression and TMB in pan-cancer. **(B)** Relationship between HSF2 expression and TMB in 9 tumors types. **(C)** The stick chart shows the associations between HSF2 expression and MSI in pan-cancer. **(D)** Relationship between HSF2 expression and MSI in 5 tumors types. Correlation analysis was performed using Spearman's method.

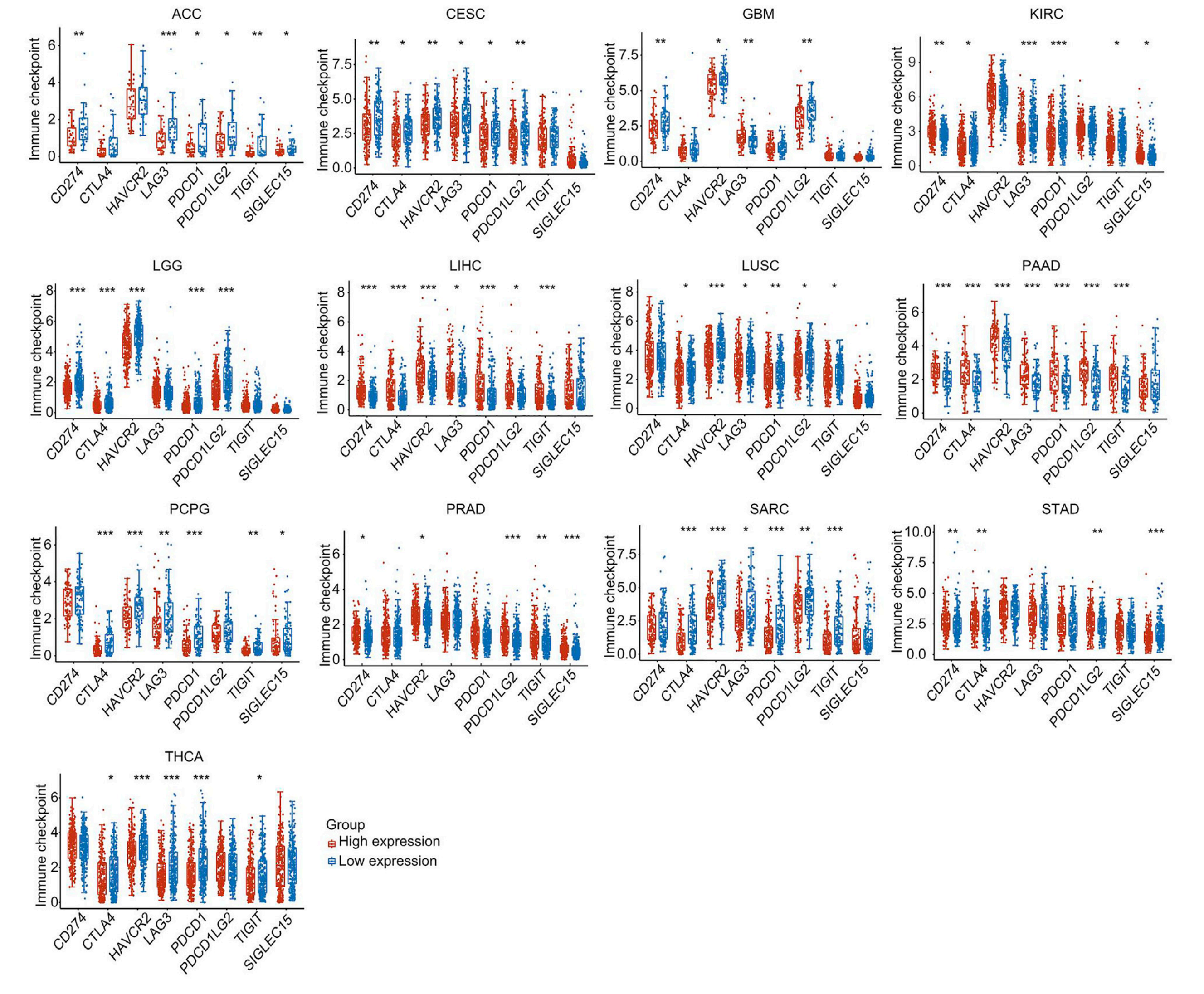


FIGURE 11 | The expression of various immune checkpoint genes between the HSF2 low-expression group and the high-expression group in different cancer types. *p < 0.05, **p < 0.01, ***p < 0.001.

CTLA4, KDR, TGFBR1, and IL10RB, in OV, PAAD, PRAD, and LIHC (**Figure 9A**). Interestingly, HSF2 expression was positively correlated with the expression of different chemokines in PAAD and PRAD but negatively associated with the expression of most chemokines in SRAC (**Figure 9B**). Moreover, we found that the expression of HSF2 was significantly and positively associated with immunostimulators in HNSC, KICH, KIRC, LIHC, OV, PAAD, and PRAD (**Figure 9C**). Additionally, HSF2 expression was positively correlated with most MHC-related genes in KIRC, OV, PAAD, and PRAD (**Figure 9D**). In contrast, HSF2 expression was negatively and strongly associated with most MHC-related genes in LGG and SARC (**Figure 9D**).

Relationships Between Heat Shock Factor 2 Expression and TMB and MSI in Pan-Cancer

TMB and MSI are two emerging biomarkers associated with the immunotherapy response. The relationships between HSF2 expression level and TMB across different cancer types were also investigated. The expression level of HSF2 was markedly and positively correlated with TMB in many cancers, including ACC, BRCA, GBM, LAML, LUAD, and SKCM, but negatively correlated with TMB in ESCA, THCA, and PAAD (**Figures 10A,B**). Additionally, the correlation of HSF2 expression with MSI was also explored in pan-cancer, among which READ, UCEC, and UCS exhibited a positive correlation while DLBC and PRAD exhibited a negative correlation with HSF2 expression (**Figures 10C,D**).

Effect of Heat Shock Factor 2 Expression on the Expression of Immune Checkpoints

Cancer patients were separated into high-expression and low-expression groups based on HSF2 expression. We then evaluated the effect of HSF2 expression on the expression of well-known immune checkpoints (CD274, CTLA4, HAVCR2, LAG3, PDCD1, PDCD1LG2, TIGHT, and SIGLEC15) to estimate the immunotherapy responses correlated with HSF2 expression. Significantly lower expression of most immune checkpoint genes was observed in the HSF2 high-expression group than in the HSF2 low-expression group in ACC, CESC, GBM, LGG, LUSC, PCPG, SARC, and THCA (**Figure 11**). In contrast, the expression of most immune checkpoint genes was higher in the HSF2 high-expression than low-expression group in LIHC, PAAD, PRAD, and STAD (**Figure 11**).

DISCUSSION

Cancer has become a serious threat to human health worldwide due to its high morbidity and mortality (Ferlay et al., 2021; Sung et al., 2021). Early detection and effective treatment are important prerequisites for improving the prognosis of cancer patients (Zhong et al., 2016; Fan et al., 2021; Liu et al., 2021). At present, the most common cancer treatments include surgical resection, radiation, and adjuvant chemotherapy, but the effectiveness is still limited. Therefore, it is urgent and necessary to identify novel tumor biomarkers and to understand their molecular mechanisms involved in tumorigenesis and progression for the development of more effective diagnostic methods and treatment strategies. HSF2, a transcription factor for the heat shock response, plays significant roles in corticogenesis and spermatogenesis by regulating various target genes, and signaling pathways (Rallu et al., 1997; Mathew et al., 1998; Widlak and Vydra, 2017). Nevertheless, HSF2 has not been largely studied in the cancer field, and its role in oncogenesis or pan-cancer is still unclear. In the present study, we employed an array of bioinformatics methods to explore the potential tumor-promoting or tumor-suppressing roles of HSF2 by investigating the significant correlation between HSF2 expression and the prognosis of cancer patients, DNA methylation, TMB, MSI, immune cell infiltration levels, and immune checkpoint genes in pan-cancer according to the results from the TCGA, GTEx, UALCAN, and cBioPortal databases.

Here, we conducted the first comprehensive systematic analysis of HSF2 across 33 cancer types. Our results showed that HSF2 was dysregulated in various human cancers, which was consistent with previous studies from other clinical and preclinical data (Mustafa et al., 2010; Li et al., 2014; Björk et al., 2016; Zhong et al., 2016; Meng et al., 2017; Yang et al., 2018; Yang et al., 2019). We investigated HSF2 expression in various types of cancers and their corresponding normal tissues according to the TCGA database and observed that HSF2 was differentially expressed in 14 types of cancer (**Figure 1**). When combining the data from TCGA and GTEx, HSF2 was dysregulated in up to 25 types of cancer (**Figure 1**). HSF2 has been reported to be expressed at high levels in patients with lung cancer and affects the growth and migration of lung cancer cells by regulating the expression of HSPs (Zhong et al., 2016). HSF2 is also dysregulated in breast cancer cells to modulate their proliferation and invasion (Li et al., 2014; Yang et al., 2018). In breast cancer cells, HSF2 has been identified to mediate transcription of the miR-183/-96/-182 cluster, which is highly expressed to promote tumorigenesis by directly regulating RAB21 expression (Li et al., 2014). Moreover, HSF2 mediates expression of the ALG3 enzyme, which subsequently promotes the growth and migration of breast cancer cells (Yang et al., 2018). ALG3 silencing significantly suppresses tumor growth and downregulates HSF2 expression, suggesting the presence of a feedback loop between these two genes (Yang et al., 2018). Additionally, previous studies have shown a higher level of HSF2 expression in HCC than in normal liver tissues (Yang et al., 2019). Mechanistically, HSF2 interacts with euchromatic histone lysine methyltransferase 2 (EHMT2) to suppress the expression of fructose-bisphosphatase 1 (FBP1) (Yang et al., 2019). Knockdown of FBP1 facilitates the HIF1 activation and upregulates the expression of glucose transporter 1 (GLUT1), lactate dehydrogenase A (LDHA), and hexokinase 2 (HK2) to increase aerobic glycolysis in HCC (Yang et al., 2019). These results reveal that HSF2 may act as an oncogene to promote the initiation and progression of HCC. In addition, in ESCC, miR-202 inhibits apoptotic cell death by directly targeting HSF2, which subsequently affects the expression of HSP70 (Meng et al., 2017). In contrast, HSF2 expression is clearly decreased in prostate cancer tissues (Björk et al., 2016). The reduced expression of HSF2 is associated with the metastasis of prostate cancer, indicating that HSF2 is a tumor suppressor in prostate cancer (Björk et al., 2016). Altogether, these previous studies suggest that HSF2 may function as an oncogenic or tumor-suppressing gene in different tumors.

In view of the pathological and clinical significance of HSF2 across different cancer types, we also investigated whether HSF2 could be used as a potential biomarker for the early diagnosis of human cancers. Therefore, we examined the relationship between HSF2 expression and OS, DSS, DFI, and PFI across different cancer types (**Figure 3**; **Supplementary Figures S2–4**). The results indicated that high expression of HSF2 was a risk factor and associated with poor OS, DSS, DFI, and PFI in some cancers but seemed to be protective in KIRC and PAAD. Moreover, a nomogram including HSF2 and clinicopathological characteristics was constructed and exhibited good predictive power for the OS of KIRP, ACC, and LIHC patients (**Figure 4**; **Supplementary Figure S5**). These observations, together with the clinicopathological features, illustrate that HSF2 is a newly identified multicancer-relevant gene with prognostic potential in cancer risk prediction and they support the possible effect of HSF2 on lymph node metastasis in COAD, ESCA, LIHC, LUSC, and STAD.

To further explore the molecular mechanism by which HSF2 affects oncogenesis, we performed KEGG and GSEA analyses. The results directly demonstrated the involvement of HSF2 in colorectal cancer, renal cell carcinoma, hepatocellular carcinoma, endometrial cancer, small cell lung cancer, chronic myeloid leukemia, and viral carcinogenesis (**Figures 6**, **7**). Moreover, HSF2 was found to be associated with ubiquitin-mediated proteolysis based on the KEGG and GSEA results (**Figures 6**, **7**). A recent study indicated that the degradation of p53 was inhibited in HSF2-depleted cells by regulating the expression of PSMD10, an oncogene that interacts with the ubiquitin ligase MDM2 (Lecomte et al., 2010). In addition to PSMD10, the expression of some proteasome subunits, including PSMD1, PSMD2, PSMC4, ubb, and ubc, was also downregulated in the absence of HSF2 (Lecomte et al., 2010). As proteasome inhibition is an important strategy for the treatments of cancers, targeting HSF2 may be a valuable tool to reduce chemoresistance to proteasome inhibition. More importantly, we found that HSF2 was associated with various oncogenesis-related pathways, such as the cell cycle, Hippo signaling pathway, mismatch repair, ErbB signaling pathway, and mTOR signaling pathway (**Figures 6**, **7**). Consistent with our previous studies, the results of the present study strengthen the important role of HSF2 in neurodegenerative diseases, including spinocerebellar ataxia (**Figures 6**, **7**). Our previous study demonstrated that HSF2 deficiency accelerated disease progression and shortened lifespan in a mouse model of Huntington's disease, suggesting that HSF2 could be a potential therapeutic target for neurodegenerative diseases by regulating the expression of αB-crystallin (CRYAB) (Shinkawa et al., 2011). Our GSEA results further implied that HSF2 was closely associated with the histone acetyltransferase complex (**Figure 6**). A previous study has shown that HSF2 can interact with WDR5, a core component of the Set1/MLL H3K4 histone methyltransferase complex (Hayashida, 2015). Moreover, HSF2 modifies active histone markers in the CRYAB promoter, including H3K4me3, H3K14Ac, and H3K27Ac (Hayashida, 2015). In fact, in addition to HSPs, HSF2 also regulates other target genes associated with oncogenesis, such as c-Fos (Wilkerson et al., 2007). Bioinformatics analysis has demonstrated that HSF2 may be involved in the oncogenesis of thyroid carcinoma by mediating the expression of SERPINA1 and FOSB (Lu and Zhang, 2016).

Oncogenesis is a complicated process accompanied by increased proliferation, resistance to cell death, enhanced angiogenesis, escape from immune surveillance, and tumor microenvironment (TME). The TME has attracted wide attention in cancer immunotherapy and has been identified as a main contributor to cancer initiation and development. It is well known that immunosurveillance affects the prognosis of cancer patients and that tumors can evade immune responses and immunotherapy by taking advantage of immune checkpoint genes, such as PD-1, PD-L1, and CTLA-4 (Gong et al., 2018; Kruger et al., 2019). Recently, immunotherapy has been recognized as an effective new strategy for cancer treatment. Although immunotherapy has made breakthroughs in cancer treatment, it still faces many challenges, and only a limited proportion of cancer patients respond well to immunotherapy (Gong et al., 2018; Kruger et al., 2019). Therefore, the identification of new targets and biomarkers is the key to further improving the efficacy of immunotherapy. Tumor-infiltrating immune cells, including B cells, T cells, dendritic cells, macrophages, and neutrophils, are the major part of the TME. Notably, our GSEA and KEGG results suggested that HSF2 was also involved in many immunity-associated pathways (IL−17 signaling pathway and the adaptive immune system) and various microbial infections (hepatitis B, shigellosis, Yersinia infection, and herpes simplex virus 1 infection) (**Figures 5**, **6**). A recently study showed that HSF2 was upregulated in ulcerative colitis and was negatively associated with colon inflammation in mice (Wang W. et al., 2020; Zhang F. et al., 2021). NLRP3 inflammasome activation and IL-1β secretion are greatly enhanced in HSF2−/− DSS model mice (Zhang et al., 2020). Consistently, overexpression of HSF2 significantly suppresses inflammation-related processes, indicating that HSF2 participates in inflammation. Moreover, the expression of HSF2 is obviously higher in the intestinal mucosa of UC patients (Miao et al., 2014). More importantly, serum HSF2 levels are positively correlated with the expression of IL-1β and TNF-α. Knockdown of HSF2 potentiates the production of IL-1β and TNF-α induced by LPS (Miao et al., 2014; Wang W. et al., 2020; Zhang et al., 2020; Santopolo et al., 2021; Zhang F. et al., 2021). Here, to further estimate the relationships between HSF2 and the TME, we first examined the correlation of HSF2 expression and the abundance of different infiltrating immune cells across different cancer types. HSF2 expression was significantly linked with the abundance of infiltrating $CD4^+$ T cells in 12 types of cancer, $CD8^+$ T cells in 16 types of cancer, B cells in 16 types of cancer, macrophages in 17 types of cancer, neutrophils in 18 types of cancer, and DCs in 18 types of cancer (**Figure 8**). In addition, HSF2 positively correlated with these six types of immune cells in KIRC, LIHC, PAAD, and PRAD but negatively correlated with them in SARC (**Figure 8A**). We also used the xCell algorithm to further estimate the relationship between HSF2 expression and the level of infiltrating immune cells and found that HSF2 expression was significantly correlated with infiltrating $CD4^+$ T helper (Th) cells and macrophages in most cancer types (**Figure 8B**). Tumor-associated macrophages (TAMs) within the TME have attracted great interest in basic science regarding their roles in metastasis, angiogenesis, and immunosuppression in various cancers (Mills et al., 2016; DeNardo and Ruffell, 2019; Anderson et al., 2021). The infiltration of TAMs in and around the tumor nest is one of the most important hallmarks of the process of cancer development. Macrophages consist of at least two subgroups, including proinflammatory M1 macrophages, and antiinflammatory M2 macrophages (Mills et al., 2016; Duan and Luo, 2021). M1 macrophages are cancer resistant due to their intrinsic phagocytosis, high antigen presenting capacity, and antitumor inflammatory activity. M1 macrophages also produce reactive oxygen species (ROS) and cytokines and are correlated with a favorable prognosis in cancer patients. In contrast, M2 macrophages are endowed with a repertoire of tumor-promoting capabilities associated with immunosuppression, angiogenesis, and neovascularization. A better understanding of their polarization into a protumoral phenotype to regulate tumor growth, angiogenesis, metastasis, and immune evasion prompted us to investigate their clinical significance as biomarkers in diverse cancers (Mills et al., 2016; DeNardo and Ruffell, 2019; Anderson et al., 2021; Duan and Luo, 2021). Here, we observed that HSF2 was significantly and negatively associated with the abundance of infiltrating macrophages, including M1 and M2 phenotypes, in most tumors, and illustrating the complexity of the TME. Moreover, accumulating evidence suggests that Th cells are essential to the development of the immune response and are involved in the response to antitumor immunotherapy (Basu et al., 2021; Renaude et al., 2021). T helper 1 (Th1) and T helper 2 (Th2) cells are the two predominant subtypes of CD4$^+$ Th cells. Th1 cells play an antitumor role by orchestrating immunity against tumor cells. Th1 cells enhance the generation and function of CD8$^+$ T cells, prevent angiogenesis, promote the senescence of tumor cells, and protect effector cytotoxic T lymphocytes from exhaustion; thus, modulating the Th1 cell response may lead to effective immune-based therapy (Basu et al., 2021; Laba et al., 2021; Renaude et al., 2021). Following differentiation, Th2 cells can produce IL-4, IL-5, IL-10, IL-13, and IL-17, not all of which are beneficial in cancer and contribute to tumor growth and metastasis. Simultaneously, infiltration of Th2 cells in the TME is usually connected with a poor prognosis in human cancers (Basu et al., 2021; Laba et al., 2021; Renaude et al., 2021). In the present study, we found that HSF2 expression was negatively associated with the infiltration level of Th1 cells but positively correlated with the level of infiltrating Th2 cells in most tumor types. More directly, we also investigated the effects of HSF2 expression on immune modulators, including immunoinhibitors, immunostimulators, chemokines, and MHC-related genes (**Figure 9**). Intriguingly, we observed that the expression of HSF2 was greatly associated with most immunoinhibitors, including PD-1, PD-L1, and CTLA4, in OV, PAAD, PRAD, and LIHC (**Figure 9**). Furthermore, HSF2 was also significantly correlated with the expression of different chemokines in PAAD, PRAD, and SARC; positively associated with immunostimulators in HNSC, KICH, KIRC, LIHC, OV, PAAD, and PRAD; and strongly correlated with most MHC-related genes in KIRC, OV, PAAD, PRAD, LGG, and SARC (**Figure 9**). Additionally, HSF2 expression significantly affected the expression of well-known immune checkpoints (**Figure 11**). TMB influences the possibility of generating immunogenic peptides, therefore affecting the response to immunotherapy in cancer patients. MSI is another important index for predicting oncogenesis and tumor development. Therefore, TMB and MSI could act as predictive factors for the efficacy of immune checkpoint inhibitors (Sha et al., 2020; Li et al., 2021). Here, we observed that the expression of HSF2 was associated with TMB and MSI in several cancer types (**Figure 10**). These results provide further clues regarding the correlation between HSF2 expression and cancer immunity. Based on our observations, targeting HSF2 may be a promising immunotherapeutic strategy for the treatment of specific cancers. Until now, there have been no small molecule drugs specifically targeting HSF2. Efforts are needed to develop novel drugs or RNAi techniques targeting HSF2 in tumor-infiltrative immune cells. Conversely, engineering tumor-specific macrophages and Th cells by modulating HSF2 expression may also be a promising strategy to increase the efficacy of immunotherapy.

Our results may provide better prognostic prediction and immune-oncological perspectives regarding the application of HSF2 as a prognostic biomarker. However, despite performing these bioinformatics analyses by collecting information from various databases, this study has several limitations. First, some contradictory findings of individual cancers in different databases were observed. It is therefore necessary to further investigate the expression and function of HSF2 using a large sample size. A deeper understanding of these differences may facilitate the development a global view to generate cancer development mechanisms with HSF2 expression. Second, although the signaling pathways and prognostic value of HSF2 in different cancer types were explored, there were no *in vitro* or *in vivo* experiments to verify these findings. Third, the effects of HSF2 on immune cell infiltration and immunotherapy in human cancer require experimental and clinical validation.

CONCLUSION

The results of the current study reveal the varied expression of HSF2 in different types and stages of cancers, which suggests that the effects of HSF2 on oncogenesis may vary across different cancer types. A significant correlation between HSF2 expression and the prognosis of cancer patients was observed. HSF2 expression was strongly related to immune cell infiltration, immune checkpoint genes, TMB, and MSI. The present study integrated existing data to explore the potential function of HSF2 in cancers and provides insights for targeting HSF2 to improve the therapeutic efficacy of 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.

AUTHOR CONTRIBUTIONS

Study concept and design: KT. Acquisition of data: KT, FC, YF, XL, JZ, and PC. Analysis and interpretation of data: KT, YF, FC, XL, JZ, YS, BZ, BL, JH, and PC. Statistical analysis: FC, YF, XL, and JZ. Drafting of the article: KT. Critical revision and final approval of the article: KT, PC, and YF. Obtained funding: YF, JZ, XL, and KT. Study supervision: KT. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the One Hundred Person Project of Hebei Province (E2016100019), Medical Science Research Project of Health Commission of Hebei Province (20210970), Natural Science Foundation of Hebei Province (C2020205003 and C2019205262), China Postdoctoral Science Foundation (2017M621099), and Hebei Normal University Foundation (L2019B27).

ACKNOWLEDGMENTS

We would like to thank TCGA, GTEx, CCLE, cBioPortal, UALCAN, TIMER, and xCell databases for the availability of the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.789703/full#supplementary-material

REFERENCES

Åkerfelt, M., Henriksson, E., Laiho, A., Vihervaara, A., Rautoma, K., Kotaja, N., et al. (2008). Promoter ChIP-Chip Analysis in Mouse Testis Reveals Y Chromosome Occupancy by HSF2. *Proc. Natl. Acad. Sci.* 105 (32), 11224–11229. doi:10.1073/pnas.0800620105

Åkerfelt, M., Morimoto, R. I., and Sistonen, L. (2010). Heat Shock Factors: Integrators of Cell Stress, Development and Lifespan. *Nat. Rev. Mol. Cel Biol* 11 (8), 545–555. doi:10.1038/nrm2938

Anderson, N. R., Minutolo, N. G., Gill, S., and Klichinsky, M. (2021). Macrophage-Based Approaches for Cancer Immunotherapy. *Cancer Res.* 81 (5), 1201–1208. doi:10.1158/0008-5472.Can-20-2990

Basu, A., Ramamoorthi, G., Albert, G., Gallen, C., Beyer, A., Snyder, C., et al. (2021). Differentiation and Regulation of TH Cells: A Balancing Act for Cancer Immunotherapy. *Front. Immunol.* 12, 669474. doi:10.3389/fimmu.2021.669474

Björk, J. K., Åkerfelt, M., Joutsen, J., Puustinen, M. C., Cheng, F., Sistonen, L., et al. (2016). Heat-shock Factor 2 Is a Suppressor of Prostate Cancer Invasion. *Oncogene* 35 (14), 1770–1784. doi:10.1038/onc.2015.241

Björk, J. K., Sandqvist, A., Elsing, A. N., Kotaja, N., and Sistonen, L. (2010). miR-18, a Member of Oncomir-1, Targets Heat Shock Transcription Factor 2 in Spermatogenesis. *Development* 137 (19), 3177–3184. doi:10.1242/dev.050955

Chang, Y., Östling, P., Åkerfelt, M., Trouillet, D., Rallu, M., Gitton, Y., et al. (2006). Role of Heat-Shock Factor 2 in Cerebral Cortex Formation and as a Regulatorof P35 Expression. *Genes Dev.* 20 (7), 836–847. doi:10.1101/gad.366906

Chen, F., Fan, Y., Cao, P., Liu, B., Hou, J., Zhang, B., et al. (2021). Pan-Cancer Analysis of the Prognostic and Immunological Role of HSF1: A Potential Target for Survival and Immunotherapy. *Oxidative Med. Cell Longevity* 2021, 1–21. doi:10.1155/2021/5551036

Dai, C., and Sampson, S. B. (2016). HSF1: Guardian of Proteostasis in Cancer. *Trends Cel Biol.* 26 (1), 17–28. doi:10.1016/j.tcb.2015.10.011

Dai, C. (2018). The Heat-Shock, or HSF1-Mediated Proteotoxic Stress, Response in Cancer: from Proteomic Stability to Oncogenesis. *Phil. Trans. R. Soc. B* 373, 20160525. doi:10.1098/rstb.2016.0525

DeNardo, D. G., and Ruffell, B. (2019). Macrophages as Regulators of Tumour Immunity and Immunotherapy. *Nat. Rev. Immunol.* 19 (6), 369–382. doi:10.1038/s41577-019-0127-6

Duan, Z., and Luo, Y. (2021). Targeting Macrophages in Cancer Immunotherapy. *Sig Transduct Target. Ther.* 6 (1), 127. doi:10.1038/s41392-021-00506-6

Fan, Y., Liu, B., Chen, F., Song, Z., Han, B., Meng, Y., et al. (2021). Hepcidin Upregulation in Lung Cancer: A Potential Therapeutic Target Associated with Immune Infiltration. *Front. Immunol.* 12, 612144. doi:10.3389/fimmu.2021.612144

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., et al. (2021). Cancer Statistics for the Year 2020: An Overview. *Int. J. Cancer*. doi:10.1002/ijc.33588

Fujimoto, M., and Nakai, A. (2010). The Heat Shock Factor Family and Adaptation to Proteotoxic Stress. *Febs j* 277 (20), 4112–4125. doi:10.1111/j.1742-4658.2010.07827.x

Gomez-Pastor, R., Burchfiel, E. T., and Thiele, D. J. (2018). Regulation of Heat Shock Transcription Factors and Their Roles in Physiology and Disease. *Nat. Rev. Mol. Cel Biol* 19 (1), 4–19. doi:10.1038/nrm.2017.73

Gong, J., Chehrazi-Raffle, A., Reddi, S., and Salgia, R. (2018). Development of PD-1 and PD-L1 Inhibitors as a Form of Cancer Immunotherapy: a Comprehensive Review of Registration Trials and Future Considerations. *J. Immunotherapy Cancer* 6 (1), 8. doi:10.1186/s40425-018-0316-z

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: the Next Generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Hayashida, N. (2015). Set1/MLL Complex Is Indispensable for the Transcriptional Ability of Heat Shock Transcription Factor 2. *Biochem. Biophysical Res. Commun.* 467 (4), 805–812. doi:10.1016/j.bbrc.2015.10.061

Jeggo, P. A., Pearl, L. H., and Carr, A. M. (2016). DNA Repair, Genome Stability and Cancer: a Historical Perspective. *Nat. Rev. Cancer* 16 (1), 35–42. doi:10.1038/nrc.2015.4

Kallio, M., Chang, Y., Manuel, M., Alastalo, T. P., Rallu, M., Gitton, Y., et al. (2002). Brain Abnormalities, Defective Meiotic Chromosome Synapsis and Female Subfertility in HSF2 Null Mice. *Embo j* 21 (11), 2591–2601. doi:10.1093/emboj/21.11.2591

Kruger, S., Ilmer, M., Kobold, S., Cadilha, B. L., Endres, S., Ormanns, S., et al. (2019). Advances in Cancer Immunotherapy 2019 - Latest Trends. *J. Exp. Clin. Cancer Res.* 38 (1), 268. doi:10.1186/s13046-019-1266-0

Laba, S., Mallett, G., and Amarnath, S. (2021). The Depths of PD-1 Function within the Tumor Microenvironment beyond CD8+ T Cells. *Semin. Cancer Biol.* doi:10.1016/j.semcancer.2021.05.022

Lecomte, S., Desmots, F., Le Masson, F., Le Goff, P., Michel, D., Christians, E. S., et al. (2010). Roles of Heat Shock Factor 1 and 2 in Response to Proteasome Inhibition: Consequence on P53 Stability. *Oncogene* 29 (29), 4216–4224. doi:10.1038/onc.2010.171

Li, L., Bai, L., Lin, H., Dong, L., Zhang, R., Cheng, X., et al. (2021). Multiomics Analysis of Tumor Mutational burden across Cancer Types. *Comput. Struct. Biotechnol. J.* 19, 5637–5646. doi:10.1016/j.csbj.2021.10.013

Li, P., Sheng, C., Huang, L., Zhang, H., Huang, L., Cheng, Z., et al. (2014). MiR-183/-96/-182 Cluster Is Up-Regulated in Most Breast Cancers and Increases Cell Proliferation and Migration. *Breast Cancer Res.* 16 (6), 473. doi:10.1186/s13058-014-0473-z

Liu, B., Song, Z., Fan, Y., Zhang, G., Cao, P., Li, D., et al. (2021). Downregulation of FPN1 Acts as a Prognostic Biomarker Associated with Immune Infiltration in Lung Cancer. *Aging* 13 (6), 8737–8761. doi:10.18632/aging.202685

Lu, J. C., and Zhang, Y. P. (2016). E2F, HSF2, and miR-26 in Thyroid Carcinoma: Bioinformatic Analysis of RNA-Sequencing Data. *Genet. Mol. Res.* 15 (1), 15017576. doi:10.4238/gmr.15017576

Mathew, A., Mathur, S. K., and Morimoto, R. I. (1998). Heat Shock Response and Protein Degradation: Regulation of HSF2 by the Ubiquitin-Proteasome Pathway. *Mol. Cel Biol* 18 (9), 5091–5098. doi:10.1128/mcb.18.9.5091

Mendillo, M. L., Santagata, S., Koeva, M., Bell, G. W., Hu, R., Tamimi, R. M., et al. (2012). HSF1 Drives a Transcriptional Program Distinct from Heat Shock to Support Highly Malignant Human Cancers. *Cell* 150 (3), 549–562. doi:10.1016/j.cell.2012.06.031

Meng, X., Chen, X., Lu, P., Ma, W., Yue, D., Song, L., et al. (2017). miR-202 Promotes Cell Apoptosis in Esophageal Squamous Cell Carcinoma by Targeting HSF2. *Oncol. Res.* 25 (2), 215–223. doi:10.3727/096504016x14732772150541

Miao, J., Niu, J., Wang, K., Xiao, Y., Du, Y., Zhou, L., et al. (2014). Heat Shock Factor 2 Levels Are Associated with the Severity of Ulcerative Colitis. *PLoS One* 9 (2), e88822. doi:10.1371/journal.pone.0088822

Mills, C. D., Lenz, L. L., and Harris, R. A. (2016). A Breakthrough: Macrophage-Directed Cancer Immunotherapy. *Cancer Res.* 76 (3), 513–516. doi:10.1158/0008-5472.Can-15-1737

Mustafa, D. A. M., Sieuwerts, A. M., Zheng, P. P., and Kros, J. M. (2010). Overexpression of Colligin 2 in Glioma Vasculature Is Associated with Overexpression of Heat Shock Factor 2. *Generegul Syst. Bio* 4, GRSB.S4546–107. doi:10.4137/grsb.S4546

Östling, P., Björk, J. K., Roos-Mattjus, P., Mezger, V., and Sistonen, L. (2007). Heat Shock Factor 2 (HSF2) Contributes to Inducible Expression of Hsp Genes through Interplay with HSF1. *J. Biol. Chem.* 282 (10), 7077–7086. doi:10.1074/jbc.M607556200

Puustinen, M. C., and Sistonen, L. (2020). Molecular Mechanisms of Heat Shock Factors in Cancer. *Cells* 9 (5), 1202. doi:10.3390/cells9051202

Rallu, M., Loones, M., Lallemand, Y., Morimoto, R., Morange, M., and Mezger, V. (1997). Function and Regulation of Heat Shock Factor 2 during Mouse Embryogenesis. *Proc. Natl. Acad. Sci.* 94 (6), 2392–2397. doi:10.1073/pnas.94.6.2392

Renaude, E., Kroemer, M., Borg, C., Peixoto, P., Hervouet, E., Loyon, R., et al. (2021). Epigenetic Reprogramming of CD4+ Helper T Cells as a Strategy to Improve Anticancer Immunotherapy. *Front. Immunol.* 12, 669992. doi:10.3389/fimmu.2021.669992

Sandqvist, A., Björk, J. K., Åkerfelt, M., Chitikova, Z., Grichine, A., Vourc'h, C., et al. (2009). Heterotrimerization of Heat-Shock Factors 1 and 2 Provides a Transcriptional Switch in Response to Distinct Stimuli. *MBoC* 20 (5), 1340–1347. doi:10.1091/mbc.e08-08-0864

Santopolo, S., Riccio, A., Rossi, A., and Santoro, M. G. (2021). The Proteostasis Guardian HSF1 Directs the Transcription of its Paralog and Interactor HSF2 during Proteasome Dysfunction. *Cell. Mol. Life Sci.* 78 (3), 1113–1129. doi:10.1007/s00018-020-03568-x

Sarge, K. D., Park-Sarge, O.-K., Kirby, J. D., Mayo, K. E., and Morimoto, R. I. (1994). Expression of Heat Shock Factor 2 in Mouse Testis: Potential Role as a Regulator of Heat-Shock Protein Gene Expression during Spermatogenesis1. *Biol. Reprod.* 50 (6), 1334–1343. doi:10.1095/biolreprod50.6.1334

Sha, D., Jin, Z., Budczies, J., Kluck, K., Stenzinger, A., and Sinicrope, F. A. (2020). Tumor Mutational Burden as a Predictive Biomarker in Solid Tumors. *Cancer Discov.* 10 (12), 1808–1825. doi:10.1158/2159-8290.Cd-20-0522

Shinkawa, T., Tan, K., Fujimoto, M., Hayashida, N., Yamamoto, K., Takaki, E., et al. (2011). Heat Shock Factor 2 Is Required for Maintaining Proteostasis against Febrile-Range thermal Stress and Polyglutamine Aggregation. *MBoC* 22 (19), 3571–3583. doi:10.1091/mbc.E11-04-0330

Sistonen, L., Sarge, K. D., and Morimoto, R. I. (1994). Human Heat Shock Factors 1 and 2 Are Differentially Activated and Can Synergistically Induce Hsp70 Gene Transcription. *Mol. Cel. Biol.* 14 (3), 2087–2099. doi:10.1128/mcb.14.3.2087-2099.1994

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Wang, G., Cao, P., Fan, Y., and Tan, K. (2020a). Emerging Roles of HSF1 in Cancer: Cellular and Molecular Episodes. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1874 (1), 188390. doi:10.1016/j.bbcan.2020.188390

Wang, G., Zhang, J., Moskophidis, D., and Mivechi, N. F. (2003). Targeted Disruption of the Heat Shock Transcription Factor (Hsf)-2 Gene Results in Increased Embryonic Lethality, Neuronal Defects, and Reduced Spermatogenesis. *Genesis* 36 (1), 48–61. doi:10.1002/gene.10200

Wang, W., Zhang, F., Li, X., Luo, J., Sun, Y., Wu, J., et al. (2020b). Heat Shock Transcription Factor 2 Inhibits Intestinal Epithelial Cell Apoptosis through the Mitochondrial Pathway in Ulcerative Colitis. *Biochem. Biophysical Res. Commun.* 527 (1), 173–179. doi:10.1016/j.bbrc.2020.04.103

Widlak, W., and Vydra, N. (2017). The Role of Heat Shock Factors in Mammalian Spermatogenesis. *Adv. Anat. Embryol. Cel Biol* 222, 45–65. doi:10.1007/978-3-319-51409-3_3

Wilkerson, D. C., Skaggs, H. S., and Sarge, K. D. (2007). HSF2 Binds to the Hsp90, Hsp27, and C-Fos Promoters Constitutively and Modulates Their Expression. *Cell Stress Chaper* 12 (3), 283–290. doi:10.1379/csc-250.1

Yang, L. N., Ning, Z. Y., Wang, L., Yan, X., and Meng, Z. Q. (2019). HSF2 Regulates Aerobic Glycolysis by Suppression of FBP1 in Hepatocellular Carcinoma. *Am. J. Cancer Res.* 9 (8), 1607–1621.

Yang, Y., Zhou, Y., Xiong, X., Huang, M., Ying, X., and Wang, M. (2018). ALG3 Is Activated by Heat Shock Factor 2 and Promotes Breast Cancer Growth. *Med. Sci. Monit.* 24, 3479–3487. doi:10.12659/msm.907461

Zhang, B., Fan, Y., Cao, P., and Tan, K. (2021a). Multifaceted Roles of HSF1 in Cell Death: A State-Of-The-Art Review. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1876 (2), 188591. doi:10.1016/j.bbcan.2021.188591

Zhang, F., Wang, W., Niu, J., Yang, G., Luo, J., Lan, D., et al. (2020). Heat-shock Transcription Factor 2 Promotes Sodium Butyrate-Induced Autophagy by Inhibiting mTOR in Ulcerative Colitis. *Exp. Cel Res.* 388 (1), 111820. doi:10.1016/j.yexcr.2020.111820

Zhang, F., Zhao, W., Zhou, J., Wang, W., Luo, J., Feng, Y., et al. (2021b). Heat Shock Transcription Factor 2 Reduces the Secretion of IL-1β by Inhibiting NLRP3 Inflammasome Activation in Ulcerative Colitis. *Gene* 768, 145299. doi:10.1016/j.gene.2020.145299

Zhong, Y.-H., Cheng, H.-Z., Peng, H., Tang, S.-C., and Wang, P. (2016). Heat Shock Factor 2 Is Associated with the Occurrence of Lung Cancer by Enhancing the Expression of Heat Shock Proteins. *Oncol. Lett.* 12 (6), 5106–5112. doi:10.3892/ol.2016.5368

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 Molecular Biosciences

ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/fmolb.2021.815243

Large-Scale Gastric Cancer Susceptibility Gene Identification Based on Gradient Boosting Decision Tree

Qing Chen[1†], Ji Zhang[1†], Banghe Bao[2], Fan Zhang[3] and Jie Zhou[4]*

[1]Department of Hepatobiliary Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2]Department of Pathology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [3]Wuhan Asia General Hospital, Wuhan, China, [4]Department of Biochemistry and Molecular Biology, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

OPEN ACCESS

Edited by:
Tianyi Zhao,
Harbin Institute of Technology, China

Reviewed by:
Ningyi Zhang,
Harbin Institute of Technology, China
Shuxuan Lyu,
Xian Jiaotong University, China

***Correspondence:**
Jie Zhou
chentianyuzj@126.com

†These authors share first authorship.

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics,
a section of the journal
Frontiers in Molecular Biosciences

Received: 15 November 2021
Accepted: 06 December 2021
Published: 13 January 2022

Citation:
Chen Q, Zhang J, Bao B, Zhang F and Zhou J (2022) Large-Scale Gastric Cancer Susceptibility Gene Identification Based on Gradient Boosting Decision Tree.
Front. Mol. Biosci. 8:815243.
doi: 10.3389/fmolb.2021.815243

The early clinical symptoms of gastric cancer are not obvious, and metastasis may have occurred at the time of treatment. Poor prognosis is one of the important reasons for the high mortality of gastric cancer. Therefore, the identification of gastric cancer-related genes can be used as relevant markers for diagnosis and treatment to improve diagnosis precision and guide personalized treatment. In order to further reveal the pathogenesis of gastric cancer at the gene level, we proposed a method based on Gradient Boosting Decision Tree (GBDT) to identify the susceptible genes of gastric cancer through gene interaction network. Based on the known genes related to gastric cancer, we collected more genes which can interact with them and constructed a gene interaction network. Random Walk was used to extract network association of each gene and we used GBDT to identify the gastric cancer-related genes. To verify the AUC and AUPR of our algorithm, we implemented 10-fold cross-validation. GBDT achieved AUC as 0.89 and AUPR as 0.81. We selected four other methods to compare with GBDT and found GBDT performed best.

Keywords: gastric cancer, susceptibility gene, gradient boosting decision tree (GBDT), random walk (RW), gastric cancer-related genes

INTRODUCTION

There are about 950,000 new cases of gastric cancer worldwide each year, and nearly 700,000 deaths. It is one of the most serious tumors (Rawla and Barsouk, 2019). The early clinical symptoms of gastric cancer are not obvious, and metastasis may have occurred at the time of treatment (Axon, 2006). Poor prognosis is one of the important reasons for the high mortality of gastric cancer (Eguchi et al., 2003). Therefore, the identification of gastric cancer-related genes can be used as relevant markers for diagnosis and treatment to improve diagnosis precision and guide personalized treatment (Duffy et al., 2014).

Identifying gastric cancer-related genes plays an important role in the treatment of gastric cancer. Research on metastasis-related genes is conducive to timely detection of early metastasis, screening of new markers and therapeutic targets, thereby improving the survival rate of patients (Arturi et al., 1997). Using animal models to screen gastric cancer metastasis-related genes (Wang and Chen, 2002), fully mimic the process of tumor metastasis *in vivo*, with high metastasis efficiency, clear phenotypic characteristics, and good clinical similarity. Cell line derived xenograft (CDX) model is a tumor model constructed by transplanting cultured tumor cells into immunodeficient mice (Georges et al., 2019). The cell lines used in the CDX model have been cultured *in vitro* for many generations, and their biological characteristics have changed significantly. Some tumor cell lines that adapt to culture *in vitro* and have metastatic potential have been selected, so it is easy to obtain the metastasis model. The establishment of the CDX model can be realized by subcutaneous injection, intraperitoneal injection, caudal vein injection, and so on (Lallo et al., 2017). Zhu et al. (2020) established a xenotransplantation model by subcutaneous injection of gastric cancer cell line BGC-823 into the hind limbs of nude mice. They found that mir-106a had the potential to promote tumor growth by targeting Smad7. At the same time, they found that mir-106a was related to peritoneal metastasis of gastric cancer. At present, studies have found that gastrin level has a strong relationship with the development of gastric cancer. Zu et al. (2018) successfully established a cell xenotransplantation model by subcutaneous injection of human gastric cancer cell line SGC-7901 in nude mice. They found that gastrin can inhibit the proliferation of poorly differentiated gastric cancer cells and enhance the inhibitory effect of cisplatin on gastric cancer by activating erk-p65-mir 23a/27a/24 axis. Tumor cells with biological enzyme markers can also be used to establish a CDX model (Agashe and Kurzrock, 2020), which is helpful to dynamically monitor tumor metastasis *in vivo* and facilitate the screening of metastasis related genes. Miwa et al. (2019) successfully established the intraperitoneal metastasis model by injecting MKN1 (MKN1 LUC) and MKN45 (MKN45 LUC) gastric cancer cells stably expressing luciferase and n87, Kato III, nugc4, and ocum-1 gastric cancer cells into the abdominal cavity of nude mice. The liver metastasis model was successfully established by injecting MKN1 Luc and MKN45 Luc directly into the portal vein of mice. Because the establishment of CDX model uses passage cell lines and lacks the microenvironment of tumor growth in human body (Lallo et al., 2017), it cannot well simulate the growth and metastasis of tumor in the human body. Patient derived cell models (PDC) use patient derived tumor cells isolated from malignant effusions such as ascites and pleural effusion (Bolck et al., 2019). Therefore, it can better reflect the individualized characteristics of patients and show unique advantages in the screening of tumor metastasis related genes and clinical drug screening. Lee et al. (2015) established a PDC model with cells collected from patients with metastatic cancer. The study found that the genomic changes of primary tumor and offspring PDC model were highly consistent, and the correlation of average variant allele frequency was 0.878. Further compared the genomic characteristics of primary tumor P0, P1, and P2 cells, and found that three samples (P0, P1, and P2 cells) were highly correlated. The drug response of the model reflects the clinical response of patients to targeted drugs. Although the PDC model established by metastatic patient derived tumor cells can reflect the individualized characteristics of patients, it is cultured *in vitro*, which is difficult to culture and cannot simulate the process of tumor metastasis *in vivo*. Therefore, the use of this model to screen metastasis related genes is limited. The metastasis related genes screened by the above CDX model and PDC model are conducive to the discovery of relevant molecules promoting gastric cancer metastasis and provide help for the early detection of gastric cancer metastasis in the clinic (Almagro et al., 2014). Patient derived xenograft (PDX) model improves the shortcomings of the CDX model and the PDC model. It is a better model to screen metastasis related genes at present. The model is a xenotransplantation model established by transplanting fresh clinical surgical specimens into immunodeficient mice. It maintains the microenvironment of primary tumor growth, so it can better simulate the biological behavior of tumors *in vivo*. Choi et al. (2016) successfully established 15 cases of gastric cancer PDX models, and found that the histological and genetic characteristics of the tumor models remained stable in subsequent passages and were highly consistent with the primary tumor. This discovery made the use of PDX models for the development of gastric cancer molecules possible. Research and individualized treatment are possible. The PDX model has relatively consistent genomics characteristics with the primary tumor, which is very conducive to the screening of individualized metastasis-related genes. Zhang et al. (2015) successfully established 32 PDX models of gastric cancer, and found that the gene amplification of FGFR2, MET, and ERBB2 is very similar between PDX models and their parent tumors, and the expression of PTEN and MET proteins are also moderately consistent. These data are *in vivo* testing of individualized therapy and screening of transfer-related genes provides a theoretical basis. There are many methods of tissue transplantation when establishing a PDX model, including subcutaneous transplantation, renal capsule transplantation, orthotopic transplantation, etc. (Okada et al., 2018). Among them, subcutaneous transplantation is the most commonly used transplantation method. Guo et al. (2019) established a PDX model of gastric cancer by subcutaneous transplantation and revealed the molecular mechanism of ISL1 that promotes gastric cancer metastasis by combining the ZEB1 promoter and the cofactor SETD7. ISL1 may be a potential prognostic marker of gastric cancer. Because the microenvironment of orthotopic transplantation tumors is closer to the human environment, orthotopic transplantation can simulate the growth of tumors in the human body better than subcutaneous transplantation, and it is easier to simulate clinical metastasis, which is beneficial to screening metastasis-related genes. Wang et al. (2018) found that 28 miRNAs are differentially expressed in invasive gastric cancer through array analysis. Among these 28 miRNAs, miR-29b is one of the most significantly down-regulated miRNAs. RNA response element (miRNA response element, MRE) binds to the negative regulation of MMP2, thereby affecting the development of gastric cancer.

However, this kind of animal model experiment method is very costly and time consuming. With the continuous enhancement of computing power, computing methods have been able to process massive amounts of biological data and mine knowledge from the data (Zhao et al., 2021). Deep learning, machine learning, and reinforcement learning have been widely used in the fields of biology and medicine (Zhao et al., 2020a; Tianyi et al., 2020). These methods use existing knowledge to

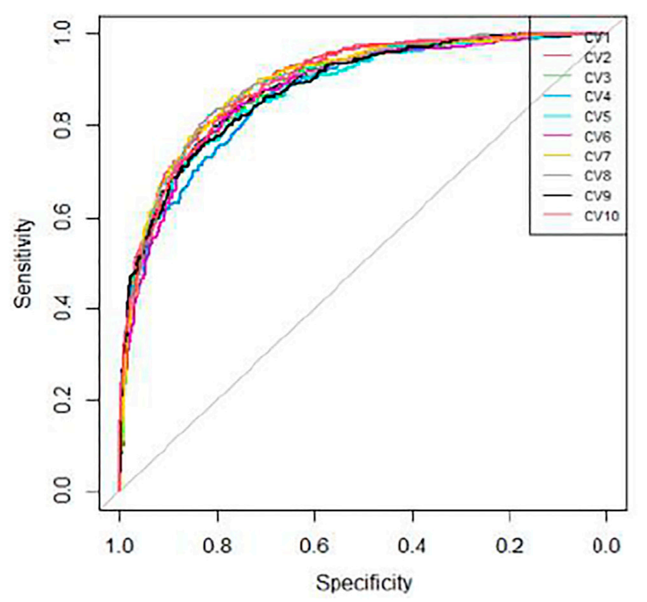


FIGURE 1 | ROC curves of 10-cross validation.

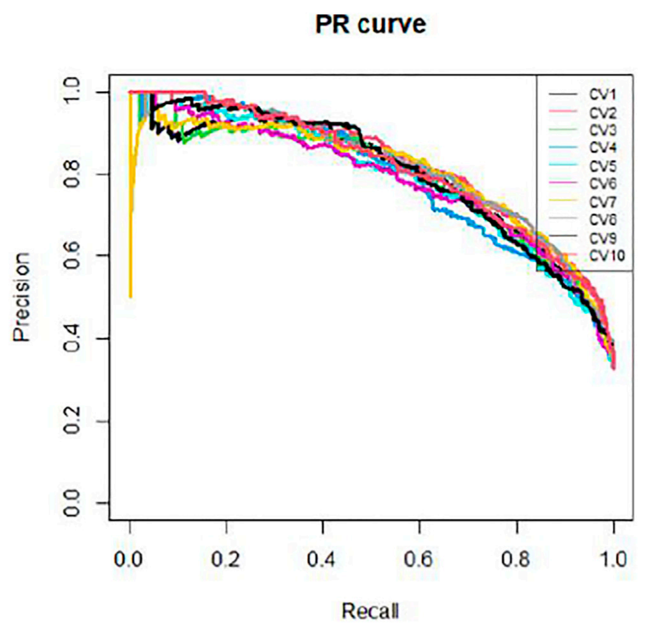


FIGURE 2 | PR curves of 10-cross validation.

construct complex mathematical models to predict new knowledge (Zhao et al., 2020b). In this paper, we extracted network association of each gene by Random Walk (RW) and used GBDT to identify the gastric cancer-related genes.

METHOD

We obtained 435 genes that are known to be related to gastric cancer in DisGeNet (Piñero et al., 2020). We collected genes that can interact with these 896 genes in HumanNet V2.0 (Hwang et al., 2019). Based on the interaction information, we built a gene interaction network. This network contains 1331 nodes, and each node is a gene.

Extracting Features by RW

The core formula of RW is as follows:

$$P_{t+1} = (1-\gamma)AP_t + \gamma P_0 \tag{1}$$

A is the adjacency matrix of the gene interaction network. P is random walk matrix. γ is a parameter that is needed to be set. We set γ as 0.5 based on experience.

If $\|P_{t+1} - P_t\| > \ell$ (we can set ℓ as arbitrarily small number), we can repeat Formula **(1)**. Otherwise, we could obtain P_{t+1} as the final RW matrix.

Identifying Gastric Cancer Susceptibility Gene by GBDT

After obtaining the feature of genes by RW, we need to build a classifier to identify whether a gene is associated with gastric cancer GBDT does not need to scale the data to build model, and it is also suitable for data sets where dual features and continuous features exist at the same time. First, the decision tree used by GBDT is a CART regression tree. Whether it is dealing with regression problems or two classifications and multiple classifications, the decision trees used by GBDT are all CART regression trees. Because the gradient value to be fitted in each iteration of GBDT is a continuous value, a regression tree is used. The most important thing for the regression tree algorithm is to find the best division point, then the division point in the regression tree contains all the desirable values of all features. The criterion for the best division point in the classification tree is entropy or Gini coefficient, which are both measured by purity, but the sample labels in the regression tree are continuous values, so it is no longer appropriate to use indicators such as entropy, instead of the square error, which can judge the degree of fit very well.

The process of constructing CART is as follows:

Input: training data set D. Output: regression tree f (x).

Recursively divide each region into two sub-regions in the input space where the training data set is located and determine the output value on each sub-region to construct a binary decision tree:

$$\min\left[\min\sum(y_i - c_1)^2 + \min\sum(y_i - c_2)^2\right] \tag{2}$$

As shown in Formula **(2)**, we need to choose (j, s) to minimize $\min\sum(y_i - c_1)^2 + \min\sum(y_i - c_2)^2$. Then, we need to introduce (j, s) to divide the area and determine the corresponding output value:

$$R1(j,s) = x|x(j) \le s, R2(j,s) = x|x(j) > s \tag{3}$$

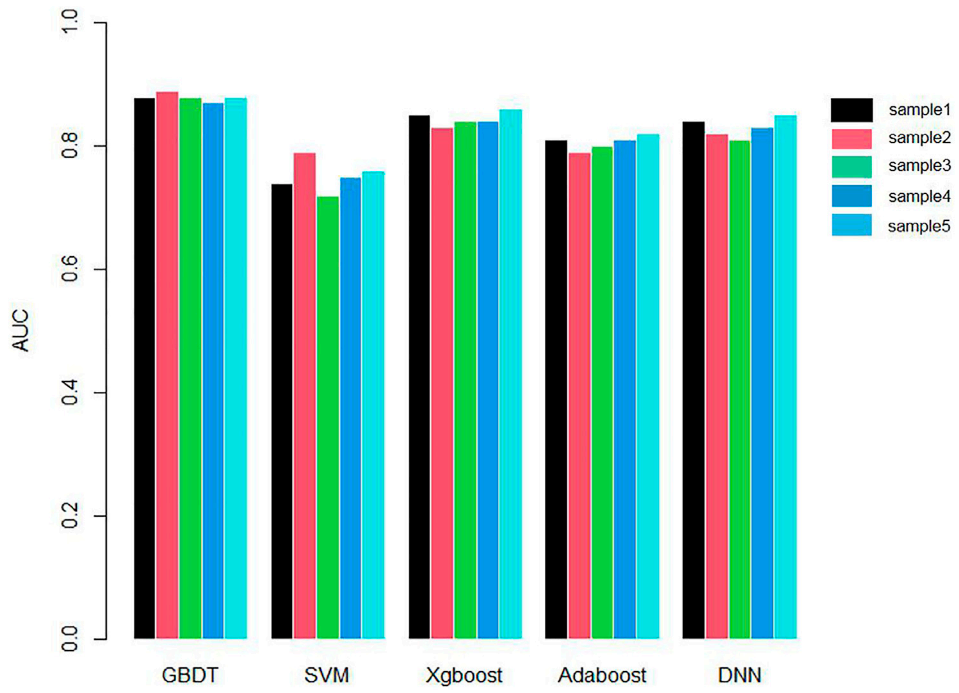


FIGURE 3 | Comparison chart of AUC values of five methods.

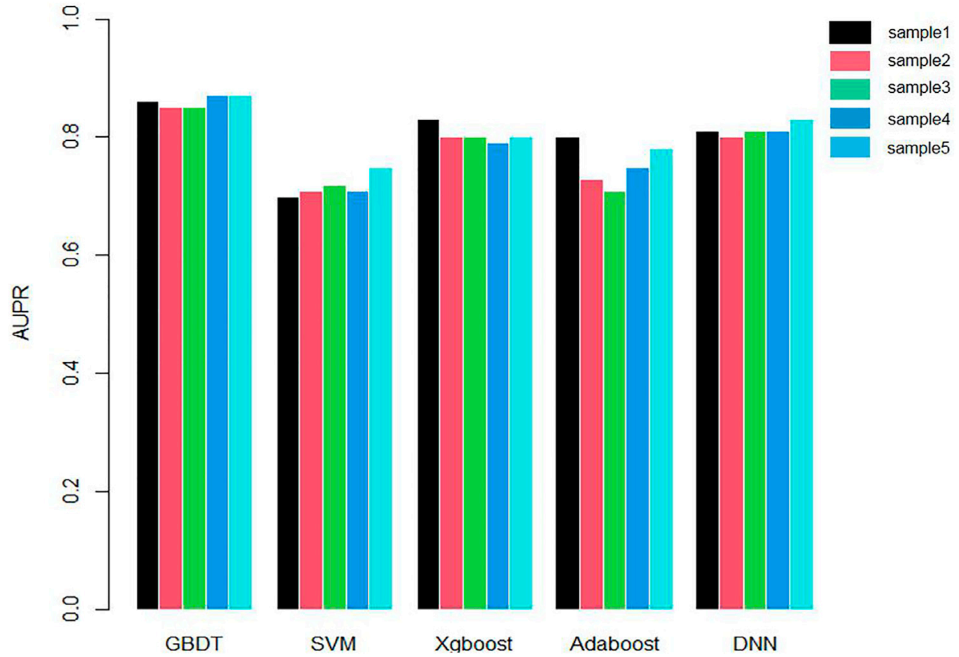


FIGURE 4 | Comparison chart of AUPR values of five methods.

$$\widehat{c}_m = \frac{1}{N} \sum_{x1\in R_m(j,s)} y_i, x \in R_m, m = 1, 2 \tag{4}$$

Continue to call Steps (1) and (2) for the two sub-regions until the stop condition is met.

Divide the input space into M regions $(R_1, R_2, ..., R_m)$, build the final decision tree.

$$f(x) = \sum_{m=1}^{M} \widehat{c}_m I(x \in R_m) \tag{5}$$

Gradient boosting is an improved algorithm of the Boosting Tree. There are three steps to implement the Boosting Tree.
Step 1 Initialize $f_0(x) = 0$.
Step 2 Calculate residual $r_{mi} = y_i - f_{m-1}(x), i = 1, 2, \ldots, N$.
Step 3 Fit the residual r_{mi} to obtain regression tree and obtain $h_m(x)$.
Step 4 Update $f_m(x)$, $f_m(x) = f_{m-1}(x) + h_m(x)$.
Step 5 The final regression boosting tree would be: $f_M(x) = \sum_{m=1}^{M} h_m(x)$.

Based on the Decision Tree and Gradient Boosting, we can combine them to obtain the final GBDT.

First, we need to initialize week learner.

$$f_0(x) = argmin_c \sum_{i=1}^{N} L(y_i, c) \tag{6}$$

For each sample i = 1, 2,..., N, we need to calculate the negative gradient (residual):

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{7}$$

Use the residual obtained in the previous step as the new true value of the sample and use (x_i, r_{im}) as the training data of the next tree to obtain the new regression tree $f_m(x)$. The leaf node area of $f_m(x)$ is $R_{jm}, j = 1, 2, \ldots, J$. J is the number of leaf nodes.

Calculate the Best Fit Value

$$\gamma_{jm} = argmin \sum_{x_i\in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \tag{8}$$

Update Strong Learner

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}) \tag{9}$$

Get the Final Learner

$$f(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M}\sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}) \tag{10}$$

RESULTS

Since we obtained 435 genes that are known to be related to gastric cancer in DisGeNet and 896 genes that have strong interaction with them, the 435 genes were used as the positive samples and 896 were used as negative samples. We used these data to build GBDT model to identify gastric cancer susceptibility genes.

We applied 10-cross validation to verify the accuracy of our model. The AUC (Area Under Curve) and AUPR (Area Under Precision Curve) of our model is shown as **Figures 1** and **2**, respectively. The average AUC of 10-cross validation is 0.89 ± 0.008 and average AUPR of 10-cross validation is 0.81 ± 0.006. Since the number of negative samples is significantly higher than positive samples, to balance the training sample set, we randomly selected 435 negative samples from 896 genes each time and repeat the 10-cross validation. In addition, we also compared our method with other methods, such as Support Vector Machine (SVM), Xgboost, Adaboost, and Deep Neural Network (DNN). We totally randomly sampled five negative sets. The performance of these methods is shown as **Figures 3** and **4**.

As shown in **Figures 3** and **4**, the AUC and AUPR of GBDT are higher than other methods, which explains the superiority of our method over other methods.

CONCLUSION

Through early detection, early diagnosis, and early treatment, the cure rate of patients with early gastric cancer can reach 85%; However, the 5-year survival rate of patients with advanced gastric cancer is less than 10%. At present, inhibitors targeting vascular endothelial growth factor (VEGF), epidermal growth factor (EGF), and tyrosine kinase have been successfully developed, showing significant curative effects on gastric cancer. This greatly encourages us to study the characteristic markers of recurrence or metastasis of gastric cancer from the perspective of genes. Few genes related to gastric cancer have been found in cohort studies and animal model experiments. However, due to the cost, such methods cannot be popularized large scale.

In this paper, we proposed a novel method to identify gastric cancer-related genes in large scale. Genes that interact more closely are more likely to be related to similar diseases. Based on this hypothesis, we considered to use the gene interaction information to build a network and infer the gastric cancer-related genes by this network. RW was applied to encode the features of genes and GBDT was implemented to identify gastric cancer-related genes. We verified our method by two kinds of 10-cross validation experiments. Our method showed high accuracy in both experiments, indicating that our method can be used to identify genes related to liver fibrosis. The method proposed in this article will provide guidance for genetic mechanism and clinical treatment of gastric cancer.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

QC, JiZ, and JeZ designed the study. QC, JiZ, BB, and FZ interpreted the data and analyzed the results. All authors read and approved the final manuscript.

FUNDING

Financial support comes from the National Natural Science Foundation of China (81371508, 81572985, and 31000471).

REFERENCES

Agashe, R., and Kurzrock, R. (2020). Circulating Tumor Cells: From the Laboratory to the Cancer Clinic. *Cancers* 12, 2361. doi:10.3390/cancers12092361

Almagro, J. C., Gilliland, G. L., Breden, F., Scott, J. K., Sok, D., Pauthner, M., et al. (2014). *Antibody Engineering and Therapeutics: December 8–12, 2013*. Huntington Beach, CA: Taylor & Francis, 577–618. doi:10.4161/mabs.28421

Arturi, F., Russo, D., Giuffrida, D., Ippolito, A., Perrotti, N., Vigneri, R., et al. (1997). Early Diagnosis by Genetic Analysis of Differentiated Thyroid Cancer Metastases in Small Lymph Nodes. *J. Clin. Endocrinol. Metab.* 82, 1638. doi:10.1210/jcem.82.5.4062

Axon, A. (2006). Symptoms and Diagnosis of Gastric Cancer at Early Curable Stage. *Best Pract. Res. Clin. Gastroenterol.* 20, 697–708. doi:10.1016/j.bpg.2006.03.015

Bolck, H. A., Pauli, C., Göbel, E., Mühlbauer, K., Dettwiler, S., Moch, H., et al. (2019). Cancer Sample Biobanking at the Next Level: Combining Tissue with Living Cell Repositories to Promote Precision Medicine. *Front. Cel Dev. Biol.* 7, 246. doi:10.3389/fcell.2019.00246

Choi, Y. Y., Lee, J. E., Kim, H., Sim, M. H., Kim, K. K., Lee, G., et al. (2016). Establishment and Characterisation of Patient-Derived Xenografts as Paraclinical Models for Gastric Cancer. *Sci. Rep.* 6, 22172. doi:10.1038/srep22172

Duffy, M. J., Lamerz, R., Haglund, C., Nicolini, A., Kalousová, M., Holubec, L., et al. (2014). Tumor Markers in Colorectal Cancer, Gastric Cancer and Gastrointestinal Stromal Cancers: European Group on Tumor Markers 2014 Guidelines Update. *Int. J. Cancer* 134, 2513–2522. doi:10.1002/ijc.28384

Eguchi, T., Fujii, M., and Takayama, T. (2003). Mortality for Gastric Cancer in Elderly Patients. *J. Surg. Oncol.* 84, 132–136. doi:10.1002/jso.10303

Georges, L. M. C., De Wever, O., Galván, J. A., Dawson, H., Lugli, A., Demetter, P., et al. (2019). Cell Line Derived Xenograft Mouse Models Are a Suitable In Vivo Model for Studying Tumor Budding in Colorectal Cancer. *Front. Med.* 6, 139. doi:10.3389/fmed.2019.00139

Guo, T., Wen, X. Z., Li, Z. Y., Han, H. B., Zhang, C. G., Bai, Y. H., et al. (2019). ISL1 Predicts Poor Outcomes for Patients with Gastric Cancer and Drives Tumor Progression Through Binding to the ZEB1 Promoter Together with SETD7. *Cell Death Dis* 10, 33–14. doi:10.1038/s41419-018-1278-2

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M., et al. (2019). HumanNet V2: Human Gene Networks for Disease Research. *Nucleic Acids Res.* 47, D573–D580. doi:10.1093/nar/gky1126

Lallo, A., Schenk, M. W., Frese, K. K., Blackhall, F., and Dive, C. (2017). Circulating Tumor Cells and CDX Models as a Tool for Preclinical Drug Development. *Transl. Lung Cancer Res.* 6, 397–408. doi:10.21037/tlcr.2017.08.01

Lee, J. Y., Kim, S. Y., Park, C., Kim, N. K. D., Jang, J., Park, K., et al. (2015). Patient-derived Cell Models as Preclinical Tools for Genome-Directed Targeted Therapy. *Oncotarget* 6, 25619–25630. doi:10.18632/oncotarget.4627

Miwa, T., Kanda, M., Umeda, S., Tanaka, H., Shimizu, D., Tanaka, C., et al. (2019). Establishment of Peritoneal and Hepatic Metastasis Mouse Xenograft Models Using Gastric Cancer Cell Lines. *In Vivo* 33, 1785–1792. doi:10.21873/invivo.11669

Okada, S., Vaeteewoottacharn, K., and Kariya, R. (2018). Establishment of a Patient-Derived Tumor Xenograft Model and Application for Precision Cancer Medicine. *Chem. Pharm. Bull.* 66, 225–230. doi:10.1248/cpb.c17-00789

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48, D845–D855. doi:10.1093/nar/gkz1021

Rawla, P., and Barsouk, A. (2019). Epidemiology of Gastric Cancer: Global Trends, Risk Factors and Prevention. *pg* 14, 26–38. doi:10.5114/pg.2018.80001

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2020). Identifying Drug–Target Interactions Based on Graph Convolutional Network and Deep Neural Network. *Brief. Bioinform.* 22, bbaa044. doi:10.1093/bib/bbaa044

Wang, J., and Chen, S. (2002). Screening and Identification of Gastric Adenocarcinoma Metastasis-Related Genes by Using cDNA Microarray Coupled to FDD-PCR. *J. Cancer Res. Clin. Oncol.* 128. 547–553. doi:10.1007/s00432-002-0379-5

Wang, T., Hou, J., Jian, S., Luo, Q., Wei, J., Li, Z., et al. (2018). miR-29b Negatively Regulates MMP2 to Impact Gastric Cancer Development by Suppress Gastric Cancer Cell Migration and Tumor Growth. *J. Cancer* 9, 3776–3786. doi:10.7150/jca.26263

Zhang, T., Zhang, L., Fan, S., Zhang, M., Fu, H., Liu, Y., et al. (2015). Patient-derived Gastric Carcinoma Xenograft Mouse Models Faithfully Represent Human Tumor Molecular Diversity. *PLoS One* 10, e0134493. doi:10.1371/journal.pone.0134493

Zhao, T., Hu, Y., and Cheng, L. (2020). Deep-DRM: A Computational Method for Identifying Disease-Related Metabolites Based on Graph Deep Learning Approaches. *Brief. Bioinform.* 22, bbaa212. doi:10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: A Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36, 4466–4472. doi:10.1093/bioinformatics/btaa428

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021). Prediction and Collection of Protein–Metabolite Interactions. *Brief. Bioinform.* 22, bbab014. doi:10.1093/bib/bbab014

Zhu, M., Zhang, N., He, S., and Lu, X. (2020). Exosomal miR-106a Derived from Gastric Cancer Promotes Peritoneal Metastasis via Direct Regulation of Smad7. *Cell Cycle* 19, 1200–1221. doi:10.1080/15384101.2020.1749467

Zu, L. D., Peng, X. C., Zeng, Z., Wang, J. L., Meng, L. L., Shen, W. W., et al. (2018). Gastrin Inhibits Gastric Cancer Progression Through Activating the ERK-P65-

miR23a/27a/24 Axis. *J. Exp. Clin. Cancer Res.* 37, 115–118. doi:10.1186/s13046-018-0782-7

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

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

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 28 January 2022
doi: 10.3389/fmolb.2021.794326

Metastatic Immune-Related Genes for Affecting Prognosis and Immune Response in Renal Clear Cell Carcinoma

Si Sun[1,2†], *Weipu Mao*[1,3,4†], *Lilin Wan*[1,2†], *Kehao Pan*[1,2†], *Liting Deng*[2], *Lei Zhang*[1,3]*, *Guangyuan Zhang*[1,3]* and *Ming Chen*[1,3,4]*

[1]Department of Urology, Zhongda Hospital, Southeast University, Nanjing, China, [2]Medical School, Southeast University, Nanjing, China, [3]Surgical Research Center, Institute of Urology, Southeast University Medical School, Nanjing, China, [4]Department of Urology, Nanjing Lishui District People's Hospital, Zhongda Hospital Lishui Branch, Southeast University, Nanjing, China

Edited by:
Chang Xu,
Qatar University, Qatar

Reviewed by:
Sachin Kumar Deshmukh,
University of South Alabama,
United States
Rong Na,
Shanghai Jiao Tong University, China
Peng Wang,
Temple University, United States

***Correspondence:**
Ming Chen
mingchenseu@126.com
Guangyuan Zhang
zgy0879@qq.com
Lei Zhang
lei198711@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics,
a section of the journal
Frontiers in Molecular Biosciences

Received: 13 October 2021
Accepted: 28 December 2021
Published: 28 January 2022

Citation:
Sun S, Mao W, Wan L, Pan K, Deng L, Zhang L, Zhang G and Chen M (2022) Metastatic Immune-Related Genes for Affecting Prognosis and Immune Response in Renal Clear Cell Carcinoma.
Front. Mol. Biosci. 8:794326.
doi: 10.3389/fmolb.2021.794326

Background: In renal clear cell carcinoma, a common cancer of the urinary system, 25–30% patients are metastatic at initial diagnosis and 20–30% patients have a tendency of recurrence and metastasis after local surgery. With the rapid development of tumor immunology, immune agents have brought new directions to tumor therapy. However, no relevant studies have explored the role of immune-related genes in kidney cancer metastasis.

Methods: Co-expressed metastatic immune-related differentially expressed genes (mIR-DEGs) were screened by GSE12606, GSE47352, and immunorelated genes. Then, differential expression analysis, prognostic analysis, and univariate and multivariate Cox regression analysis in KIRC were performed to determine independent prognostic factors associated, and the risk prognostic model was established. The correlation of hub mIR-DEGs with clinicopathological factors, immune invasion, and immune checkpoints was analyzed, and the expression of hub mIR-DEGs and their effect on tumor were re-evaluated by OCLR scores in KIRC.

Results: By comprehensive bioassay, we found that FGF17, PRKCG, SSTR1, and SCTR were mIR-DEGs with independent prognostic values, which were significantly associated with clinicopathological factors and immune checkpoint–related genes. The risk prognostics model built on this basis had good predictive potential. In addition, targeted small molecule drugs, including calmidazolium and sulfasalazine, were predicted for mIR-DEGs. Further experimental results were consistent with the bioinformatics analysis.

Conclusion: This study preliminarily confirmed that FGF17, PRKCG, SSTR1, and SCTR were targeted genes affecting renal cancer metastasis and related immune responses and can be used as potential therapeutic targets and prognostic biomarkers for renal cancer. Preliminary validation found that PRKCG and SSTR1 were consistent with predictions.

Keywords: renal clear cell carcinoma, metastatic immune-related genes, prognosis, immunotherapy, biomarkers

INTRODUCTION

Renal cell carcinoma (RCC) is the most common renal malignancy originating from tubular epithelium (Siegel et al., 2018). Kidney renal clear cell carcinoma (KIRC) accounts for approximately 80% of all clinical cases of renal cell carcinoma in adults and is the most common histological subtype (Ricketts et al., 2018). In the 2021 Global Cancer Statistics, RCC accounted for approximately 4% of all newly diagnosed cancers, ranking sixth among cancers in men and ninth among cancers in women (Siegel et al., 2021). 25%–30% of patients are metastatic at initial diagnosis (Ljungberg et al., 2011), and 20–30% of patients tend to have recurrence and metastasis after local surgery (Athar and Gentile, 2008; Mao et al., 2021a). Due to resistance to radiation and chemotherapy (Braun et al., 2021), surgical resection is still the best treatment for RCC (Escudier et al., 2019).

In recent years, the treatment of RCC has made rapid progress. Much evidence has confirmed that RCC is highly immunogenic (Şenbabaoğlu et al., 2016) and is highly responsive to immunotherapy (Escudier, 2012). Among the most advanced therapies, immunotherapy can effectively and safely treat tumors (Xie et al., 2019; Frega et al., 2020). Its characteristic is to stimulate specific immune response and inhibit and kill tumor cells, thereby reducing tumor metastasis and recurrence. As an indispensable part of immunotherapy, the tumor immune microenvironment (TIME) has attracted more and more attention. The tumor is always in a complex tissue microenvironment, and the changes of immune microenvironment may affect the occurrence, development, and metastasis of tumor in different ways. The analysis of the immune microenvironment will help improve the response of immunotherapy. Some researchers have found that the TIME can be used as an important prognostic indicator, which could also enhance the potential of precision therapy (Taube et al., 2018; Vuong et al., 2019). Although the advent of immunotherapy and targeted therapy has diversified the treatment of RCC, some patients with RCC develop symptoms only when their cancer cells have metastasized to a distant point in their body, and the five-year survival rate of these patients is usually less than 20% (Dunnick, 2016). The prognosis for patients with renal cell carcinoma remains dismal. Therefore, it is urgent to search for targeted biomarkers related to metastasis and immunity in RCC.

In this study, the comprehensive bioinformatics analysis of GSE12606 (Stickel et al., 2009), GSE47352 (Gao et al., 2017), and immune-related genes was performed, and independent prognostic factors were identified by differential expression analysis, survival analysis, and univariate and multivariate Cox regression analysis, which contributed to renal cancer metastasis. The good prognostic risk model was constructed based on metastatic immune-related independent prognostic genes. In addition, we found that hub target genes were closely associated with the tumor immune microenvironment and immune checkpoint genes. Based on the target gene, we successfully predicted the potential therapeutic drugs to prevent renal cancer progression and assist immunotherapy. In conclusion, this study provided insights into immune-related molecular mechanisms underlying the progression of renal cancer from primary to metastatic stage and identified biomarkers that might have prognostic value.

MATERIALS AND METHODS

Screening of IR-DEGs in Primary and Metastatic KIRC

To acquire metastatic immune-related differentially expressed genes (mIR-DEGs) in primary and metastatic kidney renal clear cell carcinoma (KIRC), we used the GEO database (https://www.ncbi.nlm.nih.gov). The GSE12606 and GSE47352 datasets were selected for subsequent analysis (**Supplementary Table S1**). The cut-off conditions were set to *p*-value < 0.05, and the absolute value of log-fold change ($|\log_2 FC|$) ≥ 1, which had been adjusted for multiple testing *via* the Benjamini–Hochberg procedure, was statistically significant for the DEGs. We use ImageGP to create volcano maps and Venn maps online.

Functional Enrichment Analysis of mIR-DEGs

Enrichment analysis of mIR-DEGs was performed by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis in the "ClusterProfiler" package.

Identification of Independent Prognostic mIR-DEGs

Univariate and multivariate Cox regression analyses were performed on mIR-DEGs, and the forest maps were established by the "Forestplot" R package. The univariate Cox regression analysis result was included in the multivariate regression analysis when *p* threshold value < 0.05, and the independent prognostic genes were finally identified with $p <$ 0.005. RNA sequencing data of 539 renal clear cell carcinoma samples and 72 paracancerous samples, obtained from The Cancer Genome Atlas (TCGA) database (https://cancergenome.nih.gov/), were used to evaluate mIR-DEGs' expression and prognosis by Gene Expression Profiling Interactive Analysis (GEPIA) (http://gepia2.cancer-pku.cn/index) and "survival" package. The basic information of TCGA-KIRC patients is listed in **Supplementary Table S2**.

Construction and Validation of the Hub mIR-DEGs' Prognostic Model

Hub mIR-DEGs were selected based on univariate and multivariate Cox regression analysis, differential expression analysis, and prognostic analysis of mIR-DEGs. The lasso Cox regression was used to construct the risk prognosis model of hub mIR-DEGs based on the "GLMnet" R package. Risk coefficients were calculated by centralized standardized analysis in TCGA: Risk Score = ∑7iXi × Yi (X: coefficients, Y: gene expression level). Then, t-distributed stochastic neighbor embedding (t-SNE) and principal-component analysis (PCA) were used to explore the distribution characteristics of different groups by R packages. Finally, the effectiveness of prognostic indicators was evaluated by the area under the curve (AUC) of "time receiver-operating characteristic (ROC) curve." Furthermore, Spearman correlation analysis was used to explore the relationship between the model score and the immune score by QUANTISEQ, and the R software package Pheatmap was used to verify the relationship.

Construction of Clinicopathological Correlation Analysis and the Nomogram

Based on the "survival" package in R software, combined with the clinicopathological characteristics of patients (TNM stage, pathological stage, histologic stage, and laterality), the correlation between FGF17, PRKCG, SSTR1, and SCTR in the prognostic model and clinicopathological characteristics was analyzed. Through the R package "rms," the nomogram and calibration curve were obtained. Risk scores associated with prognostic models were used as prognostic factors to evaluate one-, three-, and five-year OS.

Assessment of the Immune Microenvironment About Hub mIR-DEGs in KIRC

The correlation between FGF17, PRKCG, SSTR1, and SCTR expressions and copy number and various immune cells in KIRC were searched and analyzed through the gene module by TIMER (https://cistrome.shinyapps.io/timer/), including B cells, CD4+ T cells, CD8+ T cells, macrophages, neutrophils, and dendritic cells.

In this study, kidney cancer immune cells were investigated by the QUANTISEQ1-2 algorithm, which quantifies tumor immune status based on human RNA-seq data, and the proportion of different immune cells and other uncharacterized cells present in the sample by a deconvolution algorithm, including B cells, macrophages, M2 macrophages, monocytes, neutrophils, NK cells, CD4+ T cells, CD8+ T cells, Tregs, myeloid cells, and uncharacterized cells (Finotello et al., 2019; Plattner et al., 2020).

Relationship Between Immune Checkpoint–Related Genes and Expression of Hub mIR-DEGs in KIRC

The relationship between SIGLEC15, TIGIT, CD274, HAVCR2, PDCD1, CTLA4, LAG3, and PDCD1LG2 and hub mIR-DEGs' expression was analyzed using the "ggplot2" R package. Subsequently, the Tumor Immune Dysfunction and Exclusion (TIDE) algorithm was used to evaluate the potential ICB response of different hub mIR-DEGs' expression levels to immune checkpoint inhibitors in KIRC.

OCLR Scores of Hub mIR-DEGs in KIRC

Tumor-associated RNA-seq data were obtained from TCGA-KIRC, mRNAsi was calculated by the OCLR algorithm, and the dryness index was obtained.

Prediction of Small Molecule Drugs for Hub mIR-DEGs

The hub mIR-DEGs were used for drug prediction in Connectivity Map (www.broadinstitute.org), which was commonly used to explore potential drugs for the treatment of diseases. Therefore, Enrichment > 0.7, $p < 0.02$, and Percent non-nulld >75 were used for screening. The PubChem22 database (www.pubchem.ncbi.nlm.nih.gov) was used to retrieve the molecular structure of identified drugs.

Cell Lines, Patient Samples, RNA Extraction, and Quantitative Real-Time Polymerase Chain Reaction (qRT-PCR)

The human kidney cell line, HK-2, and human KIRC cell lines, 786-O and caki-1, were originally purchased from the cell repository of Shanghai Institute of Life Sciences. The cells were cultured in 1640 medium (GIBCO), containing 10% FBS (GIBCO), penicillin (25 U/ml), and streptomycin (25 mg/ml), and at 5% CO_2 environment.

In this study, 19 fresh samples, including tumor tissue and adjacent normal kidney tissue, were collected from patients who underwent laparoscopic radical nephrectomy for KIRC from 2019 to 2020 in the Department of Urology, Zhongda Hospital, and stored at 80°C. All patients were diagnosed with KIRC and did not receive any antitumor therapy preoperatively. Clinical characteristics of 19 KIRC patients are listed in **Supplementary Table S3**. The methodology of this study followed the criteria outlined in the Helsinki Declaration (revised in 2013), and ethical approval was obtained from the Ethics Committee and Institutional Review Board for Clinical Research of Zhongda Hospital (ZDKYSB077). All patients or their relatives who participated were informed and signed an informed consent form.

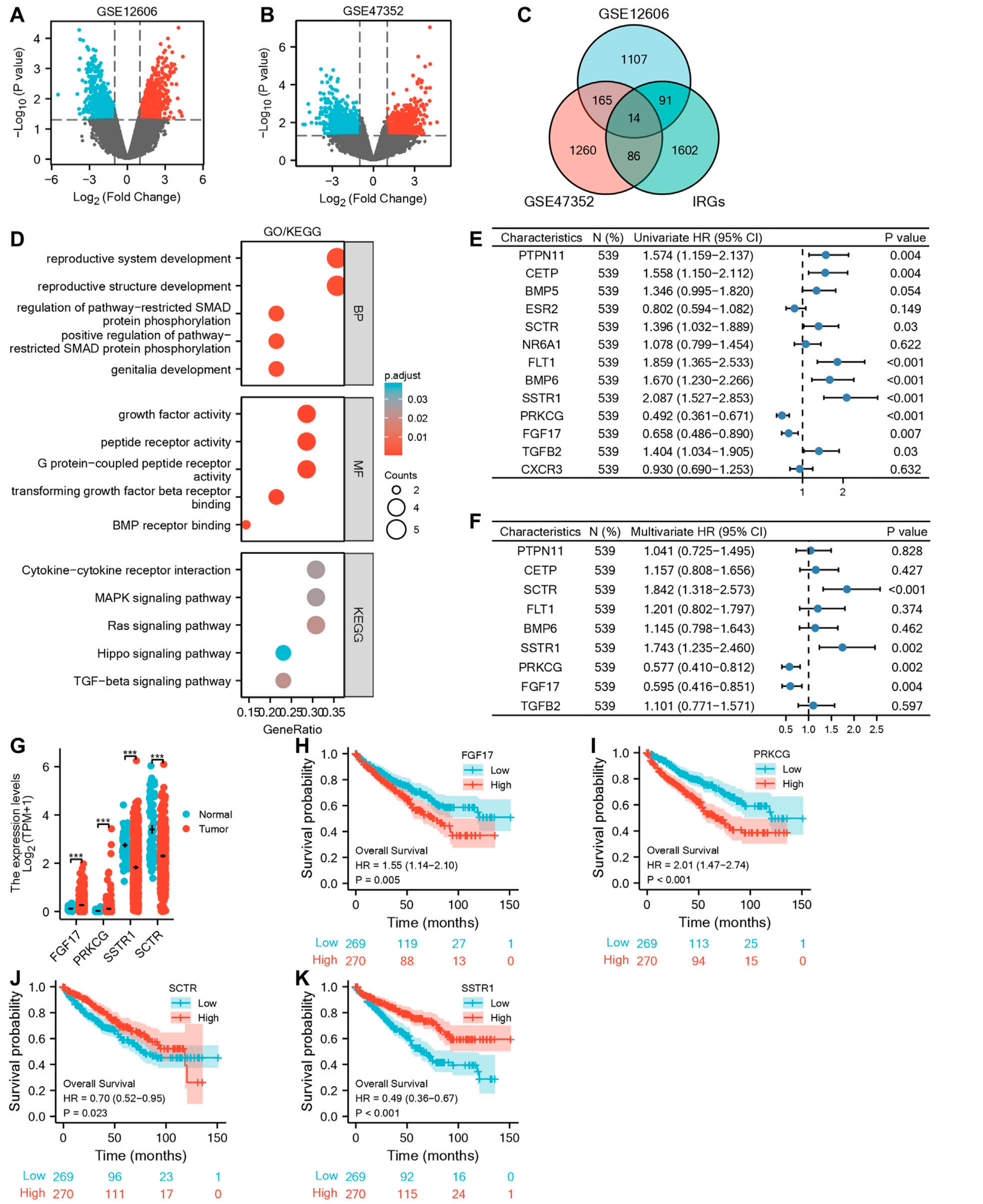
Total RNA was isolated with Total RNA Kit (OMEGAbiotec, Guangzhou, China) according to the manufacturer's instructions. Complementary DNA was synthesized using the HiScript II Q RT SuperMix (R223-01) reagent kit (Vazyme Biotech Co., Ltd., Nanjing, China). The qRT-PCR was performed using the SYBR Green PCR Mix (Vazyme Biotech Co., Ltd., Nanjing, China). The specific primers set for mIR-DEGs and GAPDH are listed in **Supplementary Table S4**. Data were normalized to GAPDH expression levels using the $2^{-\Delta\Delta Ct}$ method.

Tissue Microarray Construction and Immunohistochemistry

All specimens were fixed in 10% neutral formaldehyde solution and embedded in paraffin. Envision two-step dyeing and DAB color development were used. The primary antibody (FGF17, ab187982, Abcam; PRKCG, ab181558, Abcam; SSTR1, ab140945, Abcam) was used in this study.

Statistical Analysis

The statistical analysis was carried out by R software (version 4.0.2). The Perl programming language (version 5.30.2) was used



Characteristics	N (%)	Univariate HR (95% CI)	P value
PTPN11	539	1.574 (1.159−2.137)	0.004
CETP	539	1.558 (1.150−2.112)	0.004
BMP5	539	1.346 (0.995−1.820)	0.054
ESR2	539	0.802 (0.594−1.082)	0.149
SCTR	539	1.396 (1.032−1.889)	0.03
NR6A1	539	1.078 (0.799−1.454)	0.622
FLT1	539	1.859 (1.365−2.533)	<0.001
BMP6	539	1.670 (1.230−2.266)	<0.001
SSTR1	539	2.087 (1.527−2.853)	<0.001
PRKCG	539	0.492 (0.361−0.671)	<0.001
FGF17	539	0.658 (0.486−0.890)	0.007
TGFB2	539	1.404 (1.034−1.905)	0.03
CXCR3	539	0.930 (0.690−1.253)	0.632

Characteristics	N (%)	Multivariate HR (95% CI)	P value
PTPN11	539	1.041 (0.725−1.495)	0.828
CETP	539	1.157 (0.808−1.656)	0.427
SCTR	539	1.842 (1.318−2.573)	<0.001
FLT1	539	1.201 (0.802−1.797)	0.374
BMP6	539	1.145 (0.798−1.643)	0.462
SSTR1	539	1.743 (1.235−2.460)	0.002
PRKCG	539	0.577 (0.410−0.812)	0.002
FGF17	539	0.595 (0.416−0.851)	0.004
TGFB2	539	1.101 (0.771−1.571)	0.597

	0	50	100	150
Low	269	119	27	1
High	270	88	13	0

	0	50	100	150
Low	269	113	25	1
High	270	94	15	0

	0	50	100	150
Low	269	96	23	1
High	270	111	17	0

	0	50	100	150
Low	269	92	16	0
High	270	115	24	1

FIGURE 1 | Screening for independent prognostic genes in KIRC. **(A,B)** Volcano maps of GSE12606 and GSE47352. **(C)** Venn diagram of GSE12606, GSE47352, and immune-related genes. **(D)** GO|KEGG enrichment analysis of IR-DEGs. **(E,F)** Forest plots of univariate and multivariate Cox regression analysis of IR-DEGs. **(G)** Based on the GEPIA database for differential expression of the four IR-DEGs. **(H–K)** Survival analysis of four IR-DEGs, including SCTR **(H)**, SSTR1 **(I)**, PRKCG **(J)**, and FGF17 **(K)**.

for data processing. Multivariate Cox regression analyses were used to evaluate prognostic significance. When $p < 0.05$ or log-rank $p < 0.05$, the difference was statistically significant.

RESULTS

Identification of mIR-DEGs

Sequencing data related to primary and metastatic renal carcinoma were obtained from the GEO database (**Figures 1A,B**). 1,377 metastatic DEGs (mDEGs) were screened in GSE12606, and 1,525 mDEGs were screened in GSE47352. Then, the mDEGs and 1,793 immune-related genes, which are from the ImmPort database, were analyzed by Venn diagram, and 14 co-expressed genes were obtained by the intersection of the three gene sets (**Figure 1C**). Then, through GO/KEGG pathway enrichment analysis (**Supplementary Table S5**), it was found that the functions of 14 mIR-DEGs were mainly concentrated in "reproductive structure development," "reproductive system development," "positive regulation of pathway-restricted SMAD protein phosphorylation," "G protein-coupled peptide receptor activity," "peptide receptor activity," "growth factor activity," "TGF-beta signaling pathway," "MAPK signaling pathway," and "Cytokine-cytokine receptor interaction" (**Figure 1D**). To further clarify the correlation between mIR-DEGs and prognosis, univariate and multivariate Cox regression analysis (**Supplementary Table S6**) showed that FGF17, PRKCG, SSTR1, and SCTR were independent prognostic factors for the progression of KIRC from primary to metastatic stage (**Figures 1E,F**). Consequently, after screening hub mIR-DEGs with stringent criteria, the results conform to the Bonferroni correction significant level and minimize the inflation of Type I errors from multiple testing issues.

Differential Expression Analysis and Survival Analysis of Hub mIR-DEGs in KIRC

Using the TCGA-KIRC database, we verified the expression levels of four mIR-DEGs that were significant in univariate Cox regression analysis and found the expression levels of PRKCG and FGF17 were up-regulated and SCTR and SSTR1 were down-regulated in 539 tumors and 72 paracancerous samples (**Figure 1G**). Then, Kaplan–Meier model analysis showed that the above four mIR-DEGs were significantly associated with prognosis, and the high expressions of SCTR and SSTR1 and TGFB2 were associated with good prognosis (**Figures 1J,K**), while the high expressions of PRKCG and FGF17 were significantly associated with poor prognosis (**Figures 1H,I**). Combined with multivariate Cox regression analysis, FGF17, PRKCG, SSTR1, and SCTR were identified as the hub metastatic immune-related independent prognostic factors, which influenced the progression of primary to metastatic kidney cancer.

Construction and Validation of the Hub mIR-DEGs' Prognostic Risk Model

Based on hub mIR-DEGs, lasso Cox regression was used to construct relevant risk prognosis models, lambda.min = 0.0103, Risk Score= $(-0.1637) \times \text{SCTR} + (-0.2632) \times \text{SSTR1} + (0.1711) \times \text{PRKCG} + (0.7824) \times \text{FGF17}$ (**Figures 2A,B**). Patients were assigned into high-risk and low-risk groups according to the median risk score (50%). Survival status and hub mIR-DEGs' heatmaps in different groups were displayed by t-SNE and PCA, indicating that FGF17 and PRKCG were highly expressed in the high-risk group, while SSTR1 and SCTR were lowly expressed in the high-risk group (**Figure 2C**). The prognostic model was the risk factor model due to HR = 2.445, and the median survival time of the high-risk group was significantly shorter than that of the low-risk group (**Figure 2C**). Finally, we evaluated the prognostic prediction efficiency of the model by the ROC curve. We found that the AUC was 0.71 (one-year OS), 0.673 (three-year OS), and 0.711 (five-year OS), respectively (**Figure 2C**). In addition, Spearman correlation analysis was used to explore the correlation between the hub mIR-DEGs risk prognosis model and the tumor immune microenvironment in KIRC (**Figures 3A–K**). The risk prognosis model was significantly negatively correlated with the infiltration of M2 macrophages ($r = -0.12$, $p = 0.004$), neutrophils ($r = -0.40$, $p = 1.97\text{e-}21$), CD4+ T cells ($r = -0.26$, $p = 0.1.37\text{e-}09$), and myeloid dendritic cells ($r = -0.25$, $p = 3.91\text{e-}09$) (**Figures 3C,E,G,J**) and significantly positively correlated with the infiltration of monocytes ($r = 0.22$, $p = 4.88\text{e-}07$) and uncharacterized cells ($r = 0.23$, $p = 4.62\text{e-}08$) (**Figures 3D,K**). These results indicated that the hub mIR-DEG–based risk prognosis model, including FGF17, PRKCG, SSTR1, and SCTR, had good predictive effect and was significantly correlated with the KIRC immune microenvironment.

Relationship Between Hub mIR-DEGs and Clinicopathological Factors and the Construction Nomogram

We analyzed the correlation of FGF17, PRKCG, SSTR1, and SCTR in the risk prognosis model with clinicopathological features. The results showed that the expression of PRKCG and SSTR1 was correlated with T stage (**Figure 4A**), PRKCG was correlated with N stage (**Figures 4B,C**), and the expression of PRKCG, SSTR1, and SCTR was associated with M stage, pathologic stage, and histologic stage (**Figures 4D–F**). One-, three-, and five-year OS was predicted by the nomogram, and the potential value of M stage for prognosis was determined in KIRC patients (**Figure 4G**). Subsequently, time-dependent ROC curve analysis showed that AUCFGF17 = 0.627, AUCPRKCG = 0.694, AUCSSTR1 = 0.758, and AUCSCTR = 0.737, indicating a good prognostic value of hub mIR-DEGs for KIRC patients (**Figure 4H**). In addition, we find that the calibration curve of the predicted probability was in good agreement with the one-, three-, and five-year OS on the nomogram, and the three-year OS was the best fit (**Figures 4I–K**).

Assessment of the Immune Microenvironment About Hub mIR-DEGs in KIRC

In order to explore the potential relationship between the expression of FGF17, PRKCG, SSTR1, and SCTR in KIRC and

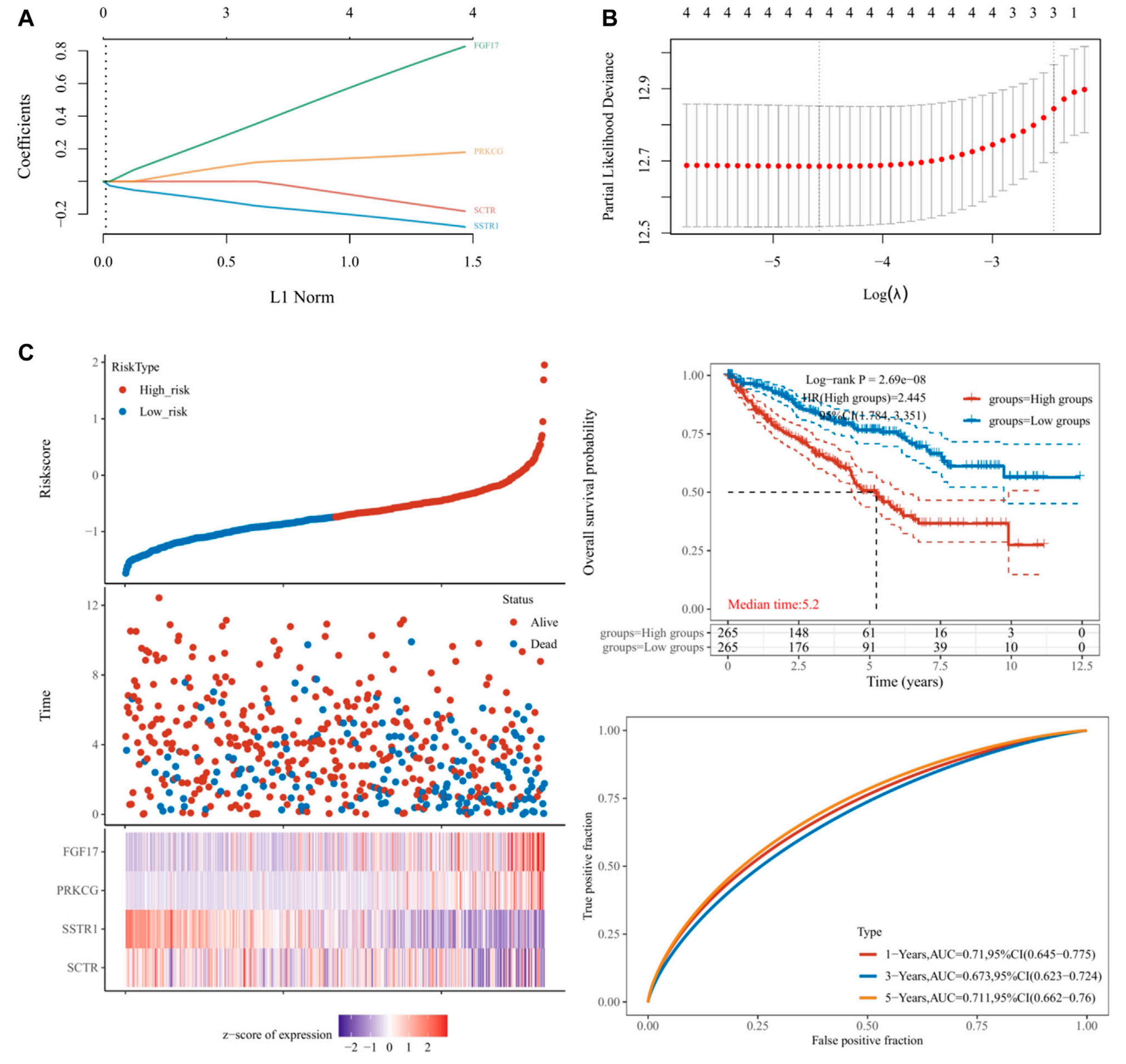


FIGURE 2 | Establishment and validation of prognostic models in KIRC. **(A,B)** Lasso regression analysis results. **(C)** Risk score distribution, survival status, and four hub IR-DEGs in low- and high-risk groups. Kaplan–Meier survival curve of two groups. Time-dependent ROC curve analyses in TCGA set.

the level of immune invasion, TIMER was used to conduct correlation analysis. First, we found positive correlations between SCTR and CD4+ T cells (R = 0.11, p = 1.88e-02). SSTR1 and CD8+ T cells (R = 0.188, p = 7.77e-05), CD4+ T cells (R = 0.172, p = 2.14e-04), macrophages (R = 0.208, p = 9.33e-06), neutrophils (R = 0.129, p = 5.70e-03), and DCs (R = 0.127, p = 6.88e-03) were positively correlated; PRKCG was positively correlated with CD4+ T cells (R = 0.209, p = 6.40e-06) and neutrophils (R = 0.102, p = 2.95e-02). FGF17 was positively correlated with CD4+ T cells (R = 0.262, p = 1.14e-08) but negatively correlated with B cells (R = −0.200, p = 1.60E-05) and DCs (R = −0.168, p = 3.08E-04) (**Figures 5E–H**). The copy numbers of SCTR, SSTR1, and PRKCG were significantly correlated with the infiltration levels of B cells, CD8+ T cells, CD4+ T cells, macrophages, neutrophils, and DCs (**Figures 5A–C**). However, FGF17 was only associated with $CD8^+$ T cells, neutrophils, and DCs (**Figure 5D**).

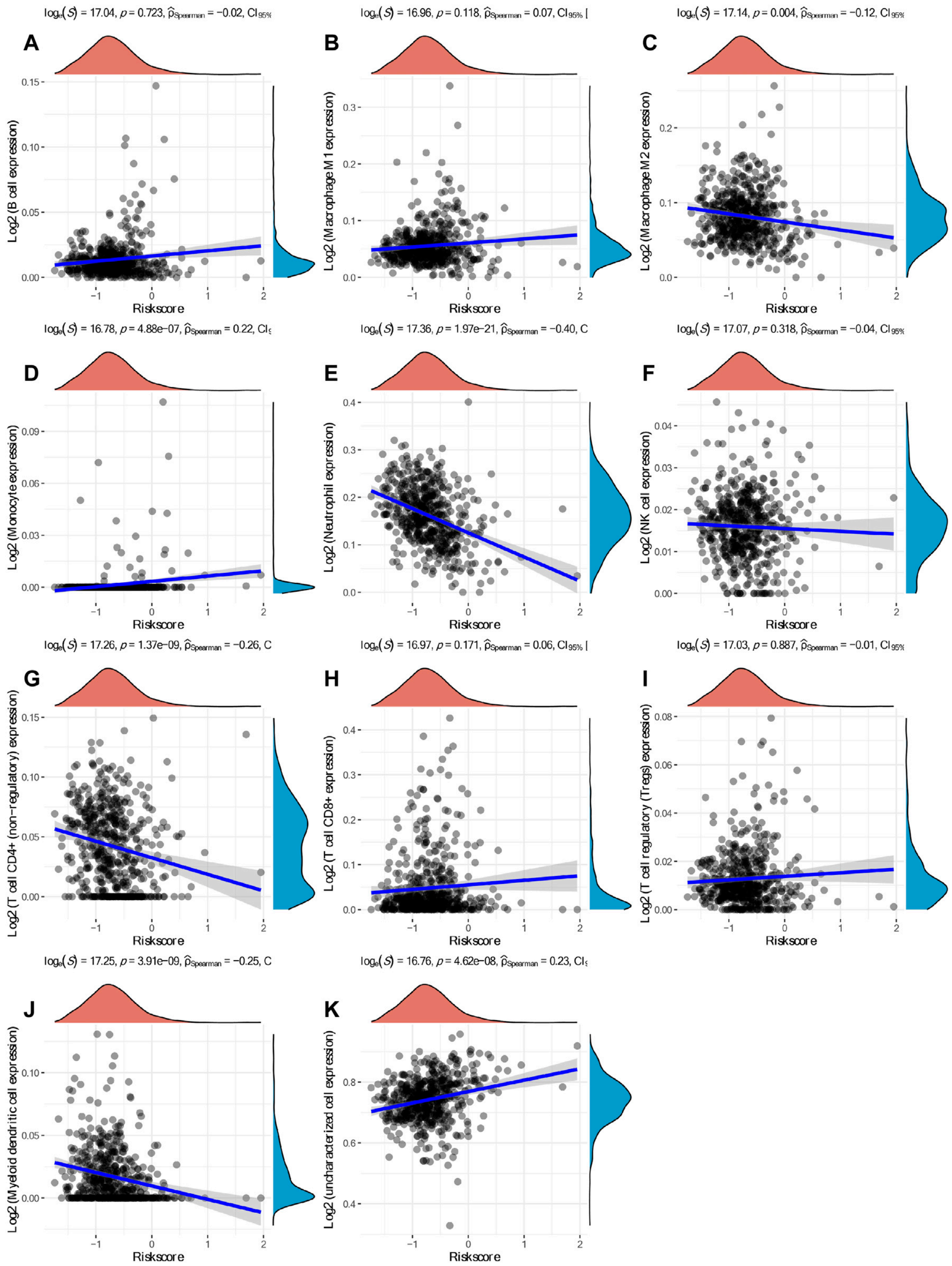


FIGURE 3 | Spearman correlation analysis between the model score and the immune score. **(A)** B cells. **(B)** M1 macrophages. **(C)** M2 macrophages. **(D)** Monocytes. **(E)** Neutrophils. **(F)** NK cells. **(G)** CD4+ T cells. **(H)** CD8+ T cells. **(I)** Tregs. **(J)** Myeloid dendritic cells. **(K)** Uncharacterized cells.

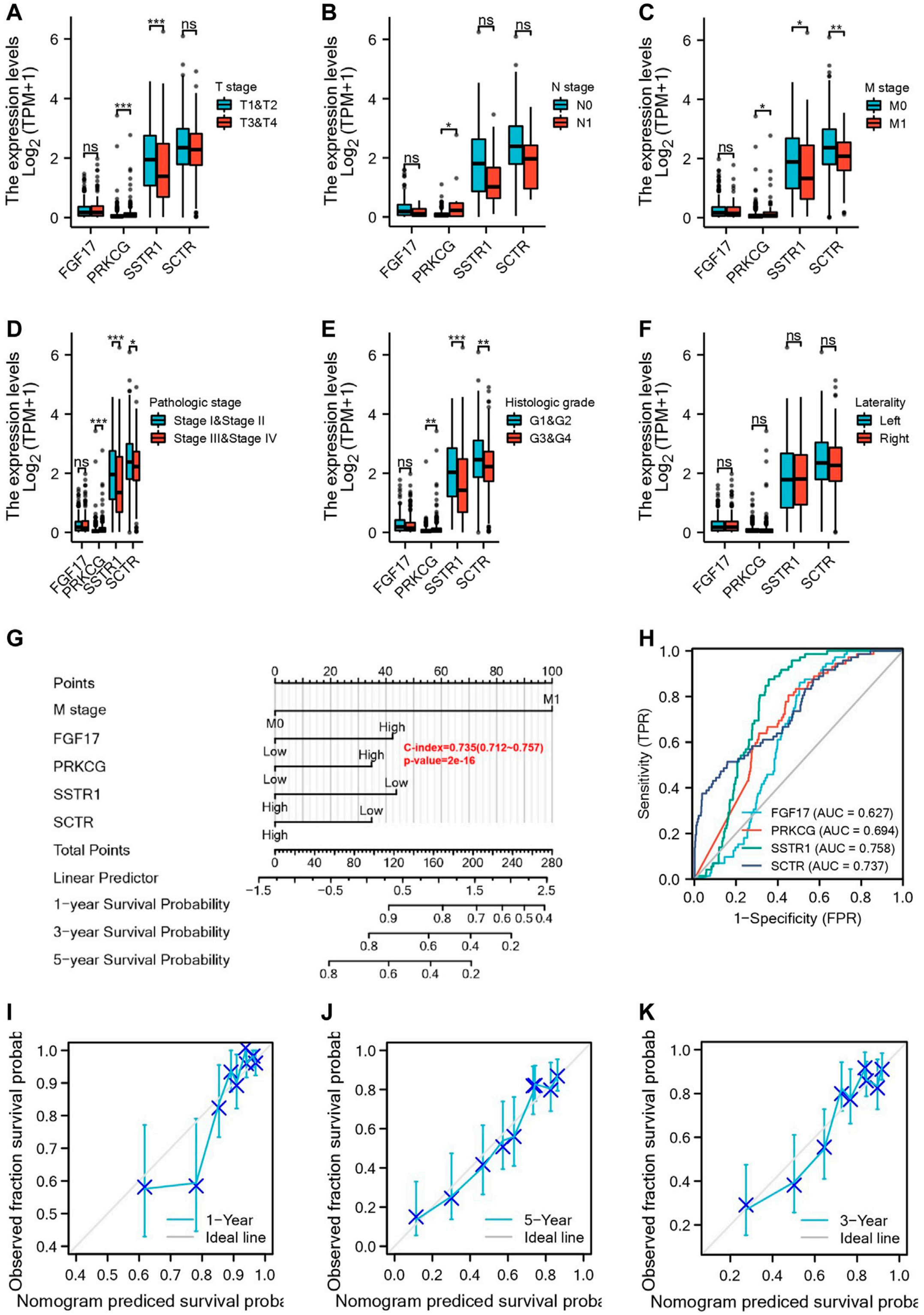


FIGURE 4 | Four IR-DEGs correlate with multiple clinicopathological factors in KIRC. Relationships between IR-DEGs and clinicopathological factors in the entire TCGA cohort, including T stage **(A)**, N stage **(B)**, M stage **(C)**, histologic grade **(D)**, pathologic stage **(E)**, and laterality **(F)**. **(G)** Nomogram for predicting one-, three-, and five-year OS in the entire TCGA cohort. **(H–K)** Calibration curves of nomogram on consistency between predicted and observed one-, three-, and five-year survival in the entire TCGA cohort. The dashed line at 45° implies a perfect prediction, and the actual performances of our nomogram are shown by blue lines.

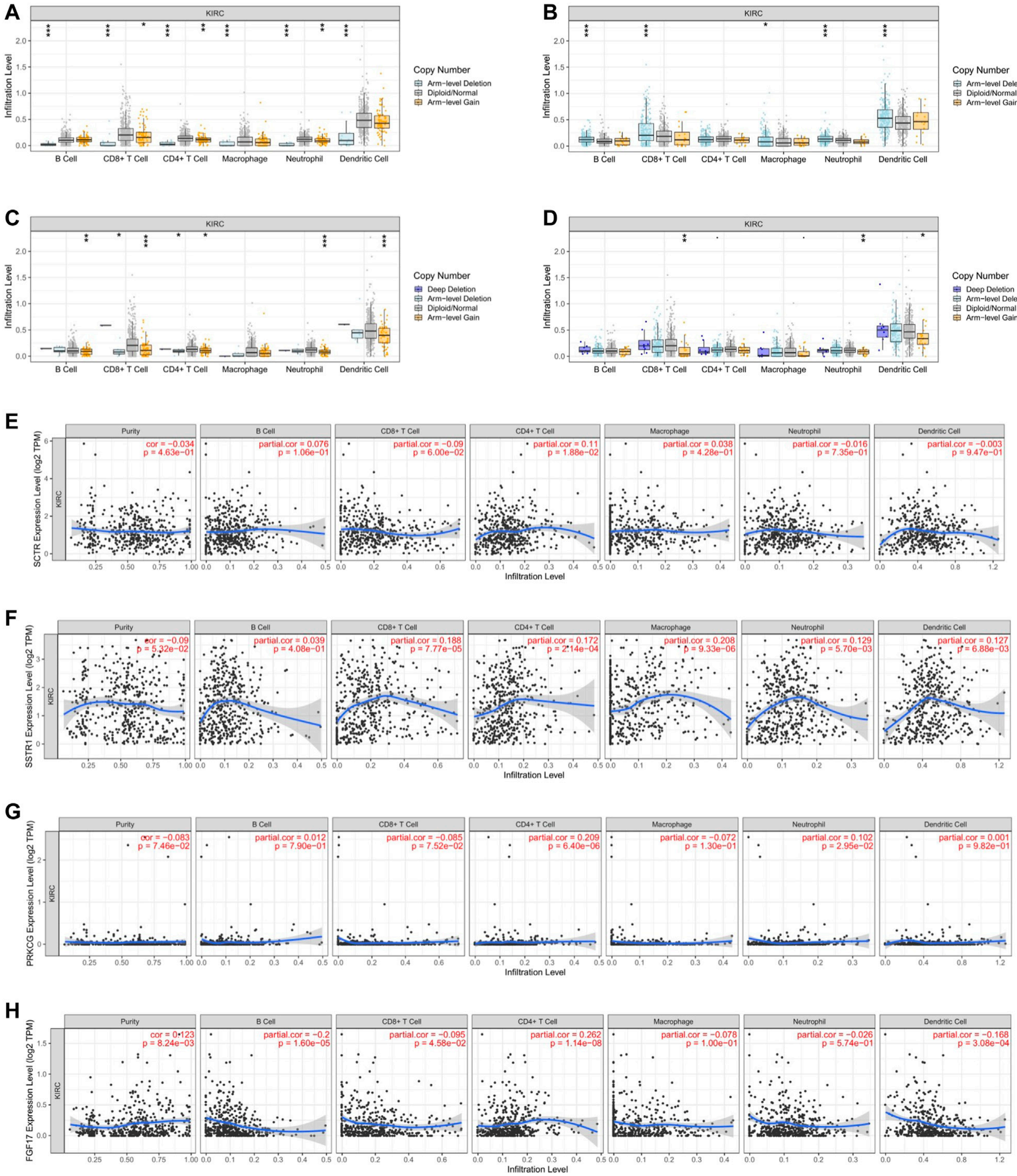


FIGURE 5 | Relationship of immune cell infiltration with IR-DEG levels in KIRC. Infiltration level of various immune cells under different copy numbers of IR-DEG levels, including SCTR **(A)**, SSTR1 **(B)**, PRKCG **(C)**, and FGF17 **(D)**. Correlation of IR-DEG expression levels with B cell, CD8+ T cell, CD4+ T cell, macrophage, neutrophil, and dendritic cell infiltration levels, including SCTR **(E)**, SSTR1 **(F)**, PRKCG **(G)**, and FGF17 **(H)**.

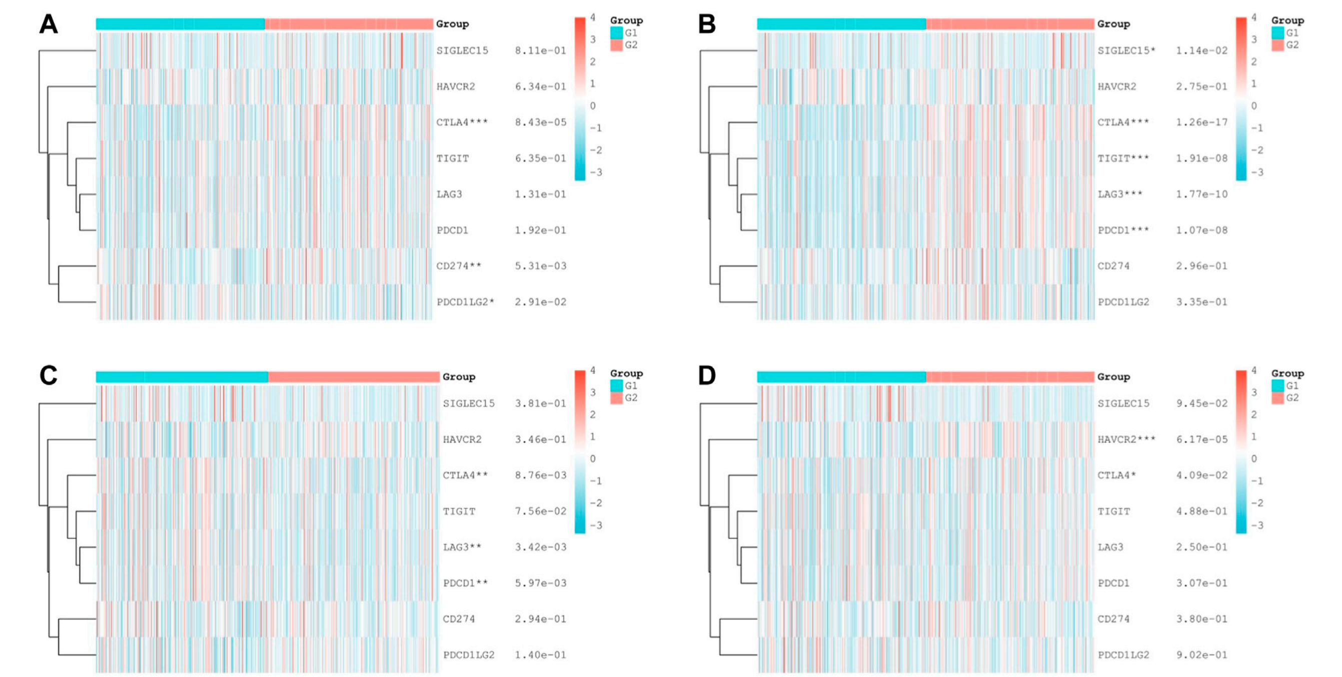


FIGURE 6 | Differential expression of immune checkpoint—related genes in KIRC tissues. **(A–D)** Comparison of immune checkpoints in different expression levels of IR-DEGs and M stage in KIRC, including FGF17 **(A)**, PRKCG **(B)**, SSTR1 **(C)**, and SCTR **(D)**. G1 is IR-DEG upexpression in non-metastatic KIRC. G2 is IR-DEG downexpression in non-metastatic KIRC.

Relationship Between Immune Checkpoint–Related Genes and Expression of Hub mIR-DEGs

Based on the apparent correlation between hub mIR-DEGs, risk prediction models, and tumor immune microenvironment, we further explored the relationship between hub mIR-DEGs and immune checkpoints, providing potential directions for future immunotherapy (**Figure 6**). We found significant differences between FGF17 and CTLA4, CD274, and PDCD1LG2 (**Figure 6A**). PRKCG was significantly different from SIGLEC15, CTLA4, TIGIT, LAG3, and PDCD1 (**Figure 6B**). SSTR1 was significantly different from CTLA4, LAG3, and PDCD1 (**Figure 6C**). SCTR was significantly different from HAVCR2 and CTLA4 (**Figure 6D**). CTLA4 was strongly correlated with four hub mIR-DEGs. Our results suggested that CTLA4 might be a potential target for preventing KIRC progression and metastasis through immune checkpoint inhibitors in the risk prognosis model.

Assessment of the OCLR Scores of Hub mIR-DEGs in KIRC

Through the dryness index, we discovered significant differences in dryness degree between hub mIR-DEGs in KIRC (**Figure 7**). These results suggested that FGF17, PRKCG, SSTR1, and SCTR might affect the degree of similarity between KIRC cells and stem cells, thus affecting tumor biological processes and degree of dedifferentiation.

Prediction of Small Molecule Drugs for Hub mIR-DEGs

Based on the former analysis we performed, we can propose an assumption that FGF17, PRKCG, SSTR1, and SCTR had a potential role in the progression and metastasis of KIRC. Therefore, based on probes of FGF17 (221376_at), PRKCG (206270_at), SSTR1 (208482_at), and SCTR (210382_at), we predicted potential targeted drugs with immunotherapeutic effects and prevention of KIRC metastasis through Connectivity Map (**Figure 8A**). The structural formula and molecular formula of targeted drugs with the most potential value were obtained through PubChem22, including 5224221, calmidazolium, sulfasalazine, carbenoxolone, and tribenoside (**Figures 8B–F**).

Validation of the Expression of mIR-DEGs in Clinical Tissue Samples

To detect the expression of four genes (FGF17, PRKCG, SSTR1, and SCTR) in KIRC, we performed the qRT-PCR in KIRC cells and clinical tissue samples. We verified the expression levels of four genes in the normal kidney cell line (HK-2 cells) and two KIRC cell lines (786-O, caki-1). The results showed that the expression levels of FGF17 and PRKCG were significantly increased in KIRC cells compared with normal kidney cells, while SSTR1 and SCTR were down-regulated in KIRC cells (**Figures 9A–D**). FGF17, PRKCG, and

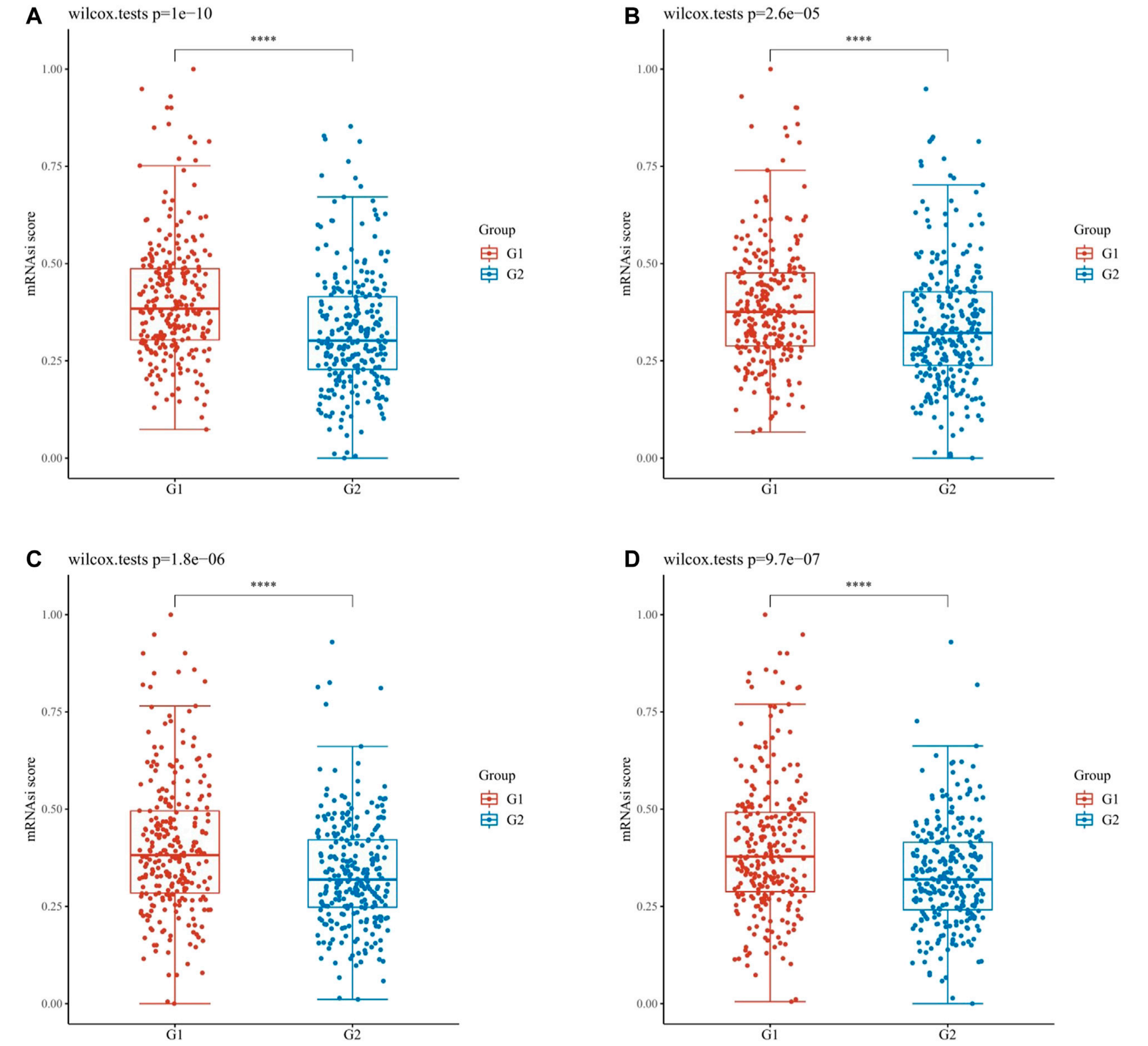
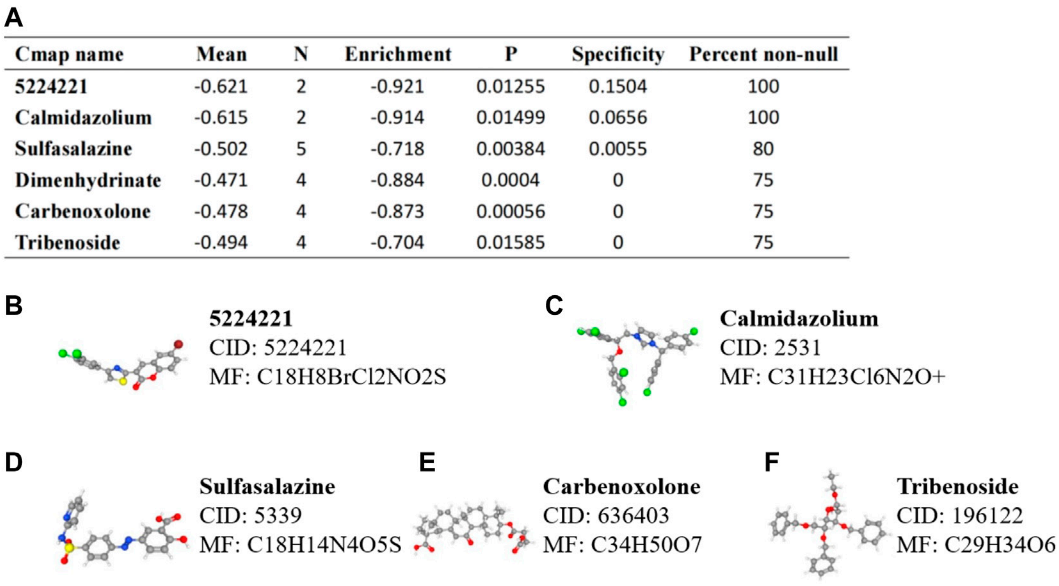


FIGURE 7 | OCLR scores of hub IR-DEGs in KIRC. OCLR scores of hub IR-DEGs at different expression levels in KIRC, including FGF17 **(A)**, PRKCG **(B)**, SSTR1 **(C)**, and SCTR **(D)**. G1 is IR-DEG down-expression and G2 is up-expression in KIRC.

SSTR1 were detected with the same results in tumor tissues and adjacent normal kidney tissues, while SCTR was not significantly different (**Figures 9E–H**). Then, we detected the protein expression of FGF17, PRKCG, and SSTR1 in the tissues by IHC. The IHC results showed that PRKCG was strongly expressed in the cytoplasm of KIRC tissues compared with adjacent normal kidney tissues. The expression of SSTR1, which was mainly expressed in the cytosol and cytoplasm, was significantly decreased. FGF17 positive expression was mainly distributed extracellularly, but FGF17 was negative in most tissues (**Figure 9I**).

DISCUSSION

Renal cell carcinoma was one of the most common urinary system tumors; about 25–30% of patients were metastatic at initial diagnosis, and 20–30% of patients had a tendency of recurrence and metastasis after local surgery, especially ccRCC (Jung et al., 2001). Many studies had shown that, in mRCC, the top three metastases were lung (45–75%), bone (15–34%), and liver (20%), whose five-year survival rates were 36–50%, 35%, and 18–43%, respectively (Staehler, 2011; Hatzaras et al., 2012). Given the rapid development of tumor immunology, a large number of



Cmap name	Mean	N	Enrichment	P	Specificity	Percent non-null
5224221	-0.621	2	-0.921	0.01255	0.1504	100
Calmidazolium	-0.615	2	-0.914	0.01499	0.0656	100
Sulfasalazine	-0.502	5	-0.718	0.00384	0.0055	80
Dimenhydrinate	-0.471	4	-0.884	0.0004	0	75
Carbenoxolone	-0.478	4	-0.873	0.00056	0	75
Tribenoside	-0.494	4	-0.704	0.01585	0	75

FIGURE 8 | Prediction of small molecule drugs targeting IR-DEGs in KIRC. **(A)** mRNA probes were used to predict potential drugs for KIRC. **(B–F)** Prediction results of targeted drugs, including 5224221 **(B)**, calmidazolium **(C)**, sulfasalazine **(D)**, carbenoxolone **(E)**, and tribenoside **(F)**.

previous studies had found that traditional immunotherapy, such as IFN-α and IL-2, could extend the OS to a certain extent, but their response duration was limited, and only a few patients could fully respond (Floros and Tarhini, 2015; Mao et al., 2021b). Currently, new immunotherapy drugs have been developed successively, such as cancer vaccine (Amin et al., 2015), adoptive cell therapy (Tang et al., 2013), and checkpoint inhibitors (Ghatalia et al., 2017). These drugs were reported to be capable of prolonging the response time of combination drugs and improving the OS significantly. Therefore, it was of great significance to elucidate the molecular mechanism of immune-related invasion and metastasis of RCC and to identify potential biomarkers for immunotherapy in RCC.

In this study, we firstly screened in GSE12606, GSE47352, and immune-related genes to analyze the co-expression of differential genes in primary and metastatic renal carcinoma. Secondly, FGF17, PRKCG, SSTR1, and SCTR were identified as metastatic immune-related independent risk factors by differential expression analysis, prognostic analysis, and univariate and multivariate Cox regression analysis. Then, the risk prognostic model was constructed based on lasso regression analysis, that is, Risk Score= (−0.1637) × SCTR + (−0.2632) × SSTR1 + (0.1711) PRKCG + (0.7824) FGF17. The predictive value of this model was favorable. There were significant correlations between the expression levels of four mIR-DEGs and clinicopathological factors, immune infiltration, and immune checkpoint. In addition, the calibration curves and nomogram showed an excellent prediction effect. Subsequently, through OCLR scores, it was further confirmed that the expressions of FGF17, PRKCG, SSTR1, and SCTR were different in KIRC, which might lead to tumor metastasis by promoting tumor dedifferentiation. Therefore, all of these results preliminary indicate that FGF17, PRKCG, SSTR1, and SCTR may impact the progression and metastasis in KIRC. Furthermore, their significant association with KIRC immune microenvironment and immune checkpoint–related genes also implied that mIR-DEGs may be potential targets and prognostic biomarkers for KIRC immunotherapy.

FGF17, as a member of the fibroblast growth factor (FGF) family, was located at 8p21.3 and played a significant role in the occurrence and progression of cancer (Tabarés-Seisdedos and Rubenstein, 2009). Studies had shown that the dual inhibition of FGF and CSF1 or VEGF signals was expected to enhance the antitumor effect by targeting immune escape and angiogenesis in the tumor microenvironment (Katoh, 2016). Protein kinase C gamma (PRKCG), as an isoenzyme of protein kinase Cs (PKCs) (Nishizuka, 1984), mediates IL-2 expression and tumor immune response (Chen et al., 1994). The 20th serine site could also be phosphorylated in p53 to activate apoptosis of colon cancer cells (Kawabata et al., 2012). Somatostatin receptor 1 (SSTR1) was a subtype of SSTR, belonging to the G-protein–coupled receptor (GPCR), which was involved in various signal transduction mechanisms in different parts of the human body (Nagarajan et al., 2020). Studies had found abnormal expression of SSTR in prostate cancer, colorectal cancer, breast cancer, and leiomyoma (Reubi et al., 1998), and high expression of SSTR1 could reduce the proliferation of acetaldehyde dehydrogenase (ALDH) positive cells, resulting in silenced and proliferation inhibition of colon cancer stem cells (Zou Y et al., 2019). Therefore, somatostatin analogs (SSAs) had been studied for immunotherapy of various cancers (Li et al., 2005). Secretin receptor (SCTR), also known as GPCR, was abnormally expressed in many cancers to affect the proliferation of tumor cells (Awasthi et al., 2012). Low expression of SCTR could stimulate tumor cell proliferation through the PI3K/AKT signaling pathway (Lee et al., 2012), and the combination of PI3K inhibitors and tumor chemoradiotherapy

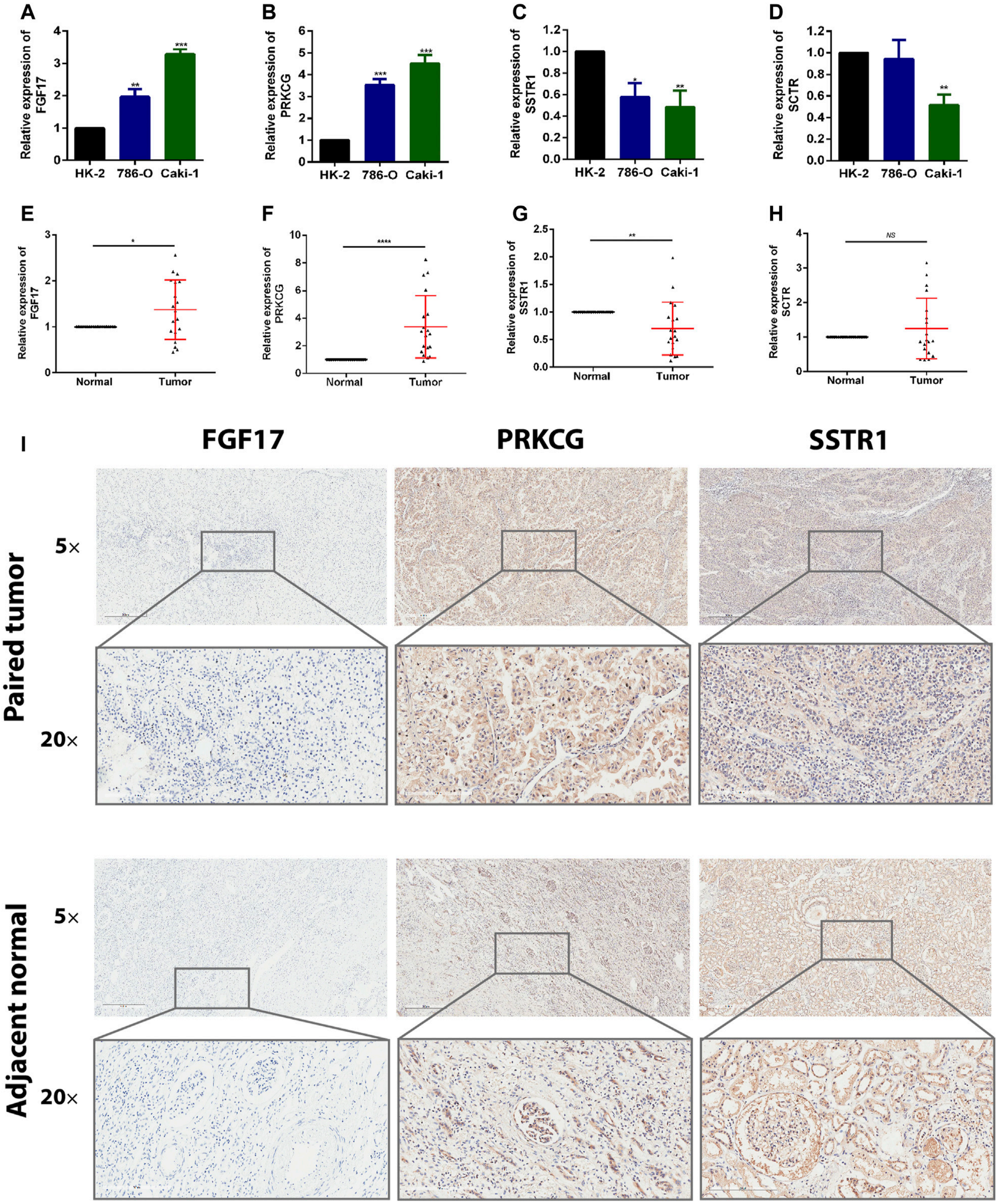

A
Relative expression of FGF17
HK-2
786-O
Caki-1
B
Relative expression of PRKCG
C
Relative expression of SSTR1
D
Relative expression of SCTR
E
F
G
H
Normal
Tumor
NS
I
FGF17
PRKCG
SSTR1
Paired tumor
Adjacent normal
5×
20×

FIGURE 9 | Expression of these hub genes in human KIRC specimens, adjacent normal tissues, and cell lines. **(A–D)** qRT-PCR analysis of FGF17 **(A)**, PRKCG **(B)**, SSTR1 **(C)**, and SCTR **(D)** in KIRC cell lines. GAPDH was used as a loading control. **(E–H)** qRT-PCR analysis of FGF17 **(E)**, PRKCG **(F)**, SSTR1 **(G)**, and SCTR **(H)** in paired KIRC tissues ($n = 19$). **(I)** Representative images of FGF17, PRKCG, and SSTR1 protein immunochemistry in KIRC tissues compared with adjacent normal kidney tissues. Magnification: ×50, ×200; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

had been shown to inhibit tumor proliferation. In summary, we found significant differences in the expression of FGF17, PRKCG, SSTR1, and SCTR in cancer, which are correlated with immune response and adjuvant therapy. However, the specific functions and potential mechanisms of these four immune-related genes in KIRC metastasis remained unclear and needed further exploration. In addition, we performed qRT-PCR analysis on clinical specimens and found that the mRNA expression levels of FGF17, PRKCG, and SSTR1 were significantly different between kidney cancer tissues and normal tissues adjacent to the cancer. However, more *in vivo* and *in vitro* experiments are needed to confirm these findings.

Interestingly, the gene probes targeting FGF17, PRKCG, SSTR1, and SCTR predicted potential targeted agents for renal cancer metastasis and adjuvant immunotherapy, including 5224221, calmidazolium, and sulfasalazine. Current studies had found that calmidazolium, as calmodulin inhibitors, could not only affect the survival status of various immune cells (Hu et al., 2019) but also affect the inositol-1,4,5-triphosphate receptor/calcium/calmodulin pathway by mediating RACK1 and regulate the proliferation of preglomerular microvascular smooth muscle cells and mesangial cells, thus treating kidney diseases (Cheng et al., 2011). In addition, calmidazolium can induce apoptosis and down-regulate stem cell–related genes to inhibit the growth of embryonal carcinoma cells (Lee et al., 2016). Sulfasalazine, as sulfonamide antibiotic, had antibacterial, anti-inflammatory, and immuno-suppressive effects. Studies had shown that sulfasalazine could be involved in cancer cell death and T cell immunity by inhibiting the ferroptosis-related NF-κB signaling pathway and systemic Xc transporters (Dixon et al., 2012; Dixon et al., 2014). At present, sulfasalazine had been found to have significant effects on tumor cells in breast cancer (Yu et al., 2019), thyroid cancer (Zou L et al., 2019), kidney cancer (Sourbier et al., 2007), and bladder cancer (Ogihara et al., 2019). Although calmidazolium and sulfasalazine had been proven to affect the occurrence, metastasis, and apoptosis of various tumors, their specific mechanisms were still unclear, and there was no relevant study on the efficacy in KIRC, which was worth further exploration.

There are also some limitations in this study. First, the retrospective study determined that there is heterogeneity in the results, so further *in vivo* and *in vitro* experiments are needed to validate the findings of this study. Second, it is necessary that we need more basic and large clinical trials to validate these findings.

CONCLUSION

In this study, we obtained hub mIR-DEGs with prognostic value through comprehensive bioinformatics analysis, including FGF17, PRKCG, SSTR1, and SCTR, which were significantly associated with methylation, ferroptosis, and immune checkpoint–related genes in KIRC. Preliminary validation found that PRKCG and SSTR1 were consistent with predictions. These indicators could be new targets and prognostic biomarkers for KIRC's metastasis and immunotherapy. Furthermore, we had predicted the formula of targeted small molecule drugs based on hub mIR-DEGs. However, this prediction still needed lots of basic experimental demonstration.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee and Institutional Review Board for Clinical Research of Zhongda Hospital. 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

MC, GZ, and LZ were responsible for study design, data acquisition and analysis, and manuscript writing. GZ, SS, and KP performed bioinformatics and statistical analyses. GZ, SS, and WM prepared the figures and tables for the manuscript. SS, LW, and LD were responsible for the integrity of the entire study and manuscript review. All authors read and approved the final manuscript.

FUNDING

This study was funded by the Natural Science Foundation of China (82170703 and 82100732), Natural Science Foundation of Jiangsu Province (BK20200360), and Excellent Youth Development Fund of Zhongda Hospital, SEU (2021ZDYYYQPY04).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.794326/full#supplementary-material

REFERENCES

Amin, A., Dudek, A. Z., Logan, T. F., Lance, R. S., Holzbeierlein, J. M., Knox, J. J., et al. (2015). Survival with AGS-003, an Autologous Dendritic Cell-Based Immunotherapy, in Combination with Sunitinib in Unfavorable Risk Patients with Advanced Renal Cell Carcinoma (RCC): Phase 2 Study Results. *J. Immunotherapy Cancer* 3, 14. doi:10.1186/s40425-015-0055-3

Athar, U., and Gentile, T. C. (2008). Treatment Options for Metastatic Renal Cell Carcinoma: a Review. *Can. J. Urol.* 15, 3954–3966.

Awasthi, N., Yen, P. L., Schwarz, M. A., and Schwarz, R. E. (2012). The Efficacy of a Novel, Dual PI3K/mTOR Inhibitor NVP-Bez235 to Enhance Chemotherapy and Antiangiogenic Response in Pancreatic Cancer. *J. Cel. Biochem.* 113, 784–791. doi:10.1002/jcb.23405

Braun, D. A., Bakouny, Z., Hirsch, L., Flippot, R., Van Allen, E. M., Wu, C. J., et al. (2021). Beyond Conventional Immune-Checkpoint Inhibition - Novel Immunotherapies for Renal Cell Carcinoma. *Nat. Rev. Clin. Oncol.* 18, 199–214. doi:10.1038/s41571-020-00455-z

Chen, W., Schweins, E., Chen, X., Finn, O. J., and Cheever, M. A. (1994). Retroviral Transduction of Protein Kinase C-Gamma into Tumor-specific T Cells Allows Antigen-independent Long-Term Growth in IL-2 with Retention of Functional Specificity *In Vitro* and Ability to Mediate Tumor Therapy *In Vivo*. *J. Immunol.* 153, 3630–3638.

Cheng, D., Zhu, X., Barchiesi, F., Gillespie, D. G., Dubey, R. K., and Jackson, E. K. (2011). Receptor for Activated Protein Kinase C1 Regulates Cell Proliferation by Modulating Calcium Signaling. *Hypertension* 58, 689–695. doi:10.1161/HYPERTENSIONAHA.111.174508

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

Dixon, S. J., Patel, D. N., Welsch, M., Skouta, R., Lee, E. D., Hayano, M., et al. (2014). Pharmacological Inhibition of Cystine-Glutamate Exchange Induces Endoplasmic Reticulum Stress and Ferroptosis. *Elife* 3, e02523. doi:10.7554/eLife.02523

Dunnick, N. R. (2016). Renal Cell Carcinoma: Staging and Surveillance. *Abdom. Radiol.* 41, 1079–1085. doi:10.1007/s00261-016-0692-0

Escudier, B., Porta, C., Schmidinger, M., Rioux-Leclercq, N., Bex, A., Khoo, V., et al. (2019). Renal Cell Carcinoma: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* 30, 706–720. doi:10.1093/annonc/mdz056

Escudier, B. (2012). Emerging Immunotherapies for Renal Cell Carcinoma. *Ann. Oncol.* 23 (Suppl. 8), viii35–viii40. doi:10.1093/annonc/mds261

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., et al. (2019). Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-Seq Data. *Genome Med.* 11, 34. doi:10.1186/s13073-019-0638-6

Floros, T., and Tarhini, A. A. (2015). Anticancer Cytokines: Biology and Clinical Effects of Interferon-A2, Interleukin (IL)-2, IL-15, IL-21, and IL-12. *Semin. Oncol.* 42, 539–548. doi:10.1053/j.seminoncol.2015.05.015

Frega, G., Wu, Q., Le Naour, J., Vacchelli, E., Galluzzi, L., Kroemer, G., et al. (2020). Trial Watch: Experimental TLR7/TLR8 Agonists for Oncological Indications. *Oncoimmunology* 9, 1796002. doi:10.1080/2162402X.2020.1796002

Gao, Y., Li, H., Ma, X., Fan, Y., Ni, D., Zhang, Y., et al. (2017). KLF6 Suppresses Metastasis of Clear Cell Renal Cell Carcinoma via Transcriptional Repression of E2F1. *Cancer Res.* 77, 330–342. doi:10.1158/0008-5472.CAN-16-0348

Ghatalia, P., Zibelman, M., Geynisman, D. M., and Plimack, E. R. (2017). Checkpoint Inhibitors for the Treatment of Renal Cell Carcinoma. *Curr. Treat. Options. Oncol.* 18, 7. doi:10.1007/s11864-017-0458-0

Hatzaras, I., Gleisner, A. L., Pulitano, C., Sandroussi, C., Hirose, K., Hyder, O., et al. (2012). A Multi-Institution Analysis of Outcomes of Liver-Directed Surgery for Metastatic Renal Cell Cancer. *HPB* 14, 532–538. doi:10.1111/j.1477-2574.2012.00495.x

Hu, J., Shi, D., Ding, M., Huang, T., Gu, R., Xiao, J., et al. (2019). Calmodulin-dependent Signalling Pathways Are Activated and Mediate the Acute Inflammatory Response of Injured Skeletal Muscle. *J. Physiol.* 597, 5161–5177. doi:10.1113/JP278478

Jung, K., Lein, M., Laube, C., and Lichtinghagen, R. (2001). Blood Specimen Collection Methods Influence the Concentration and the Diagnostic Validity of Matrix Metalloproteinase 9 in Blood. *Clin. Chim. Acta* 314, 241–244. doi:10.1016/s0009-8981(01)00679-9

Katoh, M. (2016). FGFR Inhibitors: Effects on Cancer Cells, Tumor Microenvironment and Whole-Body Homeostasis (Review). *Int. J. Mol. Med.* 38, 3–15. doi:10.3892/ijmm.2016.2620

Kawabata, A., Matsuzuka, T., Doi, C., Seiler, G., Reischman, J., Pickel, L., et al. (2012). C1B Domain Peptide of Protein Kinase Cγ Significantly Suppresses Growth of Human colon Cancer Cells *In Vitro* and in an *In Vivo* Mouse Xenograft Model through Induction of Cell Cycle Arrest and Apoptosis. *Cancer Biol. Ther.* 13, 880–889. doi:10.4161/cbt.20840

Lee, M., Waser, B., Reubi, J.-C., and Pellegata, N. S. (2012). Secretin Receptor Promotes the Proliferation of Endocrine Tumor Cells via the PI3K/AKT Pathway. *Mol. Endocrinol.* 26, 1394–1405. doi:10.1210/me.2012-1055

Lee, J., Kim, M. S., Kim, M. A., and Jang, Y. K. (2016). Calmidazolium Chloride Inhibits Growth of Murine Embryonal Carcinoma Cells, a Model of Cancer Stem-like Cells. *Toxicol. Vitro* 35, 86–92. doi:10.1016/j.tiv.2016.05.015

Li, M., Yan, S., Fisher, W. E., Chen, C., and Yao, Q. (2005). New Roles of a Neuropeptide Cortistatin in the Immune System and Cancer. *World J. Surg.* 29, 354–356. doi:10.1007/s00268-004-7811-8

Ljungberg, B., Campbell, S. C., Cho, H. Y., Jacqmin, D., Lee, J. E., Weikert, S., et al. (2011). The Epidemiology of Renal Cell Carcinoma. *Eur. Urol.* 60, 615–621. doi:10.1016/j.eururo.2011.06.049

Mao, W., Sun, S., He, T., Jin, X., Wu, J., Xu, B., et al. (2021). Systemic Inflammation Response Index Is an Independent Prognostic Indicator for Patients with Renal Cell Carcinoma Undergoing Laparoscopic Nephrectomy: A Multi-Institutional Cohort Study. *Cancer Manag. Res.* 13, 6437–6450. doi:10.2147/CMAR.S328213

Mao, W., Wang, K., Xu, B., Zhang, H., Sun, S., Hu, Q., et al. (2021). ciRS-7 Is a Prognostic Biomarker and Potential Gene Therapy Target for Renal Cell Carcinoma. *Mol. Cancer* 20, 142. doi:10.1186/s12943-021-01443-2

Nagarajan, S. K., Babu, S., Sohn, H., and Madhavan, T. (2020). Molecular-Level Understanding of the Somatostatin Receptor 1 (SSTR1)-Ligand Binding: A Structural Biology Study Based on Computational Methods. *ACS Omega* 5, 21145–21161. doi:10.1021/acsomega.0c02847

Nishizuka, Y. (1984). The Role of Protein Kinase C in Cell Surface Signal Transduction and Tumour Promotion. *Nature* 308, 693–698. doi:10.1038/308693a0

Ogihara, K., Kikuchi, E., Okazaki, S., Hagiwara, M., Takeda, T., Matsumoto, K., et al. (2019). Sulfasalazine Could Modulate the CD 44v9- xCT System and Enhance Cisplatin-induced Cytotoxic Effects in Metastatic Bladder Cancer. *Cancer Sci.* 110, 1431–1441. doi:10.1111/cas.13960

Plattner, C., Finotello, F., and Rieder, D. (2020). Deconvoluting Tumor-Infiltrating Immune Cells from RNA-Seq Data Using quanTIseq. *Methods Enzymol.* 636, 261–285. doi:10.1016/bs.mie.2019.05.056

Reubi, J. C., Schaer, J.-C., Waser, B., Hoeger, C., and Rivier, J. (1998). A Selective Analog for the Somatostatin Sst1-Receptor Subtype Expressed by Human Tumors. *Eur. J. Pharmacol.* 345, 103–110. doi:10.1016/s0014-2999(97)01618-x

Ricketts, C. J., De Cubas, A. A., Fan, H., Smith, C. C., Lang, M., Reznik, E., et al. (2018). The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Rep.* 23, 3698–4326. doi:10.1016/j.celrep.2018.06.032

Şenbabaoğlu, Y., Gejman, R. S., Winer, A. G., Liu, M., Van Allen, E. M., de Velasco, G., et al. (2016). Tumor Immune Microenvironment Characterization in clear Cell Renal Cell Carcinoma Identifies Prognostic and Immunotherapeutically Relevant Messenger RNA Signatures. *Genome Biol.* 17, 231. doi:10.1186/s13059-016-1092-z

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer Statistics, 2018. *CA: A Cancer J. Clinicians* 68, 7–30. doi:10.3322/caac.21442

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer Statistics, 2021. *CA A. Cancer J. Clin.* 71, 7–33. doi:10.3322/caac.21654

Sourbier, C., Danilin, S., Lindner, V., Steger, J., Rothhut, S., Meyer, N., et al. (2007). Targeting the Nuclear Factor- B Rescue Pathway Has Promising Future in Human Renal Cell Carcinoma Therapy. *Cancer Res.* 67, 11668–11676. doi:10.1158/0008-5472.CAN-07-0632

Staehler, M. (2011). The Role of Metastasectomy in Metastatic Renal Cell Carcinoma. *Nat. Rev. Urol.* 8, 180–181. doi:10.1038/nrurol.2011.30

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

Tabarés-Seisdedos, R., and Rubenstein, J. L. R. (2009). Chromosome 8p as a Potential Hub for Developmental Neuropsychiatric Disorders: Implications for Schizophrenia, Autism and Cancer. *Mol. Psychiatry* 14, 563–589. doi:10.1038/mp.2009.2

Tang, X., Liu, T., Zang, X., Liu, H., Wang, D., Chen, H., et al. (2013). Adoptive Cellular Immunotherapy in Metastatic Renal Cell Carcinoma: a Systematic Review and Meta-Analysis. *PLoS One* 8, e62847. doi:10.1371/journal.pone.0062847

Taube, J. M., Galon, J., Sholl, L. M., Rodig, S. J., Cottrell, T. R., Giraldo, N. A., et al. (2018). Implications of the Tumor Immune Microenvironment for Staging and Therapeutics. *Mod. Pathol.* 31, 214–234. doi:10.1038/modpathol.2017.156

Vuong, L., Kotecha, R. R., Voss, M. H., and Hakimi, A. A. (2019). Tumor Microenvironment Dynamics in Clear-Cell Renal Cell Carcinoma. *Cancer Discov.* 9, 1349–1357. doi:10.1158/2159-8290.CD-19-0499

Xie, F., Xu, M., Lu, J., Mao, L., and Wang, S. (2019). The Role of Exosomal PD-L1 in Tumor Progression and Immunotherapy. *Mol. Cancer* 18, 146. doi:10.1186/s12943-019-1074-3

Yu, H., Yang, C., Jian, L., Guo, S., Chen, R., Li, K., et al. (2019). Sulfasalazine Induced Ferroptosis in Breast Cancer Cells is Reduced by the Inhibitory Effect of Estrogen Receptor on the Transferrin Receptor. *Oncol. Rep.* 42, 826–838. doi:10.3892/or.2019.7189

Zou L, L., Gao, Z., Zeng, F., Xiao, J., Chen, J., Feng, X., et al. (2019). Sulfasalazine Suppresses Thyroid Cancer Cell Proliferation and Metastasis through T Cell Originated Protein Kinase. *Oncol. Lett.* 18, 3517–3526. doi:10.3892/ol.2019.10721

Zou Y, Y., Tan, H., Zhao, Y., Zhou, Y., and Cao, L. (2019). Expression and Selective Activation of Somatostatin Receptor Subtypes Induces Cell Cycle Arrest in Cancer Cells. *Oncol. Lett.* 17, 1723–1731. doi:10.3892/ol.2018.9773

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 Molecular Biosciences

ORIGINAL RESEARCH
published: 11 February 2022
doi: 10.3389/fmolb.2022.843150

Inferring Retinal Degeneration-Related Genes Based on Xgboost

*Yujie Xia[1], Xiaojie Li[1], Xinlin Chen[2], Changjin Lu[1] and Xiaoyi Yu[1]**

[1]*Department of Ophthalmology, The First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China,* [2]*Guangzhou University of Chinese Medicine, Guangzhou, China*

Retinal Degeneration (RD) is an inherited retinal disease characterized by degeneration of rods and cones photoreceptor cells and degeneration of retinal pigment epithelial cells. The age of onset and disease progression of RD are related to genes and environment. At present, research has discovered five genes closely related to RD. They are RHO, PDE6B, MERTK, RLBP1, RPGR, and researchers have developed corresponding gene therapy methods. Gene therapy uses vectors to transfer therapeutic genes, genetically modify target cells, and correct or replace disease-causing RD genes. Therefore, identifying the pathogenic genes of RD will play an important role in the development of treatment methods for the disease. However, the traditional methods of identifying RD-related genes are mostly based on animal experiments, and currently only a small number of RD-related genes have been identified. With the increase of biological data, Xgboost is purposed in this article to identify RP-related genes. Xgboost adds a regular term to control the complexity of the model, hence using Xgboost to find out true RD-related genes from complex and massive genes is suitable. The problem of overfitting can be avoided to some extent. To verify the power of Xgboost to identify RD-related genes, we did 10-cross validation and compared with three traditional methods: Random Forest, Back Propagation network, Support Vector Machine. The accuracy of Xgboost is 99.13% and AUC is much higher than other three methods. Therefore, this article can provide technical support for efficient identification of RD-related genes and help researchers have a deeper the understanding of the genetic characteristics of RD.

Keywords: retinitis degeneration, Xgboost, amino acids, pathogenic gene, machine learning

OPEN ACCESS

Edited by:
Tianyi Zhao,
Harbin Institute of Technology, China

Reviewed by:
Yuansong Zhao,
University of Texas Health Science Center at Houston, United States
Ningyi Zhang,
Harbin Institute of Technology, China

Correspondence:
Xiaoyi Yu
yxy741107@126.com

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *24 December 2021*
Accepted: *17 January 2022*
Published: *11 February 2022*

Citation:
Xia Y, Li X, Chen X, Lu C and Yu X (2022) Inferring Retinal Degeneration-Related Genes Based on Xgboost. Front. Mol. Biosci. 9:843150. doi: 10.3389/fmolb.2022.843150

INTRODUCTION

Hereditary eye diseases include syndromes and non-syndromic forms of retinal degeneration, hereditary glaucoma, corneal dystrophy and eye movement disorders. Retinal degeneration (RD) is a group of single-gene hereditary blindness caused by loss of function of photoreceptor cells or retinal pigment epithelium (RPE). The incidence of RDs worldwide is 1/3,000–1/2,000 (Berger et al., 2010). According to whether they are accompanied by systemic symptoms, they are divided into simple and systemic RDs (Wennström et al., 2003).The former mainly includes retinitis pigmentosa (RP), Rod cell dystrophy (cone-rod dystrophies, CORD), Leber congenital amaurosis (Leber congenital amaurosis, LCA), etc. The latter mainly includes Usher syndrome and Bardet-Biedl syndrome (Muller et al., 2010).Up to now, more than 300 pathogenic genes have been reported for RD, which suggests that RD has a high degree of clinical and genetic heterogeneity, the diagnosis of this type of disease is extremely difficult (Benayoun et al., 2009). Research on the pathogenic genes of RDs and the development and application of related molecular diagnostic techniques are the prerequisites for the diagnosis, prevention and treatment of RDs. Both single-gene Mendelian or complex hereditary eye diseases require genetic testing to determine the underlying cause. There are nearly 1,200 genes related to eye diseases in the human online Mendelian genetic database (on-line Mendelian inheritancein man, oMIM) (http://www.omim.org)(Amberger et al., 2015). RD is a type of disease with obvious clinical phenotypic heterogeneity and genetic heterogeneity, and it is also the main type of ophthalmic genetic diseases and rare and difficult ophthalmic diseases. At present, the vast majority of RD is still incurable in ophthalmology, and research on its diagnosis and treatment has always been a hot spot. Diagnosing RD at the genetic level is helpful for a deep understanding of the disease mechanism (Boycott et al., 2017). Distinguishing what kind of gene mutation causes the disease can more accurately understand the occurrence, development and outcome of the disease. This is especially important for RD with obvious heterogeneity. The genetic heterogeneity of RD requires a new disease naming and definition system. The system should include at least two main factors, namely the disease-causing gene and the name of the disease related to it. For example, EYS-related retinitis pigmentosa is more accurate than retinitis pigmentosa alone, and it is easier to explain the condition to the patient.

Because of the large number of pathogenic genes of retinal degeneration and the different mutation genes and loci in different families, it is very difficult to selectively screen candidate pathogenic genes. At present, the research on molecular genetics of hereditary eye disease is mainly family single gene research, which leads to controversy and deficiency in the genetic research of RD gene (Fan et al., 2006). A comprehensive and systematic analysis of known gene variation data may be helpful for the further study of such problems. Genes and mutations associated with retinal degeneration are controversial. Some genes were first reported to be disease-related, and then no mutations were reported. Although a large number of mutations in retinal degeneration are concentrated in a few genes, and the mutations of many genes only explain the causes of a very small number of patients, it is possible that only a very small number of patients with this gene carry mutations, but it cannot be ruled out that the previous research only found changes in a single gene and mistakenly believed that it was the cause of the disease. The controversial and questionable problems such as mutation penetrance and related risk factors reported in single gene research also bring confusion to researchers. In addition, because there was no public database containing a large number of variation data and a large number of control validation, some high-frequency SNPs were found in patients and were regarded as pathogenic mutations. These mutations are listed in the human gene mutation database (HGMD) as pathogenic mutations (Stenson et al., 2020), which mislead the follow-up molecular genetics research. At present, the reported variation analysis doubts and corrects the pathogenicity of individual Retnet genes and mutations (Pozo et al., 2015), such as the previously reported pathogenic genes fscn2 (MIM: 607643) and or2w3 of retinitis pigmentosa and hmcn1 (MIM: 608548) of macular degeneration (Fisher et al., 2007; Zhang et al., 2007; Sharon et al., 2016), and the subsequent research reports are questionable, but due to the lack of clinical phenotype analysis of patients with the same mutation, It is still impossible to completely deny its possibility as a pathogenic gene. In addition, single-gene research cannot comprehensively and systematically understand the genetic mutation spectrum of the people with hereditary retinal degeneration of this ethnic group. Different races have different gene mutation spectrums. Common disease-causing gene mutations in European and American populations are not common in Asian populations; based on common gene mutations in Asian populations, they may be very rare in European and American populations. For example, the pathogenic gene CNGA3 (MIM: 600053) of pyramidal cell dystrophy is the most frequently mutated gene in Chinese patients (Huang et al., 2016), and the most common recessive genetic mutation in foreign reports is ABCA4 (MIM: 601691) (Maugeri et al., 2000), CNGA3 only explains a small part of the cause of the disease (Wissinger et al., 2001). Even the Asian population has a different mutation spectrum. The highest mutation frequency in the Japanese retinitis pigmentosa population is EYS (MIM: 612424)(Oishi et al., 2014; Arai et al., 2015), and this gene mutation is very rare in Chinese patients (Xu et al., 2014; Chen et al., 2015). It is very important and necessary to conduct a comprehensive multi-gene systematic analysis of all retinal degeneration genes, and to understand the clinical characteristics, gene mutation frequency spectrum and discover the main pathogenic genes of the people with retinal degeneration of this nation. At the same time, it also provides important clinical evidence for the clinical diagnosis, genetic counseling, and prevention of hereditary eye diseases.

Although researchers have made great achievement in identifying RD-related genes, identifying the huge and complex acid sequences needs an algorithm which has high computational efficiency and high recognition accuracy. The generation of multi-omics data allows us to combine different data from a large number of samples to explore RD-related genes at a comprehensive level (Zhao et al., 2021a). Integrating multiple omics data to discover biological knowledge on a large scale has become a universal method. An endless stream of methods have been developed to apply to different research problems, such as identification of disease-related gene (Zhao et al., 2020; Antonarakis, 2021), identification of disease-related protein (Katako et al., 2018; Zhao et al., 2021b), identification of disease-related metabolite (Lei and Tie, 2019; Zhao et al., 2021c), disease-related drug target identification (Agamah et al., 2020; Zhao et al., 2021d), etc. Chen (Chen and Guestrin, 2016) purposed a novel method named Extreme Gradient Boosting (Xgboost) in 2004. He improved the boosting algorithm. Its multi-threaded parallel and regularization term not only improve the accuracy of the algorithm but also reduce the running time. Therefore, Xgboost is a suitable algorithm to solve the problem of identifying RD-related genes.

TABLE 1 | The six groups of the 20 amino acids.

Groups	Amino acids
Strongly hydrophilic	R,D,E,N,Q,K,H
Strongly hydrophobic	L,I,V,A,M,F
Weakly hydrophilic or Weakly hydrophobic	S,T,Y,W
Proline	P
Glycine	G
Cysteine	C

METHODS AND MATERIALS

Data Description

We searched RD-related genes from DisGeNET (Piñero et al., 2020) by the key word "Retinal Degeneration." There are 207 genes which are known to be related to RD in this database. We downloaded the sequences of these genes corresponding proteins from Uniprot (Consortium, 2019).

We also obtained 5,000 genes as genes potentially associated with RD from Genecard (Safran et al., 2010). Our aim is to identify RD-related genes from these 5,000 genes.

Feature Extraction

Compositional Analysis

Since the real constitution of RD-related genes encoded proteins is quite different from the non-related genes', the frequency of the occurrence of the all 20 amino acids in these proteins could be quite different.

We totally calculated the average amino acid composition of 207 RD-related genes encoded proteins. These proteins are richest in "L," and the composition of "G," "A," "V," "E," "S" is very high.

Dissociation Constant

The protein structure is significantly related to the chemical characteristic of amino acid, especially hydrophobic and hydrophilic (Aftabuddin and Kundu, 2007). Aftabuddin et al. divided 20 amino acids into six groups based on the ranges of the hydropathy. The reason why the gene is related to RD is significantly related to the function of the protein it encodes. Therefore, the hydrophilicity and hydrophobicity of amino acids in protein are the key to judging whether the gene is related to RD. **Table 1** shows the six groups of the 20 amino acids.

So, the sequence of every protein could be diverted to a 6-dimension sequence. Each dimension is the average composition of one of these six groups.

PEST Regions

In 1986, Rechsteiner M and Rogers SW (Rechsteiner et al., 1996) made the assumption that the amino acids of "P," "E," "S" and "T" can serve as proteolytic signals. Now more and more reports have verified that the sequence which contains PEST regions can cause the rapid degradation of proteins.

The Epestfind program can be used to identify all poor and potential PEST protein sequences. (Espreafico et al., 1992) http://emboss.bioinformatics.nl/cgi-bin/emboss/epestfind.

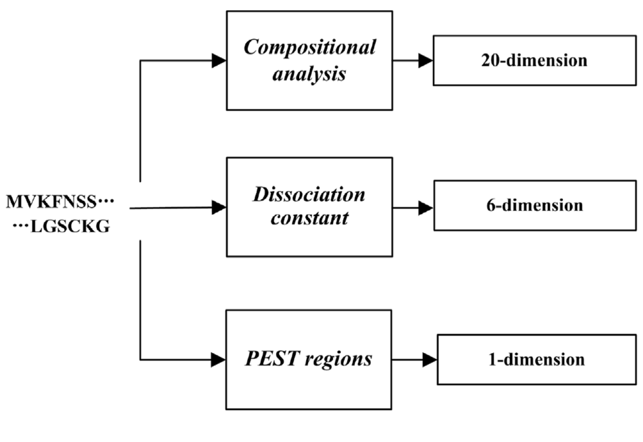


FIGURE 1 | Flow chart of Feature extraction.

We only included potential PEST protein region as a feature to identify the RD-related genes. We counted the number of potential pest regions in each sequence.

In conclusion, we totally extracted three kinds of features (**Figure 1**).

So, we used these 27-dimensions to identify the RD-related.

Methods and Framework

Extreme Gradient Boosting

The Extreme Gradient Boosting (Xgboost) is the improvement of traditional Gradient Boosting Decision Tree (GBDT). Xgboost implements the first and the two order derivatives from the loss function by applying two order Taylor expansion. However, the traditional GBDT algorithm only implements first derivative information during optimizing. Xgboost runs significantly faster than GBDT. Because it has two advantages. On the one hand, Xgboost supports automatic multi-core parallel computing through open MP. On the other hand, Xgboost proposes a new data format Dmatrix, which can be preprocessed first and then trained. This improves the efficiency of each iteration of the training process and reduces the model training time. In addition, we can input the sparse matrix into xgboost.

First, we need to obtain our train set $\{x_i, y_i\}^N$, $y_i \in \{-1, 1\}$ and set the number of leaf nodes as J. Then, we need to initialize the final function.

$$F_0(x) = \frac{1}{2}\log\frac{1+\bar{y}}{1-\bar{y}} \tag{1}$$

Then, the gradient of training samples can be obtained by:

$$\widehat{y}_i = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \tag{2}$$

Then, the CART regression tree $\{R_{jm}\}^J$ can be constructed. R_{jm} is the $\mathrm{j_{th}}$ feature space.

Then, each leaf node's regression value can be obtained by:

$$r_{jm} = \frac{\sum_{x_i \in R_{jm}} \widehat{y}_i}{\sum_{x_i \in R_{jm}} \left|\widehat{y}_i\right| \left(2 - \left|\widehat{y}_i\right|\right)} \tag{3}$$

Finally, the final model is as following:

TABLE 2 | The parameters of the Xgboost.

Setting items	The value set
Booster	gbtree
Silent	0
Learning rate	0.3
Maximum depth of a tree	6
Minimum sum of instance weight	1
Subsample ratio	1
Experimental parameter	1

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} r_{jm} I\left(x \in R_{jm}\right) \tag{4}$$

The objective function is consisted by loss function and regularization term, which can be used to show the quality of our method.

$$Obj(\Theta) = L(\theta) + \Omega(\Theta) \tag{5}$$

$L(\theta)$ represents loss function. Algorithms such as artificial neural networks only use loss function to evaluate the quality of training, which is easy to cause over fitting. The regularization parameters $\Omega(\Theta)$ are introduced into methods such as support vector machine, which can effectively reduce over fitting. However, the introduction of regularization parameters will increase the complexity of the model.

CART is the basic unit of Xgboost. Therefore, the objective function in **formula (5)** can also be represented as following:

$$Obj(\Theta) = \sum_{i}^{n} l\left(y_i, \widehat{y}_i\right) + \sum_{t=1}^{T} \Omega(f_t) \tag{6}$$

Each tree is obtained based on the last tree we constructed.

$$\begin{aligned}
\widehat{y}_i^{0} &= 0,\\
\widehat{y}_i^{1} &= f_1(x_i) = \widehat{y}_i^{0} + f_1(x_i),\\
\widehat{y}_i^{2} &= f_1(x_i) + f_2(x_i) = \widehat{y}_i^{1} + f_2(x_i),\\
&\vdots\\
\widehat{y}_i^{2} &= \sum_{k=1}^{t} f_k(x_i) = \widehat{y}_i^{t-1} + f_t(x_i),
\end{aligned} \tag{7}$$

Finally, we can obtained the first and the two order derivatives from the loss function.

$$\begin{aligned}
Obj^{(t)} = \sum_{i}^{n}\left(l\left(y_i, \widehat{y}_i^{t-1}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right) + \Omega(f_t)\\
+ \text{constant}
\end{aligned} \tag{8}$$

The next part is to obtain regularization term. Firstly, we define the decision tree as:

$$f_t(x) = w_{q(x)}, w \in R^M, q\colon R^d \rightarrow \{1, 2, \cdots, M\} \tag{9}$$

w represents leaf node's score. q(x) is used to determine the position of the input sample in the tree. The regularization term can be represented as following:

TABLE 3 | The results of the ten experiments.

		Prediction		
		1	0	Total
True Label	1	205 (TP)	2(FN)	207
	0	20(FP)	4,980 (TN)	5,000
Total		225	4,982	5,207

$$\Omega(f) = \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} w_j^2 \tag{10}$$

We need to set γ and λ to balance the complexity of the model. So $\mathrm{t_{th}}$ tree's objective function is as following:

$$\begin{aligned}
Obj^{(t)} &\approx \sum_{i=1}^{n}\left(g_i w_q(x_i) + \frac{1}{2} h_i w_q^2(x_i)\right) + \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} w_j^2\\
&= \sum_{j=1}^{M}\left(\left(\sum g_i\right) w_j + \frac{1}{2}\left(\sum h_i + \lambda\right) w_j^2\right) + \gamma M
\end{aligned} \tag{11}$$

We could define $G_j = \sum g_i$ and $H_j = \sum h_i$, then we get:

$$Obj^{(t)} = \sum_{j=1}^{M}\left(G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2\right) + \gamma M \tag{12}$$

RESULTS

Experiment Description

We totally got 207 true RD-related genes and we randomly selected 5,000 genes as the negative samples. To verify the effectiveness of Xgboost on identifying RD-related genes, we did ten-cross validation.

We randomly divided these 5,207 sequences into ten groups. For every group, we choose 520 sequences as the test set and the rest 4,687 sequences as the train set. So, we did ten experiments in total. Besides, every sequence has become a training set and a test set. We set the parameters of Xgboost as the **Table 2**.

Evaluation Criteria

We use four evaluation ways to evaluate the performance of Xgboost on identifying RD-related genes.

We put the results of the ten experiments in the **Table 2**. A total of 5,207 sequences were tested. As showed in **Table 3**, we could calculate the Accuracy = 99.13%, Precision = 99.04%, Recall = 99.23%, Specificity = 99.04%.

Experiments Result

In this study, the label of randomly selected genes is 0, and the label of RD-related genes are 1.

The **Figure 2** shows the curves of the ten times experiments' accuracy. As we can see, the experiment with the lowest accuracy is also more than 98%.

To verify the superiority of the Xgboost, we also use the same data to do the ten-cross validation by other methods. We use Back

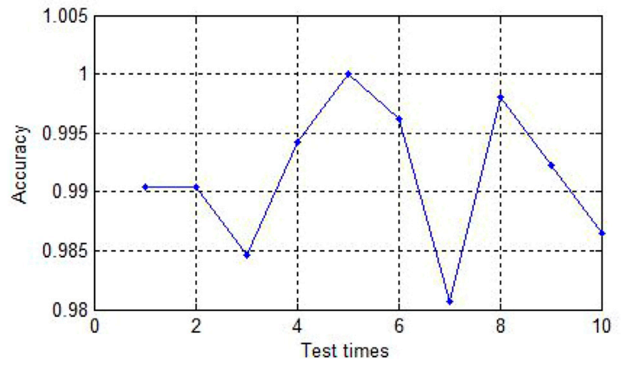


FIGURE 2 | The accuracy of ten experiments.

TABLE 4 | Comparison of the Xgboost with alternative models.

Algorithm	ACC (%)	Precision (%)	Recall (%)	Specificity (%)
Xgboost	99.13	99.04	99.23	99.04
BP	82.50	78.13	90.25	74.76
Random Forest	97.99	99.64	96.34	99.65
SVM	94.16	94.62	93.64	94.68

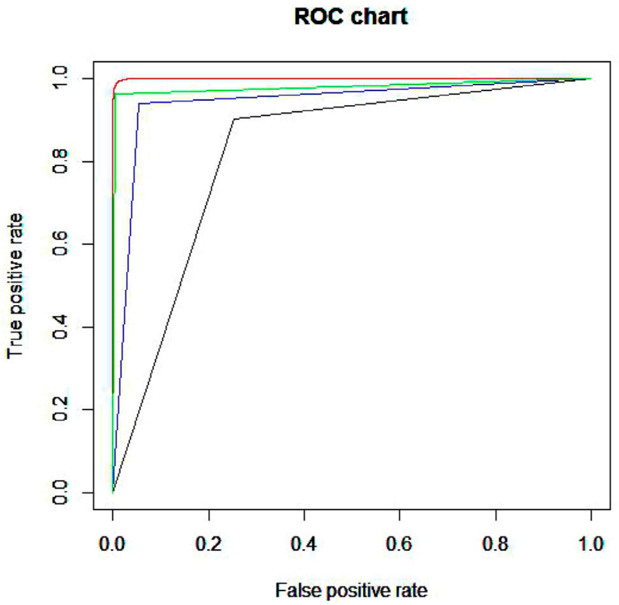


FIGURE 3 | ROC curve of four methods.

Propagation network (BP), Random Forest (RF), Support Vector Machine (SVM) respectively. The error statistics of the average results of 10 experiments are shown in the following table.

As we can see in the **Table 4**, we could see the performance of Xgboost is the best, and the performance of BP is the worst. Although RF is better than the Xgboost in the evaluations of 'Precision' and "Specificity," the accuracy of the Xgboost is the best. Besides, Xgboost uses the least time to build up the model.

Figure 3 is the ROC curve of four methods. The red line is the curve of Xgboost. The green line is the curve of RF. The blue and black one is the SVM and BP respectively. As we can see in the figure, Xgboost is the best among these four methods. Then we draw a figure of AUC in the **Figure 4**.

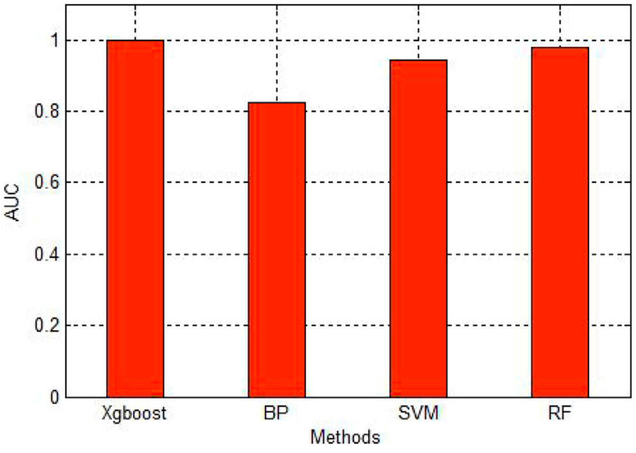


FIGURE 4 | AUC of four methods.

As we can see in the **Figure 4**, the AUC of Xgboost is very close to 1. It shows the high accuracy of the Xgboost.

CONCLUSION

Typical clinical features of RD include early night blindness, subsequent progressive vision loss and narrowing of the visual field, fundus showing osteocytic pigmentation, waxy pale atrophy of the optic disc, and electroretinogram (ERG) cone and rod Cell function decline, etc., the early rod cell response amplitude decline is more serious than the cone cell response amplitude. Due to the high degree of heterogeneity of the RP phenotype, many retinopathy have similar symptoms with RP, which is very easy to confuse.

Therefore, exploring RD from a genetic perspective is very helpful for clinical diagnosis, treatment and research on the pathogenic mechanism of diseases. With the popularization of high-throughput sequencing technology, a large amount of genome and proteomic data has been released. However, no method has been proposed to specifically identify RD-related genes. In this article, we propose a method based on XGboost to identify RD-related genes. We extracted three features of the corresponding proteins of 207 genes known to be related to RD. Each gene has 27-dimensional features, and we input these features into Xgboost for training. Through 10-fold cross-validation, we confirmed the accuracy of our method to identify RD-related genes with AUC as 0.99.

In summary, we propose a method for large-scale identification of RD-related genes. This type of machine learning method can prioritize genes that are potentially related to RD to save researchers the cost of conducting biological experiments.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

YX and XY conceived and designed the study. YX, XL, XC and CL performed the analysis procedures and analyzed the data. YX and XL wrote the article. All authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No.82174441).

REFERENCES

Aftabuddin, M., and Kundu, S. (2007). Hydrophobic, Hydrophilic, and Charged Amino Acid Networks within Protein. *Biophysical J.* 93 (1), 225–231. doi:10.1529/biophysj.106.098004

Agamah, F. E., Mazandu, G. K., Hassan, R., Bope, C. D., Thomford, N. E., Ghansah, A., et al. (2020). Computational/In Silico Methods in Drug Target and lead Prediction. *Brief. Bioinformatics* 21 (5), 1663–1675. doi:10.1093/bib/bbz103

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM), an Online Catalog of Human Genes and Genetic Disorders. *Nucleic Acids Res.* 43 (D1), D789–D798. doi:10.1093/nar/gku1205

Antonarakis, S. E. (2021). *History of the Methodology of Disease Gene Identification*. Hoboken, New Jersey, United States: Wiley Online Library.

Arai, Y., Maeda, A., Hirami, Y., Ishigami, C., Kosugi, S., Mandai, M., et al. (2015). Retinitis Pigmentosa with EYS Mutations Is the Most Prevalent Inherited Retinal Dystrophy in Japanese Populations. *J. Ophthalmol.* 2015, 819760. doi:10.1155/2015/819760

Benayoun, L., Spiegel, R., Auslender, N., Abbasi, A. H., Rizel, L., Hujeirat, Y., et al. (2009). Genetic Heterogeneity in Two Consanguineous Families Segregating Early Onset Retinal Degeneration: the Pitfalls of Homozygosity Mapping. *Am. J. Med. Genet.* 149A (4), 650–656. doi:10.1002/ajmg.a.32634

Berger, W., Kloeckener-Gruissem, B., and Neidhardt, J. (2010). The Molecular Basis of Human Retinal and Vitreoretinal Diseases. *Prog. Retin. Eye Res.* 29 (5), 335–375. doi:10.1016/j.preteyeres.2010.03.004

Boycott, K. M., Rath, A., Chong, J. X., Hartley, T., Alkuraya, F. S., Baynam, G., et al. (2017). International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am. J. Hum. Genet.* 100 (5), 695–705. doi:10.1016/j.ajhg.2017.04.003

Chen, T., and Guestrin, C. "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, August 2016, 785–794.

Chen, X., Liu, X., Sheng, X., Gao, X., Zhang, X., Li, Z., et al. (2015). Targeted Next-Generation Sequencing Reveals Novel EYS Mutations in Chinese Families with Autosomal Recessive Retinitis Pigmentosa. *Sci. Rep.* 5 (1), 8927. doi:10.1038/srep08927

Consortium, U. (2019). UniProt: a Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47 (D1), D506–D515. doi:10.1093/nar/gky1049

Espreafico, E. M., Cheney, R. E., Matteoli, M., Nascimento, A. A., De Camilli, P. V., Larson, R. E., et al. (1992). Primary Structure and Cellular Localization of Chicken Brain Myosin-V (P190), an Unconventional Myosin with Calmodulin Light Chains. *J. Cel Biol.* 119 (6), 1541–1557. doi:10.1083/jcb.119.6.1541

Fan, B. J., Tam, P. O. S., Choy, K. W., Wang, D. Y., Lam, D. S. C., and Pang, C. P. (2006). Molecular Diagnostics of Genetic Eye Diseases. *Clin. Biochem.* 39 (3), 231–239. doi:10.1016/j.clinbiochem.2005.11.010

Fisher, S. A., Rivera, A., Fritsche, L. G., Keilhauer, C. N., Lichtner, P., Meitinger, T., et al. (2007). Case-control Genetic Association Study of Fibulin-6 (FBLN6orHMCN1) Variants in Age-Related Macular Degeneration (AMD). *Hum. Mutat.* 28 (4), 406–413. doi:10.1002/humu.20464

Huang, L., Xiao, X., Li, S., Jia, X., Wang, P., Sun, W., et al. (2016). Molecular Genetics of Cone-Rod Dystrophy in Chinese Patients: New Data from 61 Probands and Mutation Overview of 163 Probands. *Exp. Eye Res.* 146, 252–258. doi:10.1016/j.exer.2016.03.015

Katako, A., Shelton, P., Goertzen, A. L., Levin, D., Bybel, B., Aljuaid, M., et al. (2018). Machine Learning Identified an Alzheimer's Disease-Related FDG-PET Pattern Which Is Also Expressed in Lewy Body Dementia and Parkinson's Disease Dementia. *Sci. Rep.* 8 (1), 13236. doi:10.1038/s41598-018-31653-6

Lei, X., and Tie, J. (2019). Prediction of Disease-Related Metabolites Using Bi-random Walks. *PloS one* 14 (11), e0225380. doi:10.1371/journal.pone.0225380

Maugeri, A., Klevering, B. J., Rohrschneider, K., Blankenagel, A., Brunner, H. G., Deutman, A. F., et al. (2000). Mutations in the ABCA4 (ABCR) Gene Are the Major Cause of Autosomal Recessive Cone-Rod Dystrophy. *Am. J. Hum. Genet.* 67 (4), 960–966. doi:10.1086/303079

Muller, J., Stoetzel, C., Vincent, M. C., Leitch, C. C., Laurier, V., Danse, J. M., et al. (2010). Identification of 28 Novel Mutations in the Bardet-Biedl Syndrome Genes: the burden of Private Mutations in an Extensively Heterogeneous Disease. *Hum. Genet.* 127 (5), 583–593. doi:10.1007/s00439-010-0804-9

Oishi, M., Oishi, A., Gotoh, N., Ogino, K., Higasa, K., Iida, K., et al. (2014). Comprehensive Molecular Diagnosis of a Large Cohort of Japanese Retinitis Pigmentosa and Usher Syndrome Patients by Next-Generation Sequencing. *Invest. Ophthalmol. Vis. Sci.* 55 (11), 7369–7375. doi:10.1167/iovs.14-15458

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48 (D1), D845–D855. doi:10.1093/nar/gkz1021

Pozo, M. G., Bravo-Gil, N., Méndez-Vidal, C., Montero-de-Espinosa, I., Millán, J. M., Dopazo, J., et al. (2015). Re-evaluation Casts Doubt on the Pathogenicity of Homozygous USH2A p.C759F. *Am. J. Med. Genet. A.* 167 (7), 1597–1600. doi:10.1002/ajmg.a.37003

Rechsteiner, M., Rogers, S. W., "Rechsteiner, M., and Rogers, S. W. (1996). PEST Sequences and Regulation by Proteolysis. *Trends Biochem. Sci.* 2121 (7), 267267–271271. doi:10.1016/s0968-0004(96)10031-1

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., et al. (2010). GeneCards Version 3: the Human Gene Integrator. *Database (Oxford)* 20102010, baq020. doi:10.1093/database/baq020

Sharon, D., Kimchi, A., and Rivolta, C. (2016). OR2W3 Sequence Variants Are Unlikely to Cause Inherited Retinal Diseases. *Ophthalmic Genet.* 37 (4), 366–368. doi:10.3109/13816810.2015.1081252

Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., et al. (2020). The Human Gene Mutation Database (HGMD®): Optimizing its Use in a Clinical Diagnostic or Research Setting. *Hum. Genet.* 139 (10), 1197–1207. doi:10.1007/s00439-020-02199-3

Wennström, A., Ericson, L., and García-Guzmán, G. (2003). The Concept of Sexually Transmitted Diseases in Plants: Definition and Applicability. *Oikos* 100 (2), 397–402. doi:10.1034/j.1600-0706.2003.12004.x

Wissinger, B., Gamer, D., Jägle, H., Giorda, R., Marx, T., Mayer, S., et al. (2001). CNGA3 Mutations in Hereditary Cone Photoreceptor Disorders. *Am. J. Hum. Genet.* 69 (4), 722–737. doi:10.1086/323613

Xu, Y., Guan, L., Shen, T., Zhang, J., Xiao, X., Jiang, H., et al. (2014). Mutations of 60 Known Causative Genes in 157 Families with Retinitis Pigmentosa Based on Exome Sequencing. *Hum. Genet.* 133 (10), 1255–1271. doi:10.1007/s00439-014-1460-2

Zhang, Q., Li, S., Xiao, X., Jia, X., and Guo, X. (2007). The 208delG Mutation inFSCN2Does Not Associate with Retinal Degeneration in Chinese Individuals. *Invest. Ophthalmol. Vis. Sci.* 48 (2), 530–533. doi:10.1167/iovs.06-0669

Zhao, T., Hu, Y., and Cheng, L. (2021). Deep-DRM: a Computational Method for Identifying Disease-Related Metabolites Based on Graph Deep Learning Approaches. *Brief Bioinform* 22 (4), bbaa212. doi:10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Valsdottir, L. R., Zang, T., and Peng, J. (2021). Identifying Drug-Target Interactions Based on Graph Convolutional Network and Deep Neural Network. *Brief. Bioinformatics* 22 (2), 2141–2150. doi:10.1093/bib/bbaa044

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021). Prediction and Collection of Protein–Metabolite Interactions. *Brief. Bioinform.* 22, bbab014. doi:10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021). SC2disease: a Manually Curated Database of Single-Cell Transcriptome for Human Diseases. *Nucleic Acids Res.* 49 (D1), D1413–D1419. doi:10.1093/nar/gkaa838

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 Molecular Biosciences

ORIGINAL RESEARCH
published: 09 March 2022
doi: 10.3389/fmolb.2022.846579

CRISPR Detection and Research on Screening Mutant Gene *TTN* of Moyamoya Disease Family Based on Whole Exome Sequencing

Yilei Xiao[1], Weidong Liu[1], Jiheng Hao[1], Qunlong Jiang[1], Xingbang Wang[2], Donghu Yu[3], Liyong Zhang[1], Zhaogang Dong[4]* and Jiyue Wang[1]**

[1]Department of Neurosurgery, Liaocheng People's Hospital, Liaocheng, China, [2]Department of Geriatric Medicine, Qilu Hospital of Shandong University, Ji'nan, China, [3]Department of Neurosurgery, Zhongnan Hospital of Wuhan University, Wuhan, China, [4]Department of Clinical Laboratory, Qilu Hospital of Shandong University, Ji'nan, China

Moyamoya disease (MMD) has a high incidence in Asian populations and demonstrates some degree of familial clustering. Whole-exome sequencing (WES) is useful in establishing key related genes in familial genetic diseases but is time-consuming and costly. Therefore, exploring a new method will be more effective for the diagnosis of MMD. We identified familial cohorts showing MMD susceptibility and performed WES on 5 affected individuals to identify susceptibility loci, which identified point mutation sites in the titin (*TTN*) gene (rs771533925, rs559712998 and rs72677250). Moreover, *TTN* mutations were not found in a cohort of 50 sporadic MMD cases. We also analyzed mutation frequencies and used bioinformatic predictions to reveal mutation harmfulness, functions and probabilities of disease correlation, the results showed that rs771533925 and rs72677250 were likely harmful mutations with GO analyses indicating the involvement of *TTN* in a variety of biological processes related to MMD etiology. CRISPR-Cas12a assays designed to detect *TTN* mutations provided results consistent with WES analysis, which was further confirmed by Sanger sequencing. This study recognized *TTN* as a new familial gene marker for moyamoya disease and moreover, demonstrated that CRISPR-Cas12a has the advantages of rapid detection, low cost and simple operation, and has broad prospects in the practical application of rapid detection of MMD mutation sites.

Keywords: moyamoya disease, *TTN*, CRISPR-Cas12a, RNF$_{213}$, MMP3

OPEN ACCESS

Edited by:
*Tianyi Zhao,
Harbin Institute of Technology, China*

Reviewed by:
*Feng Wang,
Affiliated Hospital of Nantong University, China
Cunyou Zhao,
Southern Medical University, China*

****Correspondence:***
*Liyong Zhang
13346256936@163.com
Zhaogang Dong
zhaogang.dong@email.sdu.edu.cn
Jiyue Wang
wangjiyue002100@163.com*

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *31 December 2021*
Accepted: *22 February 2022*
Published: *09 March 2022*

Citation:
Xiao Y, Liu W, Hao J, Jiang Q, Wang X, Yu D, Zhang L, Dong Z and Wang J (2022) CRISPR Detection and Research on Screening Mutant Gene TTN of Moyamoya Disease Family Based on Whole Exome Sequencing. Front. Mol. Biosci. 9:846579. doi: 10.3389/fmolb.2022.846579

INTRODUCTION

Moyamoya disease (MMD) is a chronic progressive, cerebrovascular, and occlusive disease of unknown etiology first reported by Suzuki in 1969 (Kuroda and Houkin, 2008). Compared with western country, the incidence of MMD is higher in China, Korea and Japan, among which MMD is the main cause of stroke in children and adolescents (Kim, 2016; Zhao et al., 2018; Deng et al., 2021).

Abbreviations: MMD, Moyamoya disease; RNF213, ring finger protein 213; PAM, protospacer adjacent motif; PCR, polymerase chain reactions; IRB, Institutional Review Board; IBD, identity-by-descent; SNVs, Single nucleotide variants; gVCF, Genome VCF; CNVs, Copy number variants; ACMG, American College of Medical Genetics and Genomics; CT, computed tomography; CTA, Computed Tomography Angiography; WES, whole-exome sequencing.

Previous studies have shown a higher incidence of moyamoya disease in East Asia, among which, particularly in China, the incidence of moyamoya disease in the north is significantly higher than in the south (Hu et al., 2017). In recent years, a number of studies have confirmed a genetic susceptibility for MMD, proposing that genetic factors play a major role in the pathogenesis of MMD (Liu et al., 2011; Morito et al., 2014; Kobayashi et al., 2015; Kim, 2016). For example, 10–15% of MMD patients have a family history, and the prevalence of these people with a family history is 30–40% higher than that of ordinary people (Kim, 2016). Therefore, it is easier to obtain potential genetic related genes through the research on family patient.

The first pathogenic gene to be associated with MMD was the ring finger protein 213 (RNF213) (Kamada et al., 2011). Moreover, two mutations within the *RNF213* gene (rs112735431 and rs148731719) were known to be associated with MMD pathogenesis in Chinese patients (Wu et al., 2012; Zhang et al., 2017; Wang Y. et al., 2020). In 2010, researchers discovered that the-1171 locus of the *MMP3* gene in Chinese Han patients was closely related to the onset of MMD (Li et al., 2010); this work also represented the first research on susceptibility genes in China. Other studies have also shown that 6–10% of Chinese MMD cases are likely to be familial in origin (Hishikawa et al., 2013). In addition, a novel missense mutation 377T > C and two polymorphisms (420A > G and 487C > T) in the TGIF gene were identified in a Taiwanese family segregated with holoprosencephaly (HPE) and moyamoya disease, speculated the possible association between TGIF mutation and MMD (Chen et al., 2006). An extensive genetic study on specific gene in MMD patients might shed light on the pathogenesis of MMD. Our previous studies have shown that specific gene mutations does not lead to inheritance of the disease. To some extent, our data can serve as a useful complement to family-based research.

With the development of high-throughput sequencing technology, WES has been increasingly utilized in the study of Mendelian diseases and complex diseases. The human exome region accounts for only 1% of the entire genomic sequence, but approximately 85% of known pathogenic mutations are located in coding regions (Manolio et al., 2009). Notably, traditional mutation site screening mostly uses Sanger sequencing or WES, which is time-consuming and costly, not being beneficial to the large-scale screening of samples. The CRISPR-Cas system is an important immune defense system of Archaea and bacteria against viral and plasmid infection (Ishino et al., 1987; Jansen et al., 2002; Mojica et al., 2005). Cas12a (cpf1) is a new type of programmable DNA enzyme found in the CRISPR system and contains an RuvC domain and a specific nuclease domain (Zhou et al., 2014). Some studies have found that Cas12a also has the ability to cut non-target DNA following cleavage of the target DNA (Gilbert et al., 2013; Qi et al., 2013). The CRISPR-Cas system has extremely high sensitivity and efficiency in the detection of nucleic acids, which has changed the process of molecular diagnosis of various diseases (Chertow, 2018).

In the pre-experiment, we verified the utility of the CRISPR-Cas12a and Sanger to detect specific gene (RNF213 and MMP3) mutations. In this study, we used WES to analyze familial cases of MMD from Chinese patients. The CRISPR-Cas12a system was used to screen the mutation loci of disease-related families and identify related genes, thereby uncovering the molecular basis of MMD.

METHODS

Collection of Clinical Samples

We recruited MMD patients (≥18 years old and ≤70 years old, male: female = 1:1) without previous medical history. Diagnostic criteria were based on the Japanese Research Committee on moyamoya disease of the Ministry of Health, Welfare and Labour, Japan (RCMJ) criteria (Research Committee on the Pathology and Treatment of Spontaneous Occlusion of the Circle of Willis and Health Labour Sciences Research Grant for Research on Measures for Infractable Diseases, 2012). Their clinical diagnosis was confirmed by imaging with transcranial computed tomography (CT), magnetic resonance imaging (MRI), or digital subtraction angiography (DSA) along with various clinical judgments. Fasting samples of venous blood were collected from all patients and healthy control subjects separately during the same period. All subjects signed the consent form prior to entering the trial.

Primer Design and Preparation of crRNA

Wild-type and mutant templates were designed with reference to the known mutation detection loci for the specific gene. Amplimers and crRNAs were then designed for the known mutation regions and oligonucleotides (crDNA) were synthesized. crDNA and cr-T7-F were mixed and boiled for 10 min, then the double-stranded transcription template being formed by natural cooling. The transcription template was then incubated for 16 h at 37°C under enzymatic-free conditions using the HiScribe T7 Quick High Yield RNA Synthesis Kit (NEB, Ipswich, United States). After the completion of the reaction, 2 μL of DNase 1 (TianGen, Beijing, China) was added to eliminate unreacted template before purifying the crRNA. Wild-type and mutant template sequences, amplimers, and crDNAs, were synthesized by Tianyi Huiyuan Biotechnology Co., Ltd. (**Supplementary Table S1**).

Validation of the CRISPR-Cas12a Fluorescence Detection System

Fncas12a uses 5′-KYTV-3′ 999 as protospacer adjacent motif (PAM). It was chosen as the detection protein for providing more target sequence options compared with Ascas12a and Lbcas12a (Tu et al., 2017). In brief, 50 ng of template DNA was added into the detection reagent mixture containing 0.75 μM crRNA, 1.5 μM Fncas12a, 50pM of fluorescent probe, and 3 μL of NEBuffer 3.1 (NEB, Ipswich, United States). Reactions (50 μL) were then incubated at 37°C for 1h prior to fluorescence quantification. All reactions were carried out at 37°C.

TABLE 1 | Sample information.

Specimen No.	Sex (male/ female)	Patient or Normal (*P*: Patient; N: Normal)
B1	F	*P*
B2	M	N
B3	M	N
B4	M	N
B5	F	P

Clinical Sample Testing

Following plasma separation, DNA was extracted from venous blood samples. Thereafter, polymerase chain reactions (PCR) were performed using 50 ng of DNA as the template with specific primers (**Supplmentary Table S1**) at the following cycle conditions: 95°C for 5 min; 30 cycles of 95°C for 3 min; 56°C for 10 s, and 72°C for 20 s; followed by 72°C for 5 min. PCR products were then visualized by agarose gel electrophoresis and were sequenced using the Sanger method. In parallel, 1–5 μl of amplified product was used for CRISPR-Cas12a fluorescence detection.

Collection and Selection of Samples for Whole-Exome Sequencing

We collected five samples from the familiy with clinical manifestations of the MMD phenotype from Liaocheng People's Hospital Center from June 2020 to December 2020 (**Table 1**). All five family members were subjected to WES as depicted in the flow chart in **Supplementary Figure S1**. This study was approved by the ethics committee of Liaocheng People's Hospital, Shandong Province. Informed consent for DNA analysis was obtained from patients in line with local Institutional Review Board (IRB) requirements at the time of collection.

Library Construction for Whole-Exome Sequencing

DNA extracted from peripheral blood was fragmented to an average size of 180–280 bp and subjected to DNA library creation using established Illumina paired-end protocols. The Agilent SureSelect Human All ExonV6 Kit (Agilent Technologies, Santa Clara, CA, United States) was used for exome capture according to the manufacturer's instructions. The Illumina NovaSeq 6,000 platform (Illumina Inc., San Diego, CA, United States) was utilized for genomic DNA sequencing in Novogene Bioinformatics Technology Co., Ltd. (Beijing, China) to generate 150-bp paired-end reads with a minimum coverage of 10× for 99% of the genome (mean coverage of 100×).

Whole-Exome Sequencing Data Analysis

After sequencing, base-call file conversions and demultiplexing were performed with bcl2fastq software (Illumina). The resulting fastq data were submitted to in-house quality control software to remove low quality reads; and these were then aligned to the reference human genome (hs37d5) using the Burrows-Wheeler Aligner (bwa) (Li and Durbin, 2009). Duplicate reads were marked using sambamba tools (Tarasov et al., 2015). Single nucleotide variants (SNVs) and indels were identified by samtools to generate Genome VCF (gVCF) (Li et al., 2009). Raw calls for the SNVs and INDELs were further filtered with the following inclusion thresholds: 1) a read depth > 4; 2) a root-mean-square mapping quality of covering reads that was > 30; and 3) a variant quality score > 20. Copy number variants (CNVs) were detected with CoNIFER software (Version 0.2.2) (Krumm et al., 2012). Annotation was performed using ANNOVAR (2017) (Wang et al., 2010). Annotations included minor allele frequencies from public control data sets as well as deleteriousness and conservation scores, thus enabling further filtering and assessment of the likely pathogenic variants.

Selection of Candidate Mutation Loci

Filtering for rare variants was performed as follows. First, variants with a MAF < 0.01 in 1000 genomic data (1000g_all) (Auton et al., 2015), esp6500siv2_all, and gnomAD data (gnomAD_ALL and gnomAD_EAS); (Kim, 2016) only SNVs occurring in exons or splice sites (splicing junction 10 bp) were further analyzed since we were targeting amino acid changes; (Deng et al., 2021) synonymous single nucleotide variants (SNVs) which were not relevant to the amino acid changes predicted by dbscSNV were discarded; the small fragment non-frameshift (<10bp) indel in the repeat region defined by RepeatMasker was discarded; and (Zhao et al., 2018) variations were screened according to SIFT scores (Kumar et al., 2009), PolyPhen (Adzhubei et al., 2010), MutationTaster (Schwarz et al., 2010) and CADD (Kircher et al., 2014) software packages. Potentially deleterious variations were reserved if the scores from more than half of the four software packages identified the variations as harmful (Muona et al., 2015). Sites (>2bp) that did not affect alternative splicing were also removed. To better predict the harmfulness of each variation, we applied the classification system put forward by the American College of Medical Genetics and Genomics (ACMG). The variations were classified as pathogenic, likely to be pathogenic, of uncertain significance, likely to be benign, or benign (Richards et al., 2015). Depending upon various considerations (pedigree, homozygous, and compound heterozygous), variants were considered to be candidate causal variations. The relationship between the proband and the parents was estimated using the pairwise identity-by-descent (IBD) calculation in PLINK (Purcell et al., 2007). The share of IBD between the proband and parents for all trios ranged from 45 to 55%.

Statistical Analysis

SPSS 17.0 software was used for statistical analysis. The qualitative data and the number of cases described in percentage, and the quantitative data were compared by independent sample *t*-test or analysis of variance. $p < 0.05$ indicates a significant difference.

RESULTS

The Ability of CRISPR-Cas12a to Detect Mutations

Literature searches identified *RNF213* as a susceptibility gene for MMD. In addition, two SNP loci of *RNF213*, rs112735431 and rs148731719 have been confirmed closely related to MMD (Liu et al.,

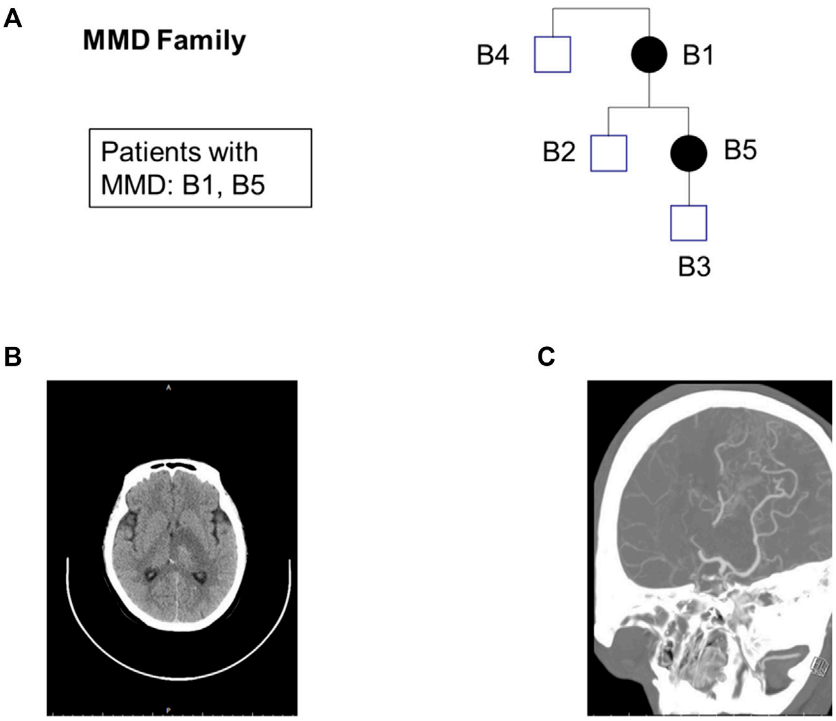


FIGURE 1 | Exome sequencing maps for the MMD family. **(A)** Pedigree charts. Squares: male; circles: female; black-filled symbols: patients; **(B)** CT of patient B1; **(C)** CTA (Computed Tomography Angiography) of patient B1.

2011; Zhang et al., 2017; Wang Y. et al., 2020). crRNA was designed to detect these two SNP point mutation loci in *RNF213*. The cleavage efficiency of the crRNAs was then verified against wild-type and mutant-target DNA (**Supplementary Figure S3**). The fluorescence levels derived from the mutant were significantly higher than the wild type ($p < 0.05$), indicating that the CRISPR-Cas12a system constructed with the indicated crRNAs could successfully detect whether there was a mutation at this locus in clinical samples.

Detection of *RNF213* Gene Locus by CRISPR-Cas12a and Sanger Sequencing

We collected 34 samples of patients who had been clinically diagnosed with MMD and 37 healthy control samples from Liaocheng People's Hospital. DNA was extracted from these samples and the *RNF213* gene of samples was tested using the CRISPR along with Sanger sequencing (**Supplementary Table S2**). The coincidence rate of the CRISPR-Cas12a system and Sanger sequencing for detecting mutation samples was 100%, indicating that the CRISPR-Cas12a detection is accurate and highly sensitive.

Analysis of the Correlation Between Gene (*RNF213*, *MMP3*) Mutations and MMD

First, the results of the Sanger test for RNF213 showed that there was a C > T mutation at locus rs112735431 and a G > A mutation at locus rs148731719 in the *RNF213* gene (**Supplementary Figure S4A**). T-tests showed that the p value for the rs112735431 locus mutation was < 0.05 when comparing between the case group and the healthy control group from the Liaocheng area. In contrast, there was no significant difference between the groups with respect to rs148731719 ($p > 0.05$) (**Supplementary Table S3**), indicating that the rs112735431 mutation within the *RNF213* gene was significant ($p < 0.05$) and that the rs112735431 was a significant mutation locus for MMD in the *RNF213* gene.

Then, we identified a base insertion mutation (rs3025058) in the *MMP3* gene (**Supplementary Figure S4B**). This mutation was identified by Sanger sequencing and detected in 67.6% of the 34 patients with MMD in Shandong province, and 5.4% of the 37 controls, indicating statistical significance ($p < 0.05$). The 1171 (6A/6A) mutation in the *MMP3* gene is associated with the risk of MMD. furthermore, the risk of the (6A/6A) genotype is higher than that of the (5A/6A) genotype (**Supplementary Table S3**.

Whole-Exome Sequencing

The pedigrees of five samples and the results of the patient's CT and CAT tests are shown in **Figure 1**. The average sequencing depth of the five samples exceeded 100×, and the coverage of regions > 10× exceeded 99%. The number of SNVs and Indels obtained from each sample after data analysis are shown in **Supplementary Table S4**.

Screening for Candidate Pathological Changes

Mutation loci were screened in accordance with the scores predicted by SIFT, PolyPhen, MutationTaster, and CADD.

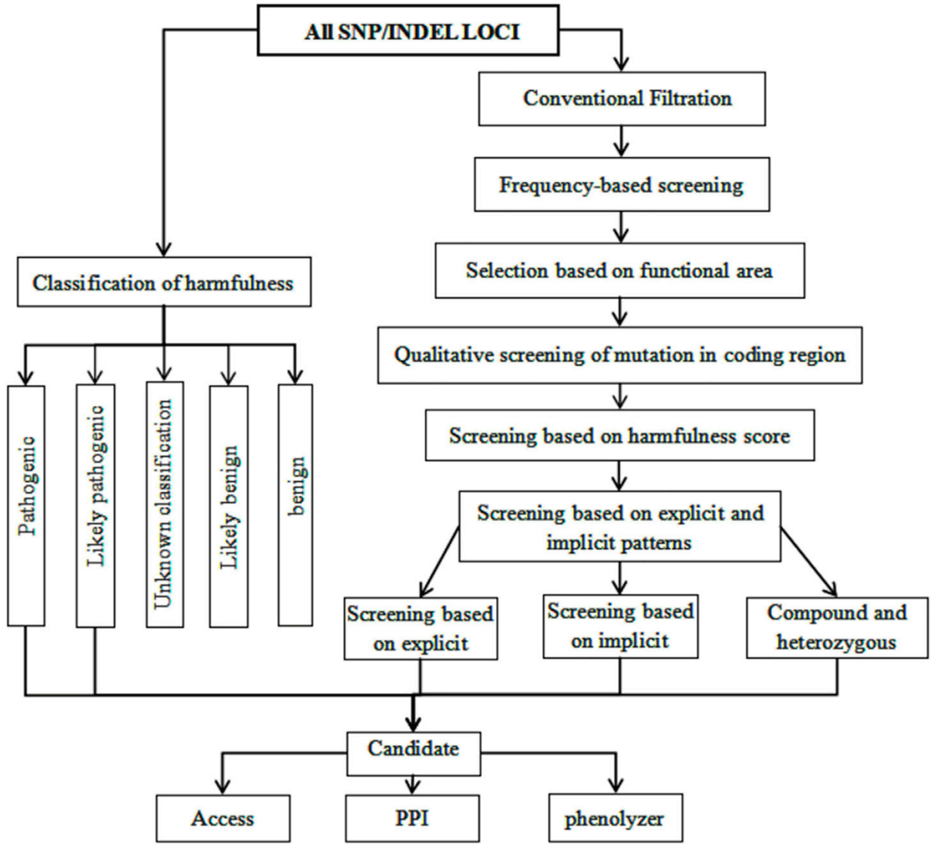

All SNP/INDEL LOCI
Conventional Filtration
Frequency-based screening
Selection based on functional area
Qualitative screening of mutation in coding region
Screening based on harmfulness score
Screening based on explicit and implicit patterns
Screening based on explicit
Screening based on implicit
Compound and heterozygous
Classification of harmfulness
Pathogenic
Likely pathogenic
Unknown classification
Likely benign
benign
Candidate
Access
PPI
phenolyzer

FIGURE 2 | Analysis Flow Chart 2. Advanced analysis pipeline: Screening based on mutation sites and their harmfulness; Screening based on sample recessive patterns; Screening based on candidate genes and relationship with disease phenotypes; Pathway enrichment of candidate genes through GO and KEGG analysis (also using DisGeNet and Phenolyzer to analyze gene-disease phenotype associations).

Candidate loci were further screened according to the process shown in **Figure 2**. The analysis identified multiple recessive pathogenic genes and notably, of these, loci mutation-related genes were within the *TTN* gene (rs771533925, rs559712998 and rs72677250) (**Table 2**).

Validation of Candidate Loci by CRISPR-Cas12a

The test results obtained by the CRISPR-Cas12a system for mutation loci in the *TTN* gene in family samples (**Figure 3**) were consistent with those obtained from WES sequencing (**Table 3**), thus verifying the presence of mutations in the samples.

Validation of Candidate Loci by CRISPR-Cas12a in Sporadic Samples

Next, CRISPR-Cas12a system was used to test a total of 50 sporadic samples for gene mutations. No mutation was found at rs771533925, rs559712998 and rs72677250 of *TTN* gene in sporadic samples (**Figure 4**).

The Deleterious Effects of rs771533925, rs559712998 and rs72677250

In addition, SIFT (Choi and Chan, 2015) PROVEAN (Vaser et al., 2016) and PolyPhen (Adzhubei et al., 2013) algorithms were used to predict the effects of amino acid substitutions on protein function (**Table 4**). All three databases showed that rs771533925 was potentially destructive. On the contrary, rs559712998 was considered tolerable according to these analyses. However, while rs72677250 was considered tolerable according to the SIFT database, it was considered to be potentially harmful according to the PROVEAN and PolyPhen databases.

TTN Mutation Sites rs72677250, rs559712998 and rs771533925 Global Population Frequency and Function Analysis

We analyzed the risk alleles (rs72677250, rs559712998 and rs771533925) in accordance with the EXAC database. We identified significant differences in frequency across the global population. The highest frequency of rs72677250 in the South

TABLE 2 | The detailed information of point mutation site.

Sample ID	Variant	RS ID	Gene	Coding DNA change	Protein change	Zygosity	ACMG	ExonicFunc	SIFT,Polyphen2_HVAR,Polyphen2_HDIV,MutationTaster,CADD
B1	2: 179412799-C-T	rs771533925	TTN	c.G66359A; c.G88631A; c.G66935A; c.G85850A; c.G66734A; c.G93554A	p.R22245H; p.R29544H; p.R22312H; p.R22120H; p.R28617H	het	.	missense SNV	D/D/D/D/24.0
B2	2: 179412799-C-T	rs771533925	TTN	c.G66359A; c.G88631A; c.G66935A; c.G85850A; c.G66734A; c.G93554A	p.R22245H; p.R29544H; p.R22312H; p.R22120H; p.R28617H	het	.	missense SNV	D/D/D/D/24.0
B1	2: 179466289-C-T	rs559712998	TTN	c.G28615A; c.G47731A; c.G28816A; c.G55435A; c.G50512A; c.G28240A	p.V16838I; p.V9539I; p.V9606I; p.V9414I; p.V15911I	het	.	missense SNV	T/B/B/D/20.2
B4	2: 179466289-C-T	rs559712998	TTN	c.G28615A; c.G47731A; c.G28816A; c.G55435A; c.G50512A; c.G28240A	p.V16838I; p.V9539I; p.V9606I; p.V9414I; p.V15911I	het	.	missense SNV	T/B/B/D/20.2
B5	2: 179466289-C-T	rs559712998	TTN	c.G28615A; c.G47731A; c.G28816A; c.G55435A; c.G50512A; c.G28240A	p.V16838I; p.V9539I; p.V9606I; p.V9414I; p.V15911I	het	.	missense SNV	T/B/B/D/20.2
B3	2: 179476144-C-T	rs72677250	TTN	c.G24193A; c.G43108A; c.G23992A; c.G45889A; c.G50812A; c.G23617A	p.E15297K; p.E8065K; p.E7998K; p.E14370K; p.E7873K	het	.	missense SNV	T/P/D/D/23.7
B5	2: 179476144-C-T	rs72677250	TTN	c.G24193A; c.G43108A; c.G23992A; c.G45889A; c.G50812A; c.G23617A	p.E15297K; p.E8065K; p.E7998K; p.E14370K; p.E7873K	het	.	missense SNV	T/P/D/D/23.7

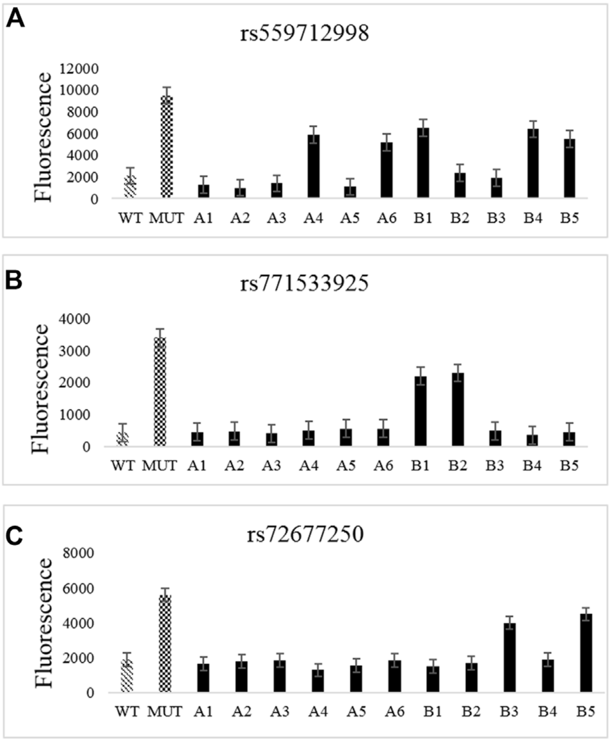


FIGURE 3 | CRISPR-Cas12a analysis of *TTN* gene mutation loci in familial samples. **(A–C)** CRISPR-Cas12a test results for rs72677250 **(A)**, rs559712998 **(B)**, and rs771533925 **(C)**.

TABLE 3 | Analysis of TTN Gene Mutation Results by CRISPR test and Sanger Sequencing in Family Samples.

TTN detection site		Sanger	CRISPR-Cas12a
		N = 5	N = 5
RS559712998	MUT	3 (60%)	3 (60%)
	WILD	2 (40%)	2 (40%)
RS771533925	MUT	2 (40%)	2 (40%)
	WILD	3 (60%)	3 (60%)
RS72677250	MUT	2 (40%)	2 (40%)
	WILD	3 (60%)	3 (60%)

Asian population was 0.00003269, the highest frequency of rs559712998 in the East Asian population was 0.002574, the highest frequency of rs771533925 in the East Asian population was 0.00005568, and the total frequency of rs559712998 mutations was 0.000192; the latter being the highest frequency of all three mutation sites (**Table 5**). According to age analysis of these three loci within the global population, we found that the rs72677250 mutation site was predominant in subjects aged 50–55 years, the rs559712998 mutation site was predominant in subjects aged 30–80 years, and the rs771533925 mutation site was predominant in subjects aged 65–70 years (**Figures 5A–C**). GO analysis was then conducted using Cytoscape 3.8.2 software with the ClueGO (Bindea et al., 2009) plugin, showing that the mutation locus for *TTN* were involved in a range of important biological processes, including myosin thick filament assembly in skeletal muscle, positive regulation of protein transport, serine/threonine kinase activity, and cardiac muscle fiber development (**Figure 5D**).

DISCUSSION

Moyamoya disease is a chronic and progressive disease that can cause cerebral ischemia, cerebral infarction, cerebral hemorrhage, etc., which is a great harm to patients (Kuroda and Houkin, 2008). Patients suffering from moyamoya disease generally could not heal themselves without scientific treatment, and even the condition may continue to aggravate, causing irreversible harm, and bringing great economic burdens to patients and their families to a certain extent (Zhang et al., 2022).

Screening family genetic patients to obtain new or known gene mutations, whole-exome sequencing has the advantages of accuracy and comprehension (Zhang et al., 2021). However, whole-exome sequencing has drawbacks such as time-consuming and high cost, which is not conducive to the large-scale screening of samples. On this basis, the CRISPR technology is used to detect new or known disease-causing gene loci, filling the blank of large-scale sample screening in terms of gene sequencing.

The CRISPR-Cas system can recognize foreign DNA or RNA, directing cleavage to silence the expression of the foreign gene (Brouns et al., 2008; Marraffini and Sontheimer, 2008; Garneau et al., 2010). It can be identified as an efficient gene editing tool for its precise targeting ability (Nelles et al., 2016). Studies have indicated that a diagnostic platform based on CRISPR-Cas represents an exciting prospect for the detection of cancer and genetic diseases (Mali et al., 2013). Cas12a (cpf1) is a new type of programmable DNA enzyme found in the CRISPR system (Zhou et al., 2014). In the presence of specific directing crRNA, Cas12a also has the ability to cut non-target DNA after cleavage of the target DNA (Gilbert et al., 2013; Qi et al., 2013). Therefore, the CRISPR-Cas12a system can be more effective for *in vitro* detection by adding a fluorescent DNA reporter (Mohanraju et al., 2016; Nelles et al., 2016; Koonin et al., 2017) which can emit detectable fluorescence after cleavage. This provides a fluorescence-based assay which only requires low technology instrumentation such as a microplate reader to provide quantitative measurements of mutations.

The rs112735431 and rs148731719 mutations in the *RNF213* gene are known to be associated with the pathogenesis of MMD in Chinese subjects (Liu et al., 2011; Morito et al., 2014; Kobayashi et al., 2015; Hu et al., 2017). *RNF213* is located on human chromosome 17 (the 17q25.3 region) and its expression occurs in different organs (Kuriyama et al., 2008). An imbalance leads to vascular smooth muscle hyperplasia and thickening, thus leading to vascular stenosis, one of the key pathogenic factors responsible for MMD (Li et al., 2010). Additionally, other studies have shown that the 1171 (6A/6A) mutation in the *MMP3* gene is associated with heightened MMD susceptibility with the risk of the (6A/6A) genotype being higher than the (5A/6A) genotype (Wang et al., 2013; Ma and You, 2015; Wang X. et al., 2020). Preliminary experiments analyzed rs112735431 and rs148731719 mutations in the *RNF213* gene in MMD patients and healthy control subjects. In the pre-experiment, we discovered that it

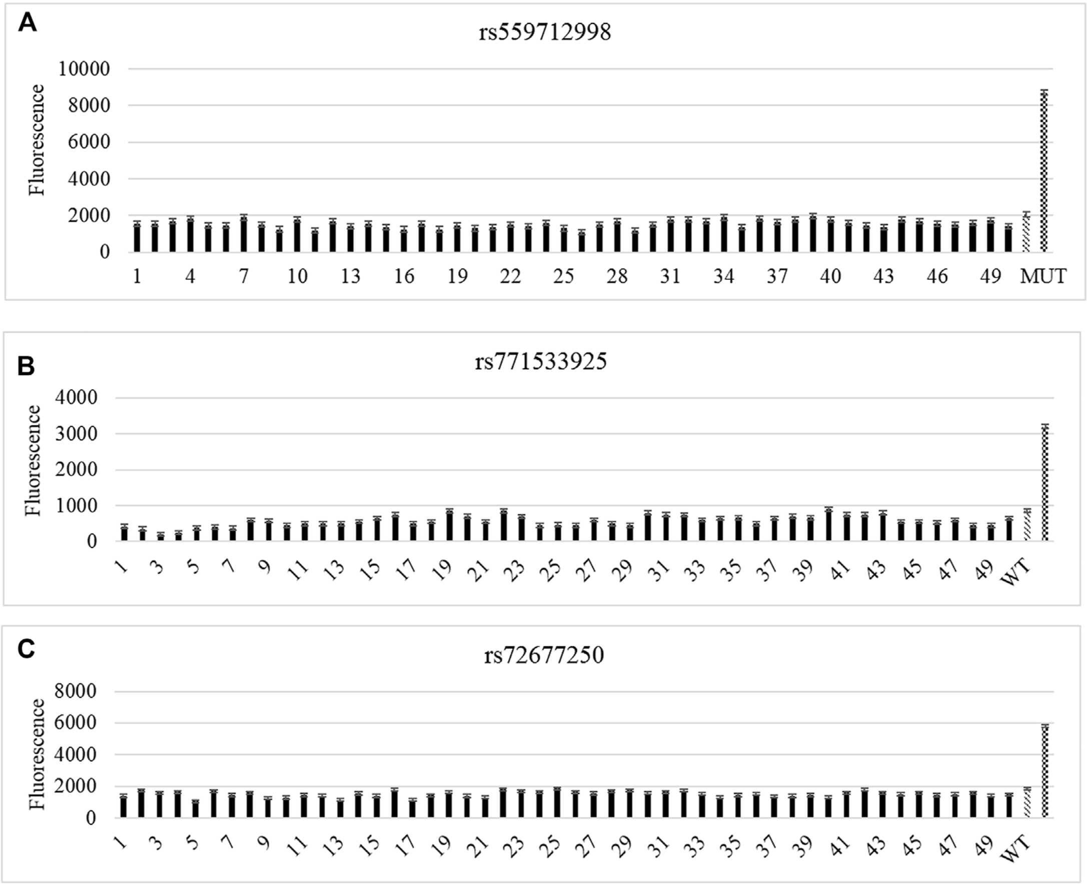


FIGURE 4 | CRISPR-Cas12a analysis of *TTN* gene mutation loci in sporadic samples. **(A–C)** CRISPR-Cas12a test results for rs559712998, rs72677250 **(A),** rs771533925 **(B)**, and rs72677250 **(C)**.

TABLE 4 | Hazard prediction of RS771533925, RS559712998 and RS72677250 mutations.

	Gene	PROVEAN prediction	SIFT prediction	Polyphen
rs771533925	TTN	Deleterious	Damaging	possibly_damaging
rs559712998	TTN	Neutral	Tolerated	benign
rs72677250	TTN	Deleterious	Tolerated	possibly_damaging

PROVEAN (Protein Variation Effect Analyzer) is a tool to predict whether biomolecular structure Variation affects Protein function; SIFT(sorts intolerant from tolerant) is a tool for predicting non-synonymous variations based on sequence homology; PolyPhen (Polymorphism Phenotyping) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.

was the rs112735431 *RNF213* gene mutation but not the rs148731719 mutation affecting the occurrence and development of MMD. At present, Sanger sequencing is mostly carried out for cerebrovascular diseases, and CRISPR technology is rarely studied. Therefore, we first used CRISPR-Cas12a system to compare the technical feasibility. The results showed that rs112735431 and rs148731719 mutations of the RNF213 gene were successfully detected by the CRISPR-Cas12a system with 100% agreement with the results of Sanger sequencing.

In this study, we performed WES on five family members of the MMD family to identify MMD genetic-related mutation loci, establishing a new candidate susceptibility loci in the *TTN* gene. We also detected mutant loci in MMD patients and healthy controls to investigate differences in the mutation loci across the population using CRISPR-Cas12a assays. Then, we compared CRISPR-Cas12a technology with Sanger sequencing and WES for the detection of mutations to highlight the diagnostic efficacy of CRISPR-Cas12a. Finally, we conducted the analysis of population frequency, harmfulness, and functional enrichment on *TTN*.

Our WES analysis also identified a number of recessive pathogenic genes in five members of two MMD families. The *TTN* was identified as the gene containing mutation-related

TABLE 5 | Analysis of RNF213 and MMP3 gene mutation.

Gene	Genotype		Sanger sequencing results	Control group
			Liao cheng	
			Case group	
			(*n* = 34)	(*n* = 37)
RNF213	rs112735431	Mutation wild	8 (23.5%)	0 (0%)
			26 (76.4%)	37 (100%)
	p value		0.0019	
	rs148731719	Mutation	4 (11.8%)	3 (8.1%)
		wild	30 (88.2%)	34 (91.9%)
	p value		0.6082	
MMP3	6A6A		23 (67.6%)	2 (5.4%)
	5A6A		11 (32.4%)	35 (94.6%)
	5A5A		0 (0%)	0 (0%)
	p value		0.00001	
	6A allele frequency		57 (83.8%)	39 (52.7%)
	5A allele frequency		11 (16.2%)	35 (47.3%)
	p value		0.0001	

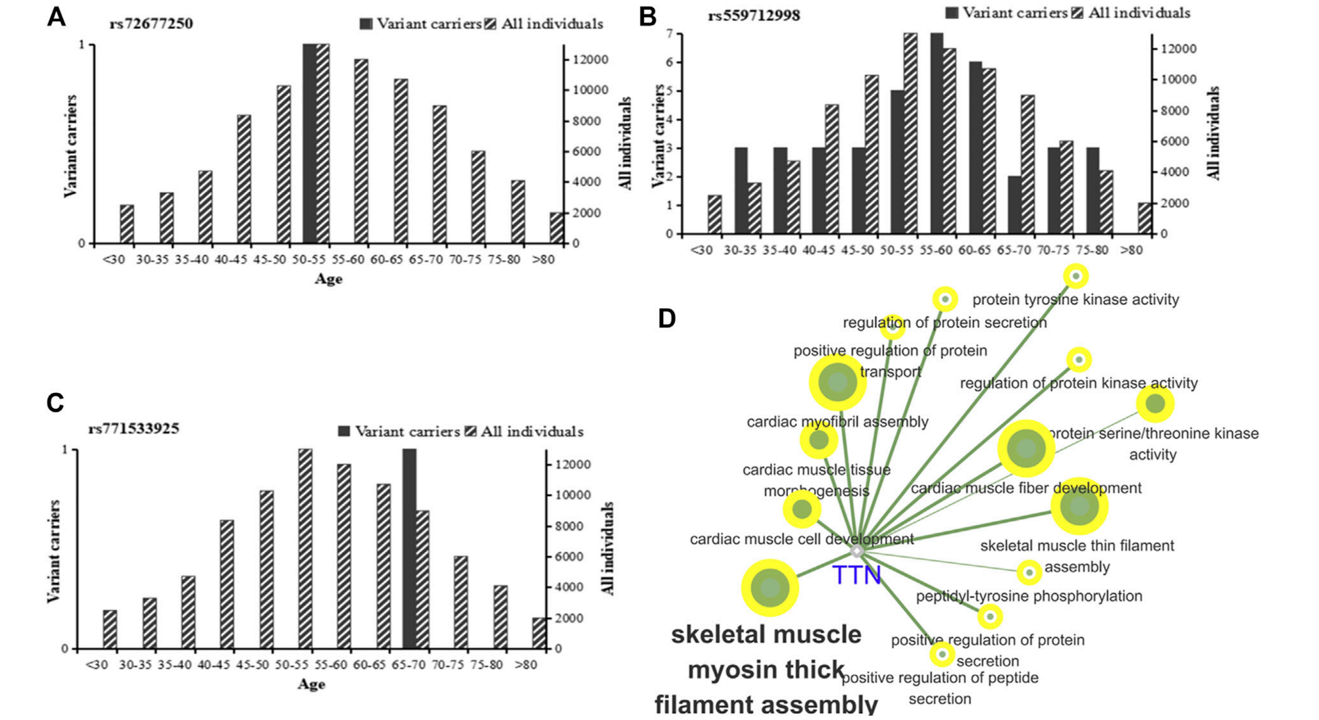


FIGURE 5 | Global population frequency and function analysis of *TTN* mutation sites. **(A–C)** Global population frequencies among different age groups for the rs72677250 mutation **(A)**, rs559712998 mutation **(B)**, and rs771533925 mutation **(C)**. Analysis included heterozygous variant carriers, homozygous variant carriers. **(D)** GO functional enrichment analysis of *TTN* using a two-sided hypergeometric test with Bonferroni correction.

loci (rs771533925, rs559712998 and rs72677250). The EXAC database was used to analyze the risk alleles (rs72677250, rs559712998 and rs771533925). Significant differences were identified in the frequencies of these alleles across the global population. Based on PROVEAN, SIFT, and PolyPhen algorithms, rs771533925 and rs72677250 were considered to be potentially damaging in all three databases where rs559712998 was considered to be tolerable in contrast. GO analysis showed that the targets of *TTN* were involved in many important biological processes. Together with actin and myosin, *TTN* constitute an important component of human cardiac muscle and skeletal muscle. Interestingly, serum antibodies directed against *TTN* were found in patients with melanoma-associated retinopathy, suggesting *TTN* was a potential biomarker for melanoma and also an association with carcinogenesis. Future studies should address the role of *TTN* gene mutations in the pathogenesis of MMD.

In the present study, CRISPR–Cas12a was developed as a novel assay that could sensitively and specifically detect MMD mutation gene loci. Moreover, compared with Sanger sequencing, the CRISPR-Cas12a method is easier, cheaper, and more sensitive for single gene mutations, so it should be promoted to use widely. Also, CRISPR-Cas12a assays were conducted to detect mutations in the candidate genes within the MMD family. Similarly, SNP loci within the *TTN* gene were readily detected with results consistent with the WES analysis. Further detection of mutations in the *TTN* gene in 50 clinical samples revealed that there was no mutation in the *TTN* gene SNP loci and no recessive genetic risk for loci mutations. We speculated that mutations at the *TTN* locus may play an important role in the familial inheritance of MMD. However, our data is limited and a large number of samples are still needed to verify. What's more, these mutations are likely suitable for identifying patient pedigrees and assessing the genetic risk of MMD in large-scale screening.

CONCLUSION

Our study identified *TTN*, a new specific candidate gene in familial moyamoya disease. We also established that CRISPR-Cas12a assays, which can effectively detect MMD mutations, and with significant advantages in time, suggest utility in the rapid detection of MMD mutations. Furthermore, with the detection technology embedded within the reagents, the instrumentation required is comparatively easy, proposing the CRISPR-Cas12a system could be readily developed as accurate, portable diagnostic tests for MMD. Therefore, the CRISPR-Cas12a system can be used to overcome obstacles created by previous platforms and provide a highly sensitive and convenient detection system for MMD mutations with DNA acquired from clinical blood samples.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Liaocheng People's Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YX, WL, and JH conceived the experiments. QJ, XW, and DY conducted the experiments. LZ, ZD, and JW analyzed the results. All authors reviewed the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant number 81701159); the Latitudinal research project of Shandong University (grant number 6010120011); the Key Research Project of Shandong Province (grant number 2018GSF118046); and the Taishan Scholar Project of Shandong Province of China (grant number tsqn202103200).

ACKNOWLEDGMENTS

We thank the technical support and guidance provided by Jiangsu Bojia Biomedical Technology Co., Ltd.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.846579/full#supplementary-material

Supplementary Figure S1 | Analysis flow chart 1. Overview of the information analysis process in three parts: Sequencing data quality assessment to determine whether library sequencing meets the required standards; Variation detection in the sample, with statistics and annotations performed for the detected variations; Variation screening and prediction of disease relevance based on the results of variation detection to identify harmful mutation sites or genes related to disease.

Supplementary Figure S2 | CRISPR-Cas12a analysis of *TTN* gene mutation loci. **(A-D)** CRISPR-Cas12a test results for rs559712998 and rs771533925 mutation type **(A,B)**; rs559712998 and rs771533925 wild type **(C, D)**.

Supplementary Figure S3 | Validation of the CRISPR-Cas12a detection system. **(A and B)** Validation of the CRISPR-Cas12a detection system.

Supplementary Figure S4 | Sequencing peak map of RNF213 and MMP3. **(A, B)** Sequencing peak maps of the RNF213 **(A)** and MMP3 **(B)** genes.

REFERENCES

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, Chapter 7, Unit7.20. doi:10.1002/0471142905.hg0720s76

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* 7 (4), 248–249. doi:10.1038/nmeth0410-248

Auton, A., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A Global Reference for Human Genetic Variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

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 (8), 1091–1093. doi:10.1093/bioinformatics/btp101

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., et al. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* 321 (5891), 960–964. doi:10.1126/science.1159689

Chen, M., Kuo, S.-J., Liu, C.-S., Chen, W.-L., Ko, T.-M., Chen, T.-H., et al. (2006). A Novel Heterozygous Missense Mutation 377T > C (V126A) ofTGIF Gene in a Family Segregated with Holoprosencephaly and Moyamoya Disease. *Prenat. Diagn.* 26 (3), 226–230. doi:10.1002/pd.1385

Chertow, D. S. (2018). Next-generation Diagnostics with CRISPR. *Science* 360 (6387), 381–382. doi:10.1126/science.aat4982

Choi, Y., and Chan, A. P. (2015). PROVEAN Web Server: a Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics* 31 (16), 2745–2747. doi:10.1093/bioinformatics/btv195

Deng, X., Ge, P., Wang, R., Zhang, D., Zhao, J., and Zhang, Y. (2021). Risk Factors for Postoperative Ischemic Complications in Pediatric Moyamoya Disease. *BMC Neurol.* 21 (1), 229–236. doi:10.1186/s12883-021-02283-9

Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., et al. (2010). The CRISPR/Cas Bacterial Immune System Cleaves Bacteriophage and Plasmid DNA. *Nature* 468 (7320), 67–71. doi:10.1038/nature09523

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., et al. (2013). CRISPR-mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* 154 (2), 442–451. doi:10.1016/j.cell.2013.06.044

Hishikawa, T., Tokunaga, K., Sugiu, K., and Date, I. (2013). Clinical and Radiographic Features of Moyamoya Disease in Patients with Both Cerebral Ischaemia and Haemorrhage. *Br. J. Neurosurg.* 27 (2), 198–201. doi:10.3109/02688697.2012.717983

Hu, J., Luo, J., and Chen, Q. (2017). The Susceptibility Pathogenesis of Moyamoya Disease. *World Neurosurg.* 101, 731–741. doi:10.1016/j.wneu.2017.01.083

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide Sequence of the Iap Gene, Responsible for Alkaline Phosphatase Isozyme Conversion in Escherichia coli, and Identification of the Gene Product. *J. Bacteriol.* 169 (12), 5429–5433. doi:10.1128/jb.169.12.5429-5433.1987

Jansen, R., Embden, J. D. A. v., Gaastra, W., and Schouls, L. M. (2002). Identification of Genes that Are Associated with DNA Repeats in Prokaryotes. *Mol. Microbiol.* 43 (6), 1565–1575. doi:10.1046/j.1365-2958.2002.02839.x

Kamada, F., Aoki, Y., Narisawa, A., Abe, Y., Komatsuzaki, S., Kikuchi, A., et al. (2011). A Genome-wide Association Study Identifies RNF213 as the First Moyamoya Disease Gene. *J. Hum. Genet.* 56 (1), 34–40. doi:10.1038/jhg.2010.132

Kim, J. S. (2016). Moyamoya Disease: Epidemiology, Clinical Features, and Diagnosis. *J. Stroke* 18 (1), 2–11. doi:10.5853/jos.2015.01627

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nat. Genet.* 46 (3), 310–315. doi:10.1038/ng.2892

Kobayashi, H., Matsuda, Y., Hitomi, T., Okuda, H., Shioi, H., Matsuda, T., et al. (2015). Biochemical and Functional Characterization of RNF213 (Mysterin) R4810K, a Susceptibility Mutation of Moyamoya Disease, in Angiogenesis *In Vitro* and *In Vivo*. *J. Am. Heart Assoc.* 4 (7), e2146–e2171. doi:10.1161/JAHA.115.002146

Koonin, E. V., Makarova, K. S., and Zhang, F. (2017). Diversity, Classification and Evolution of CRISPR-Cas Systems. *Curr. Opin. Microbiol.* 37, 67–78. doi:10.1016/j.mib.2017.05.008

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., et al. (2012). Copy Number Variation Detection and Genotyping from Exome Sequence Data. *Genome Res.* 22 (8), 1525–1532. doi:10.1101/gr.138115.112

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the Effects of Coding Non-synonymous Variants on Protein Function Using the SIFT Algorithm. *Nat. Protoc.* 4 (7), 1073–1081. doi:10.1038/nprot.2009.86

Kuriyama, S., Kusaka, Y., Fujimura, M., Wakai, K., Tamakoshi, A., Hashimoto, S., et al. (2008). Prevalence and Clinicoepidemiological Features of Moyamoya Disease in Japan. *Stroke* 39 (1), 42–47. doi:10.1161/strokeaha.107.490714

Kuroda, S., and Houkin, K. (2008). Moyamoya Disease: Current Concepts and Future Perspectives. *Lancet Neurol.* 7 (11), 1056–1066. doi:10.1016/s1474-4422(08)70240-0

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352

Li, H., Zhang, Z.-S., Liu, W., Yang, W.-Z., Dong, Z.-N., Ma, M.-J., et al. (2010). Association of a Functional Polymorphism in the MMP-3 Gene with Moyamoya Disease in the Chinese Han Population. *Cerebrovasc. Dis.* 30 (6), 618–625. doi:10.1159/000319893

Liu, W., Morito, D., Takashima, S., Mineharu, Y., Kobayashi, H., Hitomi, T., et al. (2011). Identification of RNF213 as a Susceptibility Gene for Moyamoya Disease and its Possible Role in Vascular Development. *PLoS One* 6 (7), e22542–e22561. doi:10.1371/journal.pone.0022542

Ma, J., and You, C. (2015). Association between Matrix Metalloproteinase-3 Gene Polymorphism and Moyamoya Disease. *J. Clin. Neurosci.* 22 (3), 479–482. doi:10.1016/j.jocn.2014.08.034

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., et al. (2013). RNA-guided Human Genome Engineering via Cas9. *Science* 339 (6121), 823–826. doi:10.1126/science.1232033

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the Missing Heritability of Complex Diseases. *Nature* 461 (7265), 747–753. doi:10.1038/nature08494

Marraffini, L. A., and Sontheimer, E. J. (2008). CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. *Science* 322 (5909), 1843–1845. doi:10.1126/science.1165771

Mohanraju, P., Makarova, K. S., Zetsche, B., Zhang, F., Koonin, E. V., and van der Oost, J. (2016). Diverse Evolutionary Roots and Mechanistic Variations of the CRISPR-Cas Systems. *Science* 353 (6299), aad5147. doi:10.1126/science.aad5147

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005). Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *J. Mol. Evol.* 60 (2), 174–182. doi:10.1007/s00239-004-0046-3

Morito, D., Nishikawa, K., Hoseki, J., Kitamura, A., Kotani, Y., Kiso, K., et al. (2014). Moyamoya Disease-Associated Protein mysterin/RNF213 Is a Novel AAA+ ATPase, Which Dynamically Changes its Oligomeric State. *Sci. Rep.* 4, 4442–4450. doi:10.1038/srep04442

Muona, M., Berkovic, S. F., Dibbens, L. M., Oliver, K. L., Maljevic, S., Bayly, M. A., et al. (2015). A Recurrent De Novo Mutation in KCNC1 Causes Progressive Myoclonus Epilepsy. *Nat. Genet.* 47 (1), 39–46. doi:10.1038/ng.3144

Nelles, D. A., Fang, M. Y., O'Connell, M. R., Xu, J. L., Markmiller, S. J., Doudna, J. A., et al. (2016). Programmable RNA Tracking in Live Cells with CRISPR/Cas9. *Cell* 165 (2), 488–496. doi:10.1016/j.cell.2016.02.054

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., et al. (2013). Repurposing CRISPR as an RNA-Guided Platform for Sequence-specific Control of Gene Expression. *Cell* 152 (5), 1173–1183. doi:10.1016/j.cell.2013.02.022

Research Committee on the Pathology and Treatment of Spontaneous Occlusion of the Circle of Willis; Health Labour Sciences Research Grant for Research on Measures for Infractable Diseases (2012). Guidelines for Diagnosis and Treatment of Moyamoya Disease (Spontaneous Occlusion of the circle of Willis). *Neurol. Med. Chir (Tokyo)* 52 (5), 245–266. doi:10.2176/nmc.52.245

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17 (5), 405–424. doi:10.1038/gim.2015.30

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster Evaluates Disease-Causing Potential of Sequence Alterations. *Nat. Methods* 7 (8), 575–576. doi:10.1038/nmeth0810-575

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: Fast Processing of NGS Alignment Formats. *Bioinformatics* 31 (12), 2032–2034. doi:10.1093/bioinformatics/btv098

Tu, M., Lin, L., Cheng, Y., He, X., Sun, H., Xie, H., et al. (2017). A 'new Lease of Life': FnCpf1 Possesses DNA Cleavage Activity for Genome Editing in Human Cells. *Nucleic Acids Res.* 45 (19), 11295–11304. doi:10.1093/nar/gkx783

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT Missense Predictions for Genomes. *Nat. Protoc.* 11 (1), 1–9. doi:10.1038/nprot.2015.123

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* 38 (16), e164. doi:10.1093/nar/gkq603

Wang, X., Wang, Y., Nie, F., Li, Q., Zhang, K., Liu, M., et al. (2020). Association of Genetic Variants with Moyamoya Disease in 13 000 Individuals. *Stroke* 51 (6), 1647–1655. doi:10.1161/strokeaha.120.029527

Wang, X., Zhang, Z., Liu, W., Xiong, Y., Sun, W., Huang, X., et al. (2013). Impacts and Interactions of PDGFRB, MMP-3, TIMP-2, and RNF213 Polymorphisms on the Risk of Moyamoya Disease in Han Chinese Human Subjects. *Gene* 526 (2), 437–442. doi:10.1016/j.gene.2013.05.083

Wang, Y., Zhang, Z., Wei, L., Zhang, Q., Zou, Z., Yang, L., et al. (2020). Predictive Role of Heterozygous p.R4810K of RNF213 in the Phenotype of Chinese Moyamoya Disease. *Neurology* 94 (7), e678–e686. doi:10.1212/wnl.0000000000008901

Wu, Z., Jiang, H., Zhang, L., Xu, X., Zhang, X., Kang, Z., et al. (2012). Molecular Analysis of RNF213 Gene for Moyamoya Disease in the Chinese Han Population. *PLoS One* 7 (10), e48179–e48188. doi:10.1371/journal.pone.0048179

Zhang, D., Huang, L., Huang, Z., Zhou, Q., Yang, X., Gu, H., et al. (2022). Epidemiology of Moyamoya Disease in China: A Nationwide Hospital-Based Study. *Lancet Reg. Health West. Pac.* 18, 100331. doi:10.1016/j.lanwpc.2021.100331

Zhang, Q., Liu, Y., Zhang, D., Wang, R., Zhang, Y., Wang, S., et al. (2017). RNF213 as the Major Susceptibility Gene for Chinese Patients with Moyamoya Disease and its Clinical Relevance. *J. Neurosurg.* 126 (4), 1106–1113. doi:10.3171/2016.2.jns152173

Zhang, W., Huang, C., and Zhou, W. (2021). Rapid Identification of a Pathogenic Variant of PROS1 in a Thrombophilic Family by Whole Exome Sequencing. *Medicine (Baltimore)* 100 (52), e28436. doi:10.1097/md.0000000000028436

Zhao, M., Deng, X., Zhang, D., Wang, S., Zhang, Y., Wang, R., et al. (2018). Risk Factors for and Outcomes of Postoperative Complications in Adult Patients with Moyamoya Disease. *J. Neurosurg.*130 (5), 1–12. doi:10.3171/2017.10.JNS171749

Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., et al. (2014). High-throughput Screening of a CRISPR/Cas9 Library for Functional Genomics in Human Cells. *Nature* 509 (7501), 487–491. doi:10.1038/nature13166

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 Molecular Biosciences

ORIGINAL RESEARCH
published: 16 March 2022
doi: 10.3389/fmolb.2022.807931

Identification of a 3-Gene Prognostic Index for Papillary Thyroid Carcinoma

Lin-Kun Zhong[1†], Xing-Yan Deng[2†], Fei Shen[3†], Wen-Song Cai[3], Jian-Hua Feng[3], Xiao-Xiong Gan[3], Shan Jiang[4], Chi-Zhuai Liu[1], Ming-Guang Zhang[1], Jiang-Wei Deng[1], Bing-Xing Zheng[1], Xiao-Zhang Xie[1], Li-Qing Ning[1], Hui Huang[1], Shan-Shan Chen[5], Jian-Hang Miao[1] and Bo Xu[2]**

[1]Department of General Surgery, Zhongshan City People's Hospital, Zhongshan, China, [2]Thyroid, Vascular Surgery Department, Maoming People's Hospital, Maoming, China, [3]Department of Thyroid Surgery, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China, [4]Reproductive Medicine Center, Boai Hsopital of Zhongshan, Zhongshan, China, [5]Department of Intensive Care Medicine, Zhongshan City People's Hospital, Zhongshan, China

OPEN ACCESS

Edited by:
Liangcan He,
Harbin Institute of Technology, China

Reviewed by:
Jinxuan Hou,
Wuhan University, China
Xuan Su,
Sun Yat-sen University Cancer Center (SYSUCC), China

***Correspondence:**
Jian-Hang Miao
lymjh@yeah.net
Bo Xu
eyboxu@scut.edu.cn,
https://orcid.org/0000-0001-6384-6685

†These authors share first authorship

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *02 November 2021*
Accepted: *09 February 2022*
Published: *16 March 2022*

Citation:
Zhong L-K, Deng X-Y, Shen F, Cai W-S, Feng J-H, Gan X-X, Jiang S, Liu C-Z, Zhang M-G, Deng J-W, Zheng B-X, Xie X-Z, Ning L-Q, Huang H, Chen S-S, Miao J-H and Xu B (2022) Identification of a 3-Gene Prognostic Index for Papillary Thyroid Carcinoma. Front. Mol. Biosci. 9:807931. doi: 10.3389/fmolb.2022.807931

The accurate determination of the risk of cancer recurrence is a critical unmet need in managing thyroid cancer (TC). Although numerous studies have successfully demonstrated the use of high throughput molecular diagnostics in TC prediction, it has not been successfully applied in routine clinical use, particularly in Chinese patients. In our study, we objective to screen for characteristic genes specific to PTC and establish an accurate model for diagnosis and prognostic evaluation of PTC. We screen the differentially expressed genes by Python 3.6 in The Cancer Genome Atlas (TCGA) database. We discovered a three-gene signature Gap junction protein beta 4 (GJB4), Ripply transcriptional repressor 3 (RIPPLY3), and Adrenoceptor alpha 1B (ADRA1B) that had a statistically significant difference. Then we used Gene Expression Omnibus (GEO) database to establish a diagnostic and prognostic model to verify the three-gene signature. For experimental validation, immunohistochemistry in tissue microarrays showed that thyroid samples' proteins expressed by this three-gene are differentially expressed. Our protocol discovered a robust three-gene signature that can distinguish prognosis, which will have daily clinical application.

Keywords: PTC, SVM diagnostic model, COX analysis, accurate diagnosis, prognostic evaluation

BACKGROUND

Thyroid cancer (TC) is the most common malignant tumour in the endocrine system (The Lancet, 2017), whose most popular type is papillary thyroid carcinoma (PTC), accounting for 80–90% of all thyroid malignancies (Schneider and Chen, 2013; Kennedy and Robinson, 2016). If timely detection, diagnosis, and evaluation can be achieved during the early stages of PTC, coupled with the development of corresponding surgical methods, the patient's follow-up treatment, disease surveillance, and prognosis will significantly improve. Therefore, it is of great importance to study the early screening, diagnosis, and prognosis of PTC.

Currently, the main clinical diagnostics for TC include high-resolution ultrasonography (US) and fine-needle aspiration (FNA), while FNA is the safest and most reliable test that can provide a definitive preoperative diagnosis of malignancy (Zheng et al., 2015; Ko et al., 2017). However, the sensitivity and specificity of FNA are reported to be 68–98% and 56–100%, respectively. This led to an increased rate of uncertain outcomes, underwent unnecessary diagnostic surgery, and received lifelong thyroid hormone replacement therapy with associated surgical complications. Preoperative molecular analysis using a panel of genetic alterations would overcome the limitation of FNA diagnosis (Muzza et al., 2020). Molecular markers have become a potential tool for TC management to distinguish benign from malignant lesions, predict aggressive biology, prognosis, recurrence, and identify novel therapeutic targets (Nylén et al., 2020).

Recently, with the development of genome sequencing technologies, more and more accumulating evidence has revealed that tumour biomarkers, including protein-coding genes, non-coding RNAs and immune genes, are informative for cancer detection and prognosis classification (Zhong et al., 2020; Gan et al., 2021). mRNAs have a great potential in physiological and pathological processes and predict the prognosis of various types of tumour patients (Feng et al., 2019; Tschirdewahn et al., 2019). Therefore, the dysregulated expression or mutation of RNA may be a promising predictor of poor prognosis in PTC. Thus, mRNAs' dysregulated expression or mutation may be a promising predictor of poor prognosis in PTC.

The accurate determination of cancer diagnosis and treatment risk is a significant unmet need in PTC management. Patients and physicians must weigh the benefits of currently available therapies against the potential morbidity of these treatments. Herein we screen for characteristic genes specific to PTC and establish and validate an accurate model for PTC diagnosis and prognostic evaluation.

METHODS

Patients and Tissue Samples

Tissue microarrays (TMA) of human TC (IWLT-N-58T53 TC-1503) involved in this experiment and research were purchased from Wuhan Aiwei Biological Technology Co. LTD., along with the detailed clinical information. It included 29 cases of PTC and 29 cases of para-cancer tissue. Of the 29 patients, 21 were female (aged 24–66 years), and eight were male (aged 27–60 years).

Gene-Expression Data Sets

The gene expression and clinical data used for modelling were derived from TCGA (http://www.cbioportal.org/datasets), which contained gene expression data from 568 samples and clinical information from 516 samples. From The Cancer Genome Atlas (TCGA) database, clinical information was screened via the Cancer Type Detailed PTC parameter, of which a total of 399 samples were found. In these 399 PTC patients, 395 cases had RNA-seq data, of which 52 had para-cancer tissue data creating a total number of 447 RNA-seq data points. The gene expression data used to validate our model came from GSE27155 (Giordano et al., 2005; Giordano et al., 2006) of the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The differentially expressed (DE) mRNAs between normal and PTC samples were assessed using the R Studio. software program (RStudio version 1.1.463; http://www.rproject.org), and the R package, Limma. log2FC (fold change) > 2 and p-value < 0.05 were considered for subsequent analyses (Gong et al., 2019). The project/collection had a total of 99 samples: four from normal patients, 10 cases of follicular adenomas, 13 cases of follicular thyroid carcinomas, 7 cases of eosinophilic thyroid adenomas, 8 cases of thyroid carcinomas, 51 cases of PTC, 4 cases of anaplastic thyroid carcinomas, and 2 cases of medullary thyroid carcinomas. In this study, we selected the 4 cases of normal patients and 51 cases of PTC to analyze.

Feature Selection Methods

Python 3.6 was utilized to screen TCGA expression data for DE genes. The processing steps were as follows: Delete genes with an average expression value of less than 10 reads, which are considered to be genes of no research value in survival differences. Judge whether or not there is a significant difference at $p < 0.05$ between the two comparison groups using the SciPy package (https://www.scipy.org/) to perform t-test on the different study groups. Calculate the fold change value difference between groups by taking the mean value of different groupings.

To find genes for use in modelling, we screened for characteristic genes that significantly affected PTC survival. The R 3.6 software was used to perform a univariate cox regression analysis between DE genes and clinical data (time, status) in 395 patients (Gill, 1992). Genes with a hazard ratio (HR) greater than 1, or less than 1, and a Wald test p-value of less than 0.05 were genes that significantly affected PTC survival. Therefore, selected these genes as characteristic genes for use in establishing a diagnostic model. We summarize the selection process in **Figure 1**.

Establishing the Diagnosis and Prognosis Model

This study used the sklearn package (http://scikit-learn.org) provided in Python 3.6 to establish a Support vector machine (SVM) model to differentiate between cancer and non-cancer. Use the SVM classifier model to explore the optimal three-gene signature prognosis model. Based on the univariate Cox regression analysis of the selected characteristic genes, we established a prognostic model to calculate a patient's prognosis by calculating their 'RiskScore' (Xiong et al., 2017). According to a set threshold (HR > 1 or HR < 1, $p < 0.05$), three-gene (**Table 2**) were found to be significantly associated with overall survival.

To test the diagnostic predictive power of the three-gene signature that we selected, we randomized TCGA PTC patients into a training set (312 samples, 70%) and a test set (135 samples, 30%). The training set was used for 10% cross-validation. The optimal parameters of the final model were ("C": 1, "gamma": 1,000, "kernel": "rbf"), with the final average accuracy being 0.9263 (Standard Deviation: ± 0.0117). The average accuracy of our best model using the training set was 0.9679. To verify the effectiveness of this model, we used the best model predictions that gave an average accuracy of 0.9259. In addition, to verify the diagnostic predictive power of our three-gene signature, we also used the three-gene in GSE27155 and established an SVM model in the same way. Due to the small number of negative samples, we chose to use the 3-fold

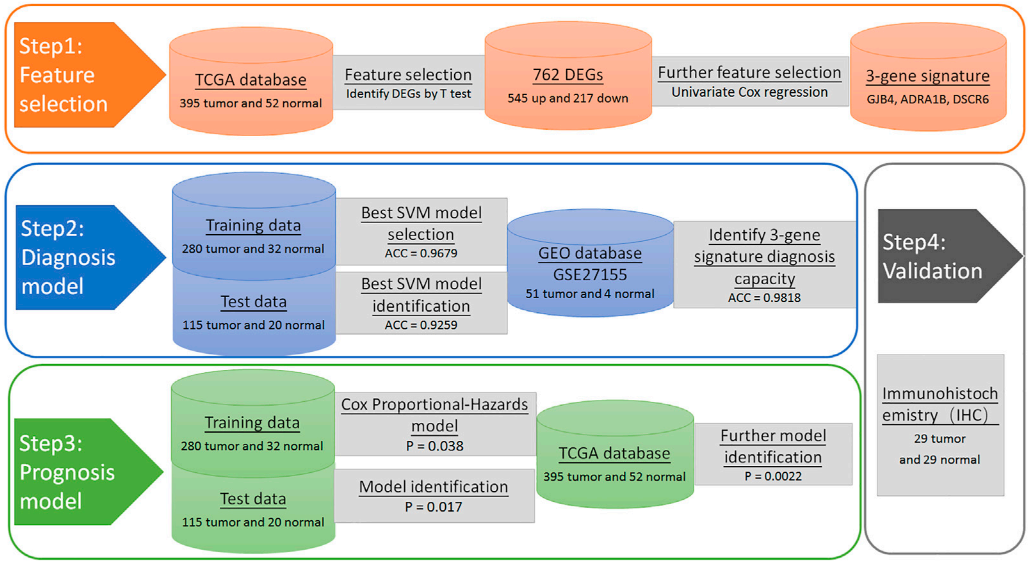


FIGURE 1 | Flowchart of PTC prognostic signatures generation and validation procedures.

cross-validation and pre-determined optimal parameters of the model ("C": 15, "gamma": 1, "kernel": "rbf"). This gave an accuracy of 0.9464 (SD: ± 0.0430).

Microarray Preprocessing

Briefly, after deparaffinization in xylene and rehydration with graded concentrations of alcohol to distilled water, the TMA slides were washed in Tris-buffered saline with 0.1% Tween 20 (TBST), the slides were incubated with the primary antibody against GJB4 (1:10, Abcam, A9888), RIPPLY3 (1:75, Sigma-Aldrich, HPA055541), ADRA1B (1:50, Abcam, ab84405) at 4 ˚C overnight. After washing three times in TBST, the specifically bound secondary antibody was detected with the DAKO EnVision detection System (Dako Diagnostics, Switzerland). Immunostaining scores were independently performed by two experienced pathologists who did not know the patient's clinical pathology data and the immediate clinical outcome. The staining intensity was scored as negative (1), weak (2), moderate (3) or strong (4). The staining extent was scored as 1 (≤10%), 2 (11–50%), 3 (51–75%) or 4 (>75%). A total expression score was calculated by multiplying the staining intensity score with that of the staining extent. ≤ 8 points were considered as low expression. Otherwise, it is considered as a high expression. Histological classification of the samples, stained with hematoxylin and eosin, was performed by two independent clinical pathologists.

Functional Enrichment Gene Ontology Analysis

GO functional annotation pathway enrichments were performed in R using the "clusterProfiler" package, and *P* adjusted (FDR) < 0.05 was statistically significant.

Construction of the Protein-Protein Interaction Network

The DE mRNA were imported into the STRING database (https://string-db.org/) (UniProt Consortium, 2010) to construct a PPI network. The network analysis plug-in in Cytoscape software was used to analyze network topological features to screen the hub nodes in the PPI network (Saito et al., 2012). Degree centrality denotes several direct connections of a node to all other nodes in the network.

Data Analysis

Statistical analysis was performed using R 3.4.0 (https://www.r-project.org/), Python 3.6 (https://www.python.org/) and Graphpad Prism version 7.0 (GraphPad Software). A two-tailed Student's t-test was used for comparisons between two independent groups. This study used the sklearn package (http://scikit-learn.org) provided in Python 3.6 to establish an SVM model to differentiate between cancer and non-cancer. All statistical analyses were two-sided. $p < 0.05$ was defined as indicating statistical significance (Ge et al., 2019).

RESULTS

Bioinformatics Analysis Was Used to Screen Differentially Expressed Genes

A total of 20,531 genes were screened from TCGA. After deleting genes with a mean expression of less than 10, 15,370 genes remained. The number of DE genes between tumour tissue and adjacent tissues was 762, of which 545 genes were upregulated, and 217 genes were down-

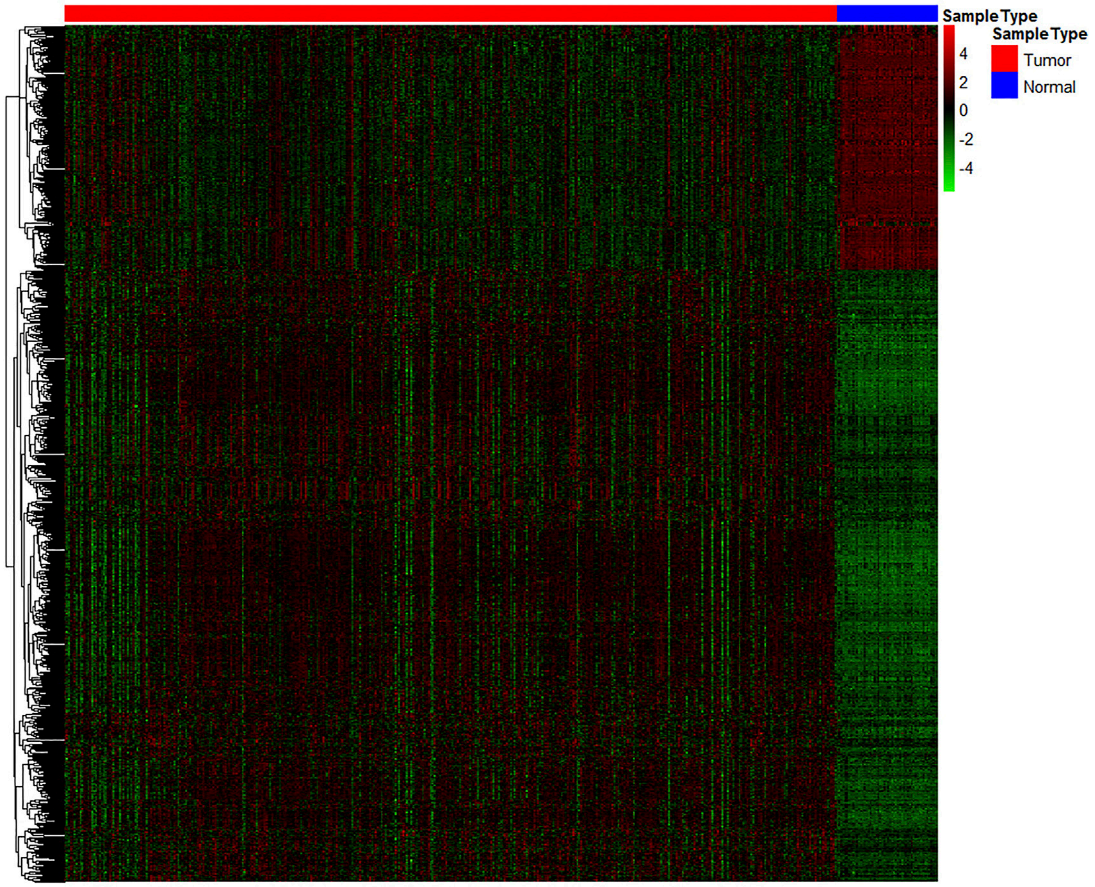


FIGURE 2 | Heatmap of significantly differentially expressed genes. Each row represents a separate gene, each column represents a separate sample, a gradient from green to red indicates a low to high level of expression, and the samples are clustered from two types of tissue: normal tissue (green) and cancer tissue (red).

regulated. The expression values of the DE genes were converted by log10 and displayed by heatmap. As seen from the heatmap (**Figure 2**), there is a significant difference between the tumour and the normal tissue, indicating that the results identified in this study are credible. The top 10 significantly upregulated and the top 10 significantly downregulated DE mRNA are displayed in **Table 1**.

TABLE 1 | The top 10 upregulated and downregulated DE mRNA genes.

Type	Genes	LogFC	p value
Up-regulated	ARHGAP36	8.894666584	<0.001
	DMBX1	8.212911341	<0.001
	SLC18A3	8.071324334	<0.001
	TRY6	7.77283886	<0.001
	TMPRSS6	7.625107474	<0.001
	PRSS1	7.59111566	<0.001
	MMP13	7.567785968	<0.001
	KLK6	7.468232832	<0.001
	LOC400794	7.390954657	<0.001
	GABRB2	7.299817152	<0.001
Down-regulated	KCNA1	−4.139432844	<0.001
	TFF3	−3.811991634	<0.001
	LRP1B	−3.692660909	<0.001
	RELN	−3.629457676	<0.001
	IPCEF1	−3.521246594	<0.001
	ZNF804B	−3.519727733	<0.001
	CNTN5	−3.507769597	<0.001
	AGR3	−3.492012695	<0.001
	VIT	−3.43067668	<0.001
	FAM180B	−3.414101394	<0.001

DE,diferentially expressed;FC,fold change.

Cox Regression Analysis Was Used to Screen Characteristic Genes

Through univariate Cox regression analysis, a hazard ratio was calculated for each gene according to the set threshold (HR > 1 or HR < 1, $p < 0.05$), To screen-specific biomarkers with accurate diagnostic ability. There were three genes found to be significantly related to overall survival (**Table 2**). These genes were GJB4, RIPPLY3, and ADRA1B. These three genes will be used as feature genes for subsequent modelling, and the specific information of genes is shown in **Table 4**.

TABLE 2 | Univariate Cox regression analysis results.

Gene symbol	Beta	HR (95% CI)	*p*. value
GJB4	−0.057	0.94 (0.91–0.98)	0.0066
ADRA1B	−0.021	0.98 (0.96–0.99)	0.0067
RIPPLY3	−0.11	0.9 (0.81–0.99)	0.0360

Establishing and Validate the Cox Prognostic Model

To test the prognostic prediction ability of the screened three-gene signature, we calculated the Risk-Score of each patient in the TCGA training set by using the established prognostic model. The risk assessment score formula was as follows: risk score= (-0.057×expression value of GJB4) + (−0.021×expression value of ADRA1B) + (-0.110×expression value of RIPPLY3). **Table 3** Then the patients were ranked according to risk-score, and the risk-score median (−0.7766) was taken as the threshold. The 312 patients were divided into two groups, 156 patients with low risk and 156 patients with high risk. Of these 312 persons, 280 had survival information (OS), of which 156 were low risk, and 124 were high risk. Kaplan-Meier survival curves and ROC curves were used to examine the predictive power of three-gene biomarkers. Kaplan-Meier survival curves showed that the survival rate in the high-risk group was significantly lower than that in the low-risk group (log-rank *p* value = 0.038), **Figure 3**. The ROC curve was shown in **Figure 4**, and the AUC value was 0.7513, indicating that the three-gene signatures had a certain prognostic predictive ability. To further verify the prognostic capability of the three-gene signatures, we calculated the risk score of the patients in the test set and the PTC patient data set of the whole TCGA respectively and divided the patients into high-risk and low-risk patients' groups according to the same threshold (−0.7766), **Table 4**. The Kaplan-Meier survival curves for both data sets showed significantly lower survival rates in the high-risk group than in the low-risk group (log-rank *p* value = 0.017 and 0.0022), with AUC values of 0.9023 and 0.7910, respectively. The training set and the results of two confirmations showed that the three-gene biomarkers screened had a strong prognostic capability.

Immunohistochemical Verification Results of Characteristic Gene Tissue Microarray

To further verify the protein expression level of three-gene signatures in PTC tissue and analyze its relationship with clinicopathological features, immunohistochemistry was used to detect the protein expression level of tri-factor in 29 PTC tissue chips. The results of this study showed that the protein

TABLE 3 | Differential expression information of characteristic genes.

Gene symbol	mRNA description	logFC	P	UP DOWN
GJB4	gap junction protein beta 4	4.0572	<0.001	UP
ADRA1B	adrenoceptor alpha 1B	2.4496	<0.001	UP
RIPPLY3	ripply transcriptional repressor 3	2.1379	<0.001	UP

FC,fold change.

TABLE 4 | Survival analysis sample.

Data set	High risk	Low risk	High risk for OS	Low risk for OS	*p* value
Training set	156	156	156	124	0.038
Test set	72	63	52	63	0.017
All data	228	219	176	219	0.002

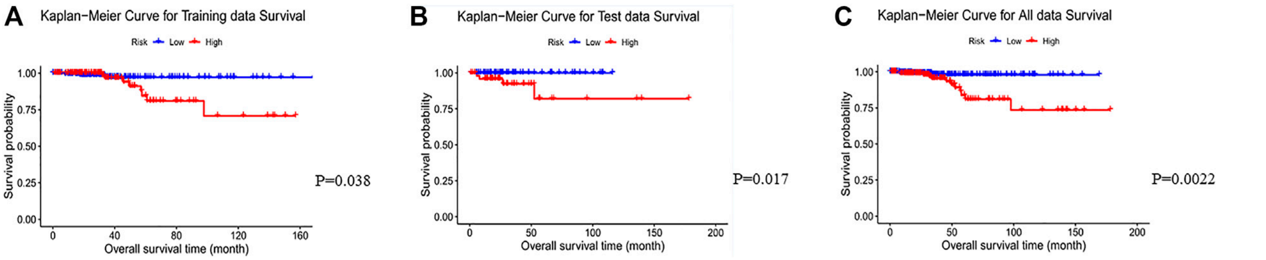


FIGURE 3 | Kaplan-Meier curves for the low- and high-risk groups separated by the Risk-Score of the 3-gene signature in the TCGA PTC data. The blue line represents the patients with low risk and the others represent patients with high risk. Significant differences in overall survival between the two groups were analyzed by log-rank test. **(A**) Kaplan-Meier curves for training data survival; **(B)**Kaplan-Meier curves for test data; **(C)** Kaplan-Meier curves for all data.

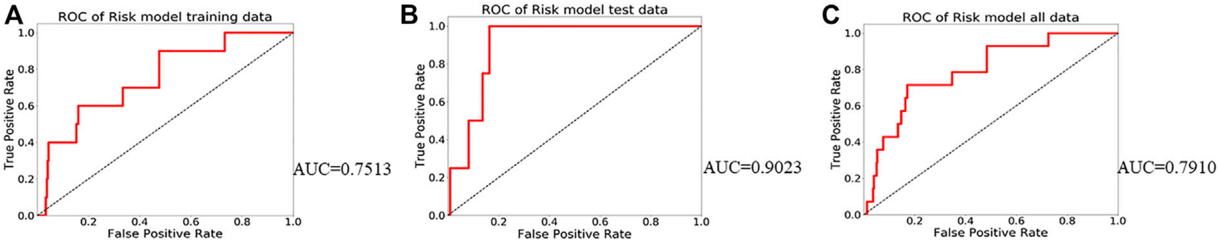


FIGURE 4 | Receiver operating characteristic curves (ROC) for the prognosis models. **(A)** ROC fited based on training data; **(B)** ROC fited based on test data; **(C)** ROC fited based on all TCGA PTC data.

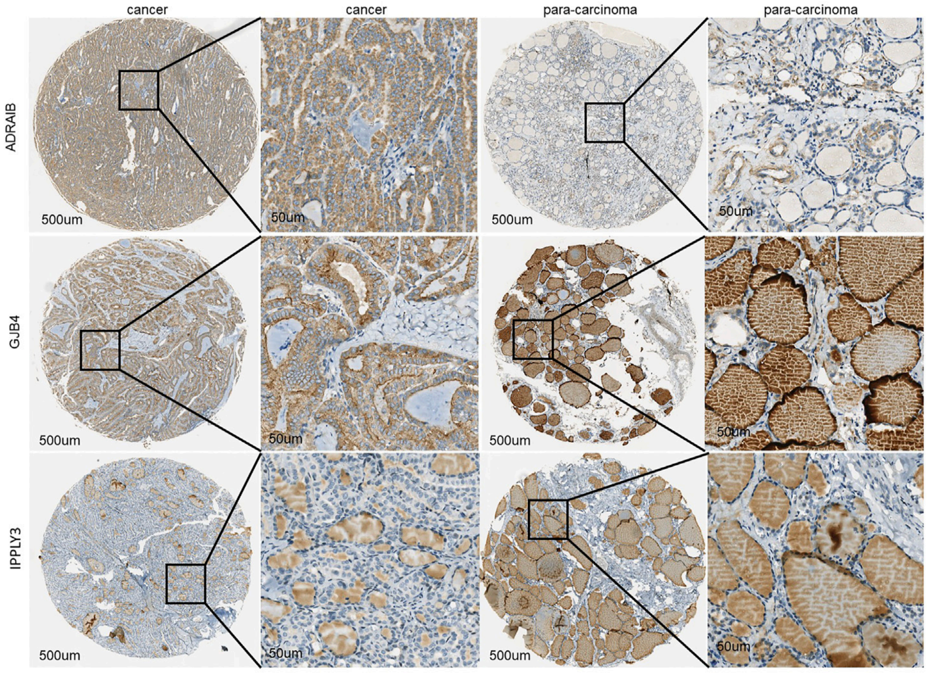


FIGURE 5 | The immunohistochemical results of GJB4, RIPPLY3, and ADRA1B characteristic gene tissue chips indicated that the expression in tumor tissues was significantly higher than that in adjacent tissues.

expression levels of GJB4 and ADRA1B in cancer cells were significantly higher than those in paired para-cancer cells ($p < 0.05$) (**Figure 5**), and the protein expression level of RIPPLY3 was not significant difference between the cancer cells and the para-cancer cells.

GO Enrichment Analyses and Construction of the PPI Network

Functional enrichment GO analyses were performed to investigate the underlying mechanisms of the DE mRNA genes' prognostic effects. Our results demonstrate that the DE mRNA genes are linked with activating pathways, such as tumour development and progression regulation, based on GO analysis of three cohorts (**Figure 6**). PPI network reveal that the potential connection between key mRNA genes (**Figure 7**).

DISCUSSION

In our study, the results above clearly demonstrate that the three-gene signatures we screened for and selected have a strong ability to distinguish between cancerous and non-cancerous samples. The genes signature include GJB4, RIPPLY3, and ADRA1B. It was reported that GJB4 and ADRA1B genes play an essential role in developing many malignant tumours. It has been shown that GJB4 is involved in tumorigenesis and may act as a tumour promoter, Wang et al. (Wang et al., 2019) indicated that miR-492 promoted cancer progression by targeting GJB4 and was a novel

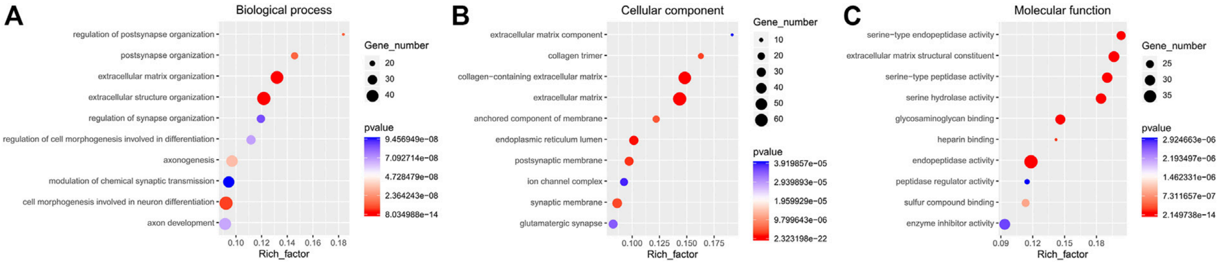


FIGURE 6 | The GO enrichment analyses. **(A)**. Biological process; **(B)**. Cellular component; **(C)**. Molecular function.

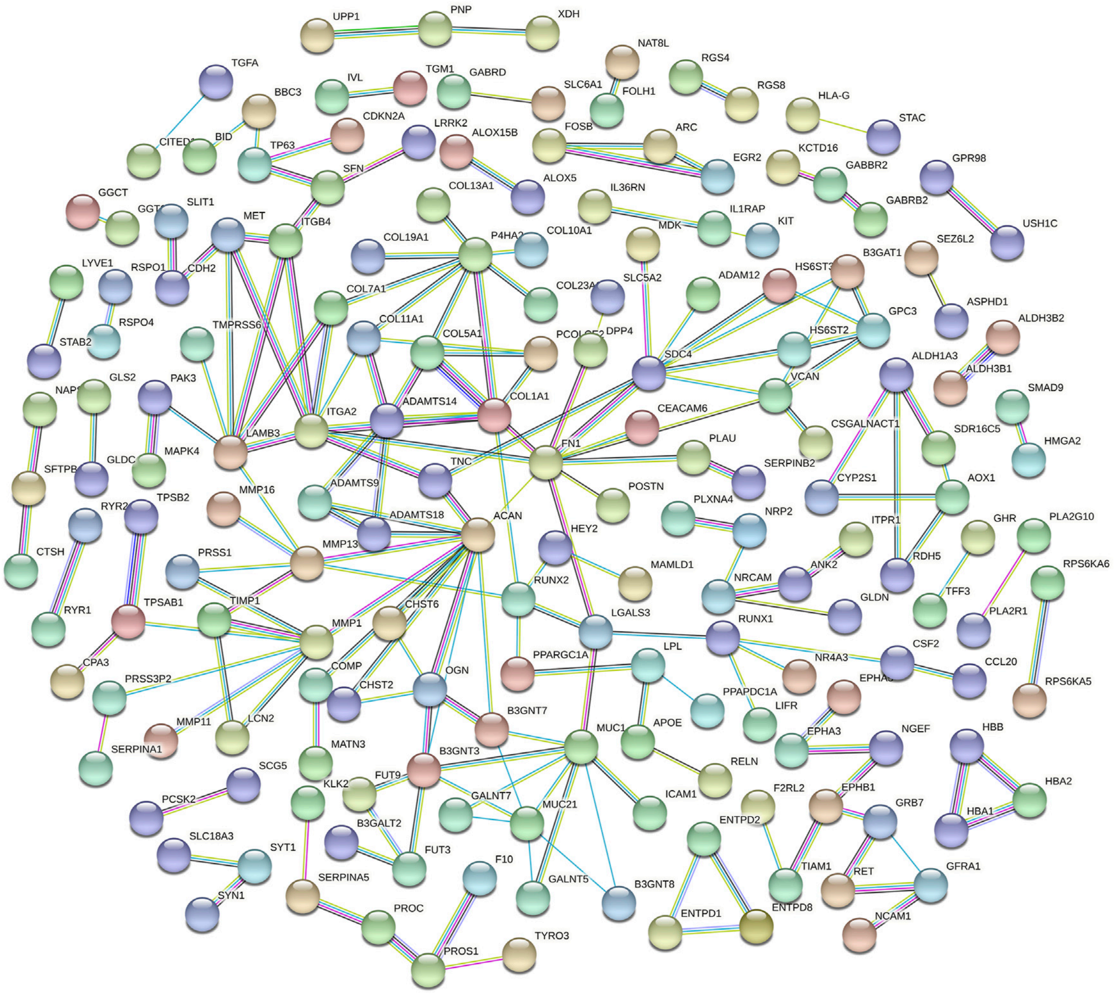


FIGURE 7 | The PPI networks of DE mRNA.

biomarker for bladder cancer. Liu et al. (Liu et al., 2019) showed that GJB4 was highly expressed in gastric cancer tissues and cells, the high expression of GJB4 was significantly correlated with the overall survival of gastric cancer patients, and the cell proliferation and migration of gastric cancer cells were significantly inhibited by knockout GJB4. At the same time, targeting GJB4 may be exploited as a modality for improving lung cancer therapy had been proved (Lin et al., 2019). The ADRA1B gene is a member of the adrenergic receptor alpha 1 (ADRA1) subfamily, which also includes ADRA1A and ADRA1D, and has been shown to promote the development of cancer in the epinephrine cell pathway. Adrenergic receptor antagonists have also been shown to be useful in the treatment of various types of cancer, including prostate and breast cancer (Freudenberger et al., 2006; Harris et al., 2007). In present, no studies have been reported on the relationship between GJB4 and ADRA1B gene in TC. However, our study demonstrated that the GJB4 and ADRA1B genes may not be able to promote the thyroid cancer progression and development, and it may even be a protective gene.

Although no direct studies have proved that RIPPLY3 (also known as DSCR6) is closely related to PTC or other malignant tumours, current studies have found that the RIPPLY3 gene plays a role in developing the pharynx and its derivatives in vertebrates (Tsuchiya et al., 2018). Li et al. (Li et al., 2013) showed that RIPPLY3 is closely associated with Down syndrome (DS). Studies have found that people with DS have an increased risk of thyroid disease (mainly autoimmune), with a lifetime prevalence of between 13 and 63% (AlAaraj et al., 2019). These results suggest that RIPPLY3 may affect the development of the thyroid gland, and its abnormal expression may lead to the occurrence and development of PTC. The GO enrichment analysis revealed that the chief pathways regulated the cell-molecular function and the enzyme activity. Previous studies have demonstrated the gene effect on thyroid cell function and cell morphology (Yu et al., 2017; Rudzińska et al., 2019).

At present, the preoperative diagnosis of PTC is still mainly FNA. However, according to the Bethesda grading standard, the proportion of FNA diagnosis results is suspicious or uncertain is 3–18% (Misiakos et al., 2016). In the era of precision therapy, we need an accurate diagnosis, which requires an accurate prognosis. To find prognostic molecular markers of PTC, this study obtained the gene expression characteristic of tumor prognosis through TCGA to screen characteristic genes and carry out an effective risk assessment of tumour prognosis.

However, our study has some limitations, the most important one is the limited number of patients in our database group confined to the limitation to TCGA、GEO. There is still a lack of large sample data sets and clinical samples to verify the accuracy of the three-signature prognosis model. Also, we should further investigate the correlation between the three gene expression levels and the clinicopathological features. Fortunately, gene sequencing technology is gradually maturing and becoming faster and less expensive. We will continue to collect cases of TC tissue to verify our signature further. Furthermore, although we believe that the three-gene signature is promising in selecting patients who will benefit from the three-gene prognostic model, its significant value still needs to be verified in prospective studies.

CONCLUSION

This study screened for DE genes (GJB4, RIPPLY3, ADRA1B) that were significantly related to the diagnosis and prognosis of PTC. The three-gene diagnostic model could accurately predict the occurrence of PTC and guide prognosis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

Conception: L-kZ, BX, and J-HM. Design and revise the manuscript: L-kZ, X-YD, FS, W-SC, J-HF, X-XG, SJ, C-ZL, M-GZ, J-WD, B-XZ, X-ZX, L-QN, HH and S-SC. Analysis and interpretation of data: L-kZ, X-YD, FS, W-SC, J-HF, and X-XG.

FUNDING

This research was supported by the Zhongshan Medical Research Project (2021A020606), The Major Research Plan and Young Top-notch Talents of Zhongshan City People’s Hospital (2021) and Guangzhou Medicine and Health Care Technology Projects (20211A011010).

ACKNOWLEDGMENTS

Special thanks to Tian-Xiong Shi for support and encouragement.

REFERENCES

AlAaraj, N., Soliman, A. T., Itani, M., Khalil, A., and De Sanctis, V. (2019). Prevalence of Thyroid Dysfunctions in Infants and Children with Down Syndrome (DS) and the Effect of Thyroxine Treatment on Linear Growth and Weight Gain in Treated Subjects versus DS Subjects with normal Thyroid Function: a Controlled Study. *Acta Biomed.* 90 (8-S), 36–42. doi:10.23750/abm.v90i8-S.8503

Feng, G., Ma, H.-M., Huang, H.-B., Li, Y.-W., Zhang, P., Huang, J.-J., et al. (2019). Overexpression of COL5A1 Promotes Tumor Progression and Metastasis and Correlates with Poor Survival of Patients with clear Cell Renal Cell Carcinoma. *Cmar* Vol. 11, 1263–1274. doi:10.2147/cmar.s188216

Freudenberger, R. S., Kim, J., Tawfik, I., and Sonnenberg, F. A. (2006). Optimal Medical Therapy Is superior to Transplantation for the Treatment of Class I, II, and III Heart Failure: a Decision Analytic Approach. *Circulation* 114 (1 Suppl. l), I62–I66. doi:10.1161/CIRCULATIONAHA.105.001412

Gan, X., Guo, M., Chen, Z., Li, Y., Shen, F., Feng, J., et al. (2021). Development and Validation of a Three-Immune-Related Gene Signature Prognostic Risk Model in Papillary Thyroid Carcinoma. *J. Endocrinol. Invest.* 44 (10), 2153–2163. doi:10.1007/s40618-021-01514-7

Ge, W., Cai, W., Bai, R., Hu, W., Wu, D., Zheng, S., et al. (2019). A Novel 4-gene Prognostic Signature for Hypermutated Colorectal Cancer. *Cmar* Vol. 11, 1985–1996. doi:10.2147/cmar.s190963

Gill, R. D. (1992). Multistate Life-tables and Regression Models. *Math. Popul. Stud.* 3 (4), 259–276. doi:10.1080/08898489209525345

Giordano, T. J., Au, A. Y., Kuick, R., Thomas, D. G., Rhodes, D. R., Wilhelm, K. G., Jr., et al. (2006). Delineation, Functional Validation, and Bioinformatic Evaluation of Gene Expression in Thyroid Follicular Carcinomas with the PAX8-PPARG

Translocation. *Clin. Cancer Res.* 12 (7 Pt 1), 1983–1993. doi:10.1158/1078-0432.CCR-05-2039

Giordano, T. J., Kuick, R., Thomas, D. G., Misek, D. E., Vinco, M., Sanders, D., et al. (2005). Molecular Classification of Papillary Thyroid Carcinoma: Distinct BRAF, RAS, and RET/PTC Mutation-specific Gene Expression Profiles Discovered by DNA Microarray Analysis. *Oncogene* 24 (44), 6646–6656. doi:10.1038/sj.onc.1208822

Gong, Y., Zou, B., Chen, J., Ding, L., Li, P., Chen, J., et al. (2019). Potential Five-MicroRNA Signature Model for the Prediction of Prognosis in Patients with Wilms Tumor. *Med. Sci. Monit.* 25, 5435–5444. doi:10.12659/msm.916230

Harris, A. M., Warner, B. W., Wilson, J. M., Becker, A., Rowland, R. G., Conner, W., et al. (2007). Effect of α1-Adrenoceptor Antagonist Exposure on Prostate Cancer Incidence: An Observational Cohort Study. *J. Urol.* 178 (5), 2176–2180. doi:10.1016/j.juro.2007.06.043

Kennedy, J. M., and Robinson, R. A. (2016). Thyroid Frozen Sections in Patients with Preoperative FNAs. *Am. J. Clin. Pathol.* 145 (5), 660–665. doi:10.1093/ajcp/aqw042

Ko, Y. S., Hwang, T. S., Kim, J. Y., Choi, Y. L., Lee, S. E., Han, H. S., et al. (2017). Diagnostic Limitation of Fine-Needle Aspiration (FNA) on Indeterminate Thyroid Nodules Can Be Partially Overcome by Preoperative Molecular Analysis: Assessment of RET/PTC1 Rearrangement in BRAF and RAS Wild-type Routine Air-Dried FNA Specimens. *Int. J. Mol. Sci.* 18 (4). doi:10.3390/ijms18040806

Li, H.-Y., Grifone, R., Saquet, A., Carron, C., and Shi, D.-L. (2013). The Xenopus Homologue of Down Syndrome Critical Region Protein 6 Drives Dorsoanterior Gene Expression and Embryonic axis Formation by Antagonising Polycomb Group Proteins. *Development* 140 (24), 4903–4913. doi:10.1242/dev.098319

Lin, Y.-P., Wu, J.-I., Tseng, C.-W., Chen, H.-J., and Wang, L.-H. (2019). Gjb4 Serves as a Novel Biomarker for Lung Cancer and Promotes Metastasis and Chemoresistance via Src Activation. *Oncogene* 38 (6), 822–837. doi:10.1038/s41388-018-0471-1

Liu, G., Pang, Y., Zhang, Y., Fu, H., Xiong, W., and Zhang, Y. (2019). GJB4 Promotes Gastric Cancer Cell Proliferation and Migration via Wnt/CTNNB1 Pathway. *Ott* Vol. 12, 6745–6755. doi:10.2147/ott.s205601

Misiakos, E. P., Margari, N., Meristoudis, C., Machairas, N., Schizas, D., Petropoulos, K., et al. (2016). Cytopathologic Diagnosis of fine Needle Aspiration Biopsies of Thyroid Nodules. *Wjcc* 4 (2), 38–48. doi:10.12998/wjcc.v4.i2.38

Muzza, M., Colombo, C., Pogliaghi, G., Karapanou, O., and Fugazzola, L. (2020). Molecular Markers for the Classification of Cytologically Indeterminate Thyroid Nodules. *J. Endocrinol. Invest.* 43 (6), 703–716. doi:10.1007/s40618-019-01164-w

Nylén, C., Mechera, R., Maréchal-Ross, I., Tsang, V., Chou, A., Gill, A. J., et al. (2020). Molecular Markers Guiding Thyroid Cancer Management. *Cancers (Basel)* 12 (8). doi:10.3390/cancers12082164

Rudzińska, M., Grzanka, M., Stachurska, A., Mikula, M., Paczkowska, K., Stępień, T., et al. (2019). Molecular Signature of Prospero Homeobox 1 (PROX1) in Follicular Thyroid Carcinoma Cells. *Int. J. Mol. Sci.* 20 (9). doi:10.3390/ijms20092212

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., et al. (2012). A Travel Guide to Cytoscape Plugins. *Nat. Methods* 9 (11), 1069–1076. doi:10.1038/nmeth.2212

Schneider, D. F., and Chen, H. (2013). New Developments in the Diagnosis and Treatment of Thyroid Cancer. *CA Cancer J. Clin.* 63 (6), 374–394. doi:10.3322/caac.21195

The Lancet, L. (2017). Thyroid Cancer Screening. *The Lancet* 389 (10083), 1954. doi:10.1016/s0140-6736(17)31349-1

Tschirdewahn, S., Panic, A., Püllen, L., Harke, N. N., Hadaschik, B., Riesz, P., et al. (2019). Circulating and Tissue IMP3 Levels Are Correlated with Poor Survival in Renal Cell Carcinoma. *Int. J. Cancer* 145 (2), 531–539. doi:10.1002/ijc.32124

Tsuchiya, Y., Mii, Y., Okada, K., Furuse, M., Okubo, T., and Takada, S. (2018). Ripply3 Is Required for the Maintenance of Epithelial Sheets in the Morphogenesis of Pharyngeal Pouches. *Develop. Growth Differ.* 60 (2), 87–96. doi:10.1111/dgd.12425

UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38 (Suppl. l_1), D142–D148. doi:10.1093/nar/gkp846

Wang, K., Lü, H., Qu, H., Xie, Q., Sun, T., Gan, O., et al. (2019). miR-492 Promotes Cancer Progression by Targeting GJB4 and Is a Novel Biomarker for Bladder Cancer. *Ott* Vol. 12, 11453–11464. doi:10.2147/ott.s223448

Xiong, Y., Wang, R., Peng, L., You, W., Wei, J., Zhang, S., et al. (2017). An Integrated lncRNA, microRNA and mRNA Signature to Improve Prognosis Prediction of Colorectal Cancer. *Oncotarget* 8 (49), 85463–85478. doi:10.18632/oncotarget.20013

Yu, Y., Liu, C., Zhang, J., Zhang, M., Wen, W., Ruan, X., et al. (2017). Rtfc (4931414P19Rik) Regulates *In Vitro* Thyroid Differentiation and *In Vivo* Thyroid Function. *Sci. Rep.* 7, 43396. doi:10.1038/srep43396

Zheng, B., Liu, J., Gu, J., Lu, Y., Zhang, W., Li, M., et al. (2015). A Three-Gene Panel that Distinguishes Benign from Malignant Thyroid Nodules. *Int. J. Cancer* 136 (7), 1646–1654. doi:10.1002/ijc.29172

Zhong, L. K., Gan, X. X., Deng, X. Y., Shen, F., Feng, J. H., Cai, W. S., et al. (2020). Potential five-mRNA S-ignature M-odel for the P-rediction of P-rognosis in P-atients with P-apillary T-hyroid C-arcinoma. *Oncol. Lett.* 20 (3), 2302–2310. doi:10.3892/ol.2020.11781

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 XS declared a shared parent affiliation with the author(s) to the handling editor at the time of the review.

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

frontiers
in Molecular Biosciences

REVIEW
published: 31 March 2022
doi: 10.3389/fmolb.2022.847812

Network Pharmacology and Experimental Validation to Reveal the Pharmacological Mechanisms of Chongcaoyishen Decoction Against Chronic Kidney Disease

*Zhenliang Fan[1†], Jingjing Chen[2†], Qiaorui Yang[3] and Jiabei He[4]**

[1]Nephrology Department, The First Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China, [2]Department of Rheumatology and Immunology, The First Hospital Affiliated to Army Medical University, Chongqing, China, [3]Graduate School, Heilongjiang University of Chinese Medicine, Harbin, China, [4]Department of Oncology Radiotherapy, Affiliated Zhongshan Hospital to Dalian University, Liaoning, China

OPEN ACCESS

Edited by:
Tianyi Zhao,
Harbin Institute of Technology, China

Reviewed by:
Hongzhe Guo,
Harbin Institute of Technology, China
Eunhee Choi,
Columbia University Irving Medical Center, United States

****Correspondence:***
Jiabei He
hejiabei1128@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics,
a section of the journal Frontiers in Molecular Biosciences

Received: *03 January 2022*
Accepted: *14 February 2022*
Published: *31 March 2022*

Citation:
Fan Z, Chen J, Yang Q and He J (2022) Network Pharmacology and Experimental Validation to Reveal the Pharmacological Mechanisms of Chongcaoyishen Decoction Against Chronic Kidney Disease.
Front. Mol. Biosci. 9:847812.
doi: 10.3389/fmolb.2022.847812

Objective: To explore the pharmacological mechanisms of Chongcaoyishen decoction (CCYSD) against chronic kidney disease (CKD) *via* network pharmacology analysis combined with experimental validation.

Methods: The bioactive components and potential regulatory targets of CCYSD were extracted from the TCMSP database, and the putative CKD-related target proteins were collected from the GeneCards and OMIM database. We matched the active ingredients with gene targets and conducted regulatory networks through Perl5 and R 3.6.1. The network visualization analysis was performed by Cytoscape 3.7.1, which contains ClueGO plug-in for GO and KEGG analysis. *In vivo* experiments were performed on 40 male SD rats, which were randomly divided into the control group (n = 10), sham group (n = 10), UUO group (n = 10), and CCYSD group (n = 10). A tubulointerstitial fibrosis model was constructed by unilateral ureteral obstruction through surgery and treated for seven consecutive days with CCYSD (0.00657 g/g/d). At the end of treatment, the rats were euthanized and the serum and kidney were collected for further detection.

Results: In total, 53 chemical compounds from CCYSD were identified and 12,348 CKD-related targets were collected from the OMIM and GeneCards. A total of 130 shared targets of CCYSD and CKD were acquired by Venn diagram analysis. Functional enrichment analysis suggested that CCYSD might exert its pharmacological effects in multiple biological processes, including oxidative stress, apoptosis, inflammatory response, autophagy, and fiber synthesis, and the potential targets might be associated with JAK-STAT and PI3K-AKT, as well as other signaling pathways. The results of the experiments revealed that the oxidative stress in the UUO group was significantly higher than that in normal state and was accompanied by severe tubulointerstitial fibrosis (TIF), which could be effectively reversed by CCYSD ($p < 0.05$). Meanwhile, aggravated mitochondrial injury and autophagy was observed in the epithelial cells of the renal tubule in the UUO group, compared to the normal ones

($p < 0.05$), while the intervention of CCYSD could further activate the autophagy and reduce the mitochondrial injury ($p < 0.05$).

Conclusion: We provide an integrative network pharmacology approach combined with *in vivo* experiments to explore the underlying mechanisms governing the CCYSD treatment of CKD, which indicates that the relationship between CCYSD and CKD is related to its activation of autophagy, promotion of mitochondrial degradation, and reduction of tissue oxidative stress injury, promoting the explanation and understanding of the biological mechanism of CCYSD in the treatment of CKD.

Keywords: Chongcaoyishen decoction, chronic kidney disease, oxidative stress injury, autophagy, mitochondrial injury, tubulointerstitial fibrosis

INTRODUCTION

Chronic kidney disease (CKD) mainly refers to the irreversible structural and/or functional impairment of the kidney resulted from multiple causes for more than 3 months. With the change of people's lifestyle, CKD has become a chronic disease seriously affecting human health along with chronic diseases, such as diabetes and hypertension (Webster et al., 2017). The increasing prevalence of CKD has been accompanied by the increasing healthcare costs. According to incomplete statistics, the annual cost of CKD patients in the world is up to five billion dollars, and nearly one million CKD patients die because they cannot afford the high cost of treatment (Couser et al., 2011; Djudjaj and Boor, 2019).

However, the treatment of chronic kidney disease (CKD) is not a breakthrough which deserves celebrating at present. The mainstay of therapy for CKD contains angiotensin-converting enzyme inhibitor and angiotensin II inhibitor that can alleviate glomerular "three highs" status (Aggarwal and Singh, 2020) by means of controlling the blood pressure and blood glucose, correcting metabolic acidosis with sodium bicarbonate, restricting protein intake, and taking alpha-keto acid. However, these conventional treatment measures are mainly symptomatic treatment and the suboptimal therapy delaying renal function decline, and treating related complications does not always work effectively because it is hard to grasp the key of pathological changes and pathogenesis of CKD (Collins et al., 2012; National Kidney Foundatio, 2012; Humphreys, 2018). Therefore, it has been a hot issue in relevant research fields to find out drugs that can effectively delay the deterioration of renal function and improve the prognosis by targeting the core pathological changes in the progression of CKD (Gewin, 2018).

With the continuous promotion of traditional Chinese medicine (TCM) in the clinical treatment of CKD, its efficacy in delaying the deterioration of renal function and improving the prognosis of patients has gradually been approved by clinicians and patients. For this reason, many emerging studies in recent years have gradually revealed the mechanism of action and intervention targets of TCM in the treatment of CKD. According to the current research results, TCM compound has the advantage of multiple components and multiple targets in the treatment of CKD, which is incomparable to Western drugs with single chemical components, and can simultaneously intervene multiple targets closely related to the progress of CKD.

Chongcaoyishen decoction (CCYSD) has been applied in clinical practice for many years. Based on the pathological characteristics of patients with CKD, it emphasizes that the treatment should focus on "supplementing and removing the deficiency and combining reinforcement with elimination." Previous studies have fully confirmed that CCYSD can effectively delay renal deterioration in patients with CKD, relieve clinical symptoms, and improve the prognosis of patients (Mo et al., 2019b; Ziyang, 2019; Ma et al., 2020). Although the researchers have carried out many basic studies before, we still cannot figure out a definite explanation to the specific target and exact mechanism of CCYSD in the treatment of chronic kidney disease. In this study, we used network pharmacology to explore the bioactive ingredients in CCYSD and its mechanism of action in treating CKD. Subsequently, experimental verification was carried out on the results of the network pharmacology study to further explore the specific mechanisms of CCYSD in treating CKD.

DATA AND METHODS

Network Pharmacology Analysis

Data Sources

In this study, the bioactive ingredients of Chongcaoyishen decoction and the possible intervention targets of CCYSD were screened from the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP). The targets related to chronic kidney disease were extracted from GeneCards and OMIM databases, and ClueGO plug-in from Cytoscape 3.7.1 was used for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis in order to analyze the biological functions and signaling pathways involved in the drug regulatory network.

Screening of Active Components and Targets of Chongcaoyishen Decoction

The bioactive ingredients of *Cordyceps sinensis*, *Astragalus membranaceus*, leeches, rhubarb in wine, cardamom, and *Sergium sergii* were searched in the TCMSP database, and the

TABLE 1 | Serum oxidative stress levels of rats in each group ($\bar{X}$ ± S).

	Malondialdehyde (nmol/mL)	Superoxide dismutase (U/mL)	Reduced glutathione (μg/mL)
Control group	7.90 ± 0.80	263.89 ± 35.01	645.42 ± 84.98
Sham group	7.38 ± 0.49	244.25 ± 24.51	655.68 ± 82.92
UUO group	8.68 ± 0.42*	193.80 ± 32.52*	406.60 ± 55.72*
CCYSD group	7.39 ± 0.45#	265.18 ± 28.85#	589.18 ± 65.45#

Compared with the sham group, p < 0.05; #compared with the UUO group, p < 0.05.

bioactive ingredients and their corresponding targets were reserved with oral bioavailability (OB) ≥30% and drug-like property (DL) ≥0.18 as screening conditions.

CKD Target Screening

Taking "chronic kidney diseases" as the keyword, we surveyed the related targets of CKD from GeneCards (https://www.genecards.org/) and retrieved 12,183 genes associated with CKD. A total of 202 related gene targets were retrieved from Online Mendelian Inheritance in Man (OMIM, https://omim.org/). After removing the duplicates from GeneCards and OMIM databases, we obtained 12,348 non-repeating gene targets.

Construction of a Drug Regulatory Network and Functional Enrichment Analysis

Based on the collected data, we constructed the drug regulatory network using Perl5 and R 3.6.1 to showcase the correlation between the common targets of active ingredients in CCYSD and potential targets for CKD.

The common targets of active ingredients in CCYSD and potential targets for CKD were collected and input into Cytoscape 3.7.1 to construct a CKD–CCYSD–ingredients–target interaction network, which showcases the abundance of bioactive constituents from CCYSD exerts therapeutic effects on CKD through multiple gene targets.

Meanwhile, the plug-in from Cytoscape 3.7.1, ClueGO, was used to perform visual analysis of KEGG and GO functional enrichment. The KEGG and GO pathway analyses were screened for kappa >0.74 and $p < 0.01$. The top 30 items of GO analysis and the top 30 items of KEGG analysis were mapped as bar plots and bubble plots, and the positions of relevant nodes were adjusted, aiming to obtain a clearer network graph.

Experimental Research

Animals and Experimental Groups

A total of 40 SPF Sprague Dawley (SD) rats (180–220 g) were supplied by the Experimental Animal Center of Heilongjiang University of Chinese Medicine [SCXK (Black) 2017-014]. All these rats were housed in an environmentally controlled room (22 ± 2°C, humidity 60 ± 10%, 12 h/12 h light/dark cycle) with free access to water and lab chow.

According to the random number table, all the animals were randomly and equally divided into four groups: the control group, sham group, UUO group, and CCYSD group ($n = 10$ in each group).

Main Reagents

ELISA 96-well kit: superoxide dismutase (SOD, Solebo BC0170), reduced glutathione (GSH, Solebo BC1170), and malondialdehyde (MDA, Solebo BC0020). Western blot primary antibody: GAPDH (Cymofei MA5-15738), α-SMA (Cymofei MA1-06110), COL-III (CSI007-01-02), and LC3B (PA1-46286).

TIF Modeling and Drug Administration

Unilateral ureteral obstruction (UUO) was established by the surgical method. In brief, rats were anesthetized with pentobarbital, and then, the left renal ureter was separated from the abdomen. After ureter ligation and severance, the muscles and skin were sutured layer by layer to allow the rats to recover. The rats in the UUO group and CCYSD group received the UUO model, while the rats in the sham group received a similar surgical approach without ligation or severance, and the rats in the control group did not undergo any treatment.

The day after the surgery, the rats in the CCYSD group were administered intragastrically with CCYSD at a dose of 0.00657 g/g/d once a day for seven consecutive days, which is equal to the decoction with a dose of 0.657 g/ml (Jie, 2015; Fan et al., 2019; Mo et al., 2019a). The rats in the sham group and UUO group were given the same dosage of normal saline (2 ml). The control group received nothing. Seven days after the surgery, all rats were euthanized, and the blood and tissues were collected for further analysis.

Sample Collection

2 h after the last administration, 5 ml of inferior venous blood was collected from anesthetized rats and centrifuged at 3,000 r/min for 15 min at 4°C to separate the serum. The left kidney was separated into three parts and stored in a 2.5% glutaraldehyde electron microscopy fixative solution, 10% neutral formalin fixative solution, and liquid nitrogen, respectively.

Enzyme Linked Immunosorbent Assay

The serum oxidative stress-related markers were detected by ELISA: the spectrophotometer was preheated for 30 min and zeroed with distilled water. Superoxide dismutase (SOD, Solebo BC0170), reduced glutathione (GSH, Solebo BC1170), and malondialdehyde (MDA, Solebo BC0020) were tested according to the corresponding instructions.

Histological Analysis

Pathological observation: the renal tissues were fixed by formalin, embedded by paraffin, and sectioned at 3 μm. H&E staining was performed with hematoxylin and eosin staining after gradient elution. Then, the specimen was

TABLE 2 | Biological active ingredients in CCYSD.

Mol ID	Molecule Name	OB (%)	Caco-2[a]	DL
MOL000096	(-)-Catechin	49.68	–0.03	0.24
MOL000228	(2R)-7-hydroxy-5-methoxy-2-phenylchroman-4-one	55.23	0.87	0.2
MOL000438	(3R)-3-(2-hydroxy-3,4-dimethoxyphenyl) chroman-7-ol	67.67	0.96	0.26
MOL000033	(3S,8S,9S,10R,13R,14S,17R)-10,13-dimethyl-17-[(2R,5S)-5-propan-2-yloctan-2-yl] c-2,3,4,7,8,9,11,12,14,15,16,17-dodecahydro-1H-cyclopenta [a] phenanthren-3-ol	36.23	1.45	0.78
MOL000224	(4E,6E)-1,7-bis (3,4-dihydroxyphenyl) hepta-4,6-dien-3-one	33.06	0.29	0.31
MOL000380	(6aR,11aR)-9,10-dimethoxy-6a,11a-dihydro-6H-benzofurano [3,2-c] chromen-3-ol	64.26	0.93	0.42
MOL000442	1,7-Dihydroxy-3,9-dimethoxy pterocarpene	39.05	0.89	0.48
MOL000235	1,7-Diphenyl-3,5-dihydroxy-1-heptene	49.01	0.61	0.18
MOL000238	1,7-Diphenyl-5-hydroxy-6-hepten-3-one	32.65	0.8	0.18
MOL000371	3,9-Di-O-methylnissolin	53.74	1.18	0.48
MOL000260	5-[(2R,3R)-7-methoxy-3-methyl-5-[(E)-prop-1-enyl]-2,3-dihydrobenzofuran-2-yl]-1,3-benzodioxole	65.55	1.27	0.4
MOL000374	5’-Hydroxyiso-muronulatol-2’,5’-di-O-glucoside	41.72	–2.47	0.69
MOL000242	7-O-Methyleriodictyol	56.56	0.46	0.27
MOL000378	7-O-Methylisomucronulatol	74.69	1.08	0.3
MOL000379	9,10-Dimethoxypterocarpan-3-O-β-D-glucoside	36.74	–0.63	0.92
MOL000471	Aloe emodin	83.38	–0.12	0.24
MOL000243	Alpinolide peroxide	87.67	0.51	0.19
MOL001439	Arachidonic acid	45.57	1.2	0.2
MOL000358	Beta-sitosterol	36.91	1.32	0.75
MOL000387	Bifendate	31.1	0.15	0.67
MOL000417	Calycosin	47.75	0.52	0.24
MOL008998	Cerevisterol	39.52	0.35	0.77
MOL008999	Cholesteryl palmitate	31.05	1.45	0.45
MOL000953	CLR	37.87	1.43	0.68
MOL000274	Cordycepin	45.37	0.79	0.87
MOL002297	Daucosterol_qt	35.89	1.35	0.7
MOL000258	Dehydrodiisoeugenol	56.84	1.19	0.29
MOL002288	Emodin-1-O-beta-D-glucopyranoside	44.81	–1.12	0.8
MOL002235	EUPATIN	50.8	0.53	0.41
MOL000433	FA	68.96	–1.5	0.71
MOL000392	Formononetin	69.67	0.78	0.21
MOL000554	Gallic acid-3-O-(6’-O-galloyl)-glucoside	30.25	–1.96	0.67
MOL000296	Hederagenin	36.91	1.32	0.75
MOL000398	Isoflavanone	109.99	0.53	0.3
MOL000439	Isomucronulatol-7,2’-di-O-glucosiole	49.28	–2.22	0.62
MOL000354	Isorhamnetin	49.6	0.31	0.31
MOL000239	Jaranol	50.83	0.61	0.29
MOL000422	Kaempferol	41.88	0.26	0.24
MOL001645	Linoleyl acetate	42.1	1.36	0.2
MOL000006	Luteolin	36.16	0.19	0.25
MOL000211	Mairin	55.38	0.73	0.78
MOL002251	Mutatochrome	48.64	1.97	0.61
MOL002303	Palmidin A	32.45	–0.36	0.65
MOL011169	Peroxyergosterol	44.39	0.86	0.82
MOL002259	Physciondiglucoside	41.65	–2.64	0.63
MOL000230	Pinocembrin	57.56	0.38	0.2
MOL002260	Procyanidin B-5,3’-O-gallate	31.99	–1.61	0.32
MOL000098	Quercetin	46.43	0.05	0.28
MOL002268	Rhein	47.07	–0.2	0.28
MOL002293	Sennoside D_qt	61.06	–0.7	0.61
MOL002276	Sennoside E_qt	50.69	–0.74	0.61
MOL002280	Torachrysone-8-O-beta-D-(6’-oxayl)-glucoside	43.02	–1.23	0.74
MOL002281	Toralactone	46.46	0.86	0.24

[a]*Caco-2: permeability.*

washed with running water again, dehydrated by graded ethanol, and vitrified by dimethylbenzene, and the neutral resin was used for sealing. Weigert ferric hematoxylin staining solution, Masson cyanating solution, and lichunred fuchsin staining solution were used for Masson staining in sequence. Three fields were randomly selected for each section, and the proportion of interstitial fibrosis area was calculated with ImageJ software. Electron microscopy: the renal tissue was fixed with glutaraldehyde first and with 2% osmium tetroxide solution later. Then, the samples were dehydrated by epoxy propylene, and the epoxy resin 828 was substituted for it at 35 and 45°C for 12 h, respectively.

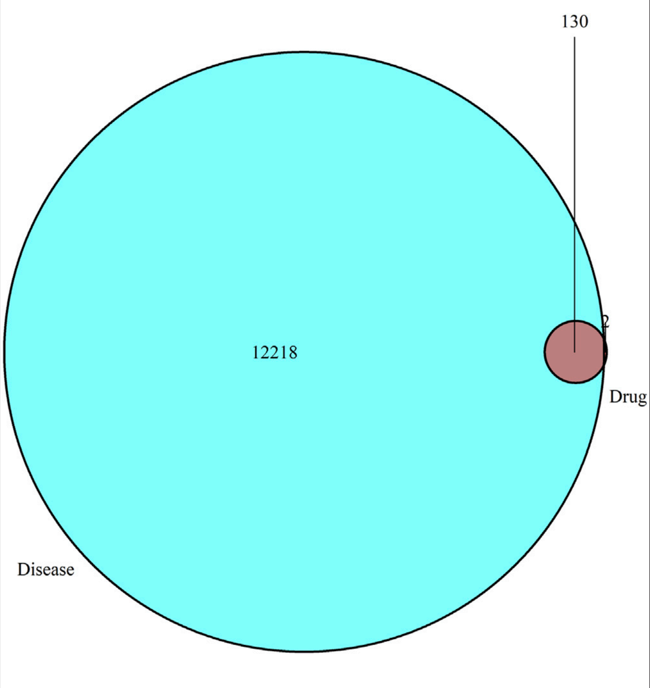


FIGURE 1 | Venn diagram summarizing the intersection targets of CCYSD and CKD.

The specimen was heated at 60°C for 48 h before repairing and cutting into semi-thin slices. Methylene blue was applied to dye and locate the specimens, which were then cut into ultra-thin slices. Uranyl acetate and lead citrate were used for double staining in order to observe the slices under transmission electron microscopy (JED1400PLUS).

Western Blot

Immunoblotting was performed using standard protocols with frozen tissues which were ground in liquid nitrogen. RIPA lysis buffer, benzonase nuclease, protease, and phosphatase inhibitors, used in the study, were added to lysate at room temperature, and the supernatant was extracted by centrifugation at 13,000 r/min for 10 min. Total protein concentration was determined by the BCA method. After 150 V electrophoresis, the membrane was incubated with primary antibody and then secondary antibody. Immunoblots were treated with a chemiluminescence detection system followed by the exposure to Hyperfilm ECL.

Statistical Analysis

All data collected in this study were analyzed using IBM-SPSS Statistics 26.0 software, and the continuous measurement data were expressed as the mean ± standard deviation ($\bar{x} \pm s$). The data, which coincide with normal distribution and satisfy homogeneity variance, between groups were compared through one-way ANOVA (**Table 1**) and that between groups with the Tukey method. Otherwise, significant differences were analyzed by the Kruskal–Wallis test and tested with the Kruskal–Wallis test and Mann–Whitney U rank-sum test (**Figures 7**, **8B**). $p < 0.05$ was considered statistically significant.

RESULTS

Active ingredients in CCYSD: by retrieving the TCMSP database, a total of 53 kinds of non-repeating bioactive ingredients of CCYSD were selected, including organic acids, lipids, biophenols, sterols, flavonoids, and other

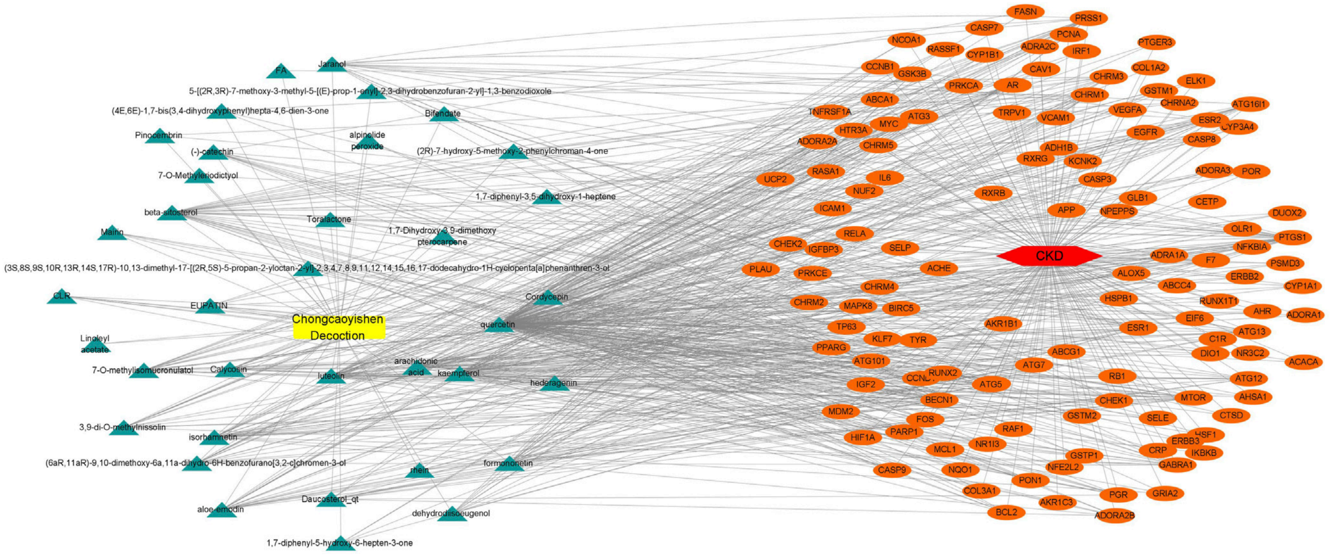


FIGURE 2 | Regulatory network of CCYSD in the treatment of CKD.

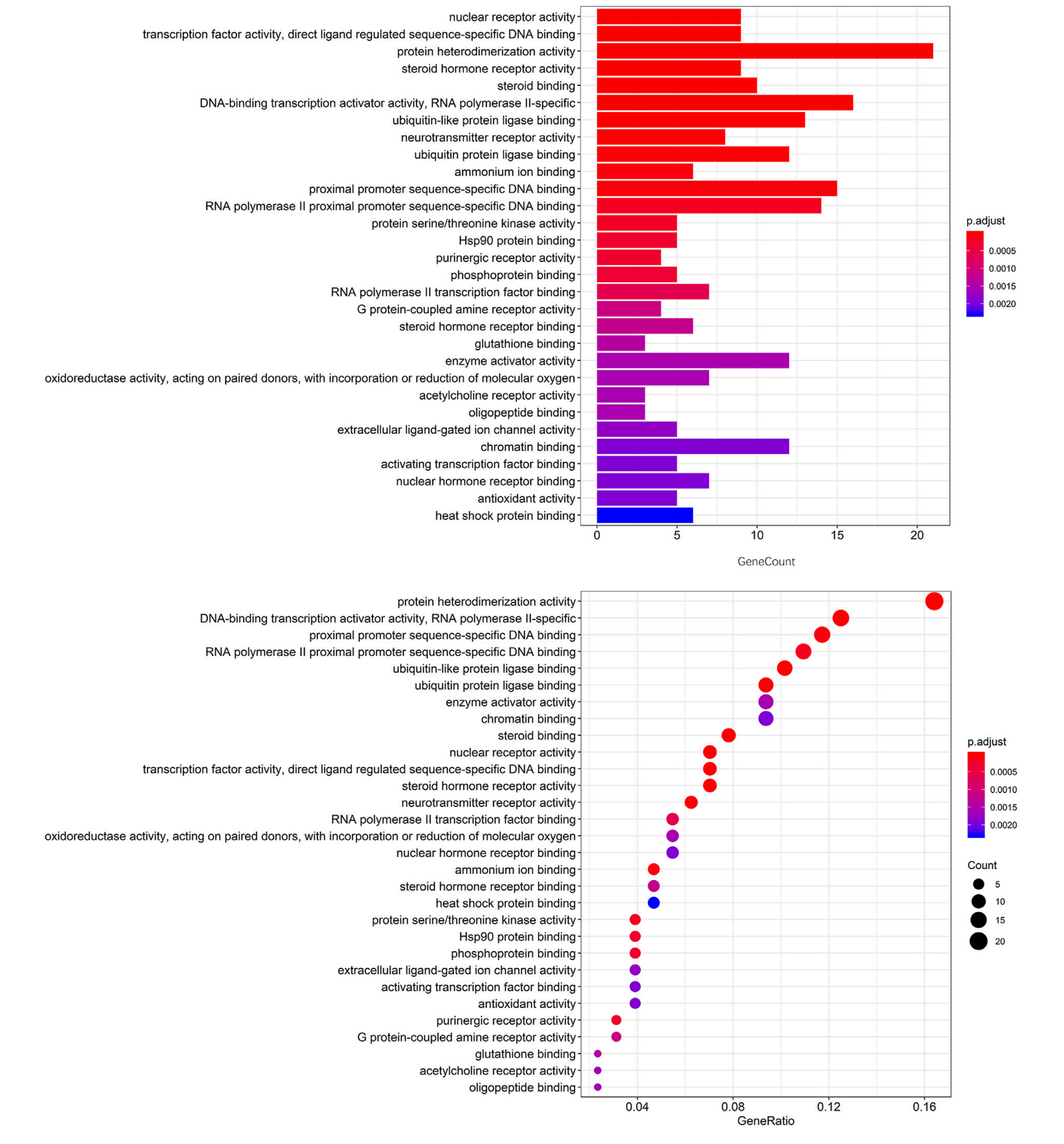


FIGURE 3 | The top 30 of GO enrichment analysis.

main compounds, for example, catechuic acid, arachidonic acid, procyanidin B-5,3’-O-gallic acid, rhein acid, allyl peroxide, biphenyl diester, toralactone, dehydrodiisoeugenol, kaempferol, β-sitosterol, cerevisterol, ergosterol peroxide, calycosin, isoflavone, and foxglove flavonoid (**Table 2**).

Potential Target Prediction of CCYSD

To identify the intersection of CCYSD ingredients and CKD targets, a Venn diagram analysis was carried out. A total of 132 potential targets were identified, which matched with the related targets of 53 active ingredients. 12,348 non-repeating targets closely related to CKD were screened from GeneCards and

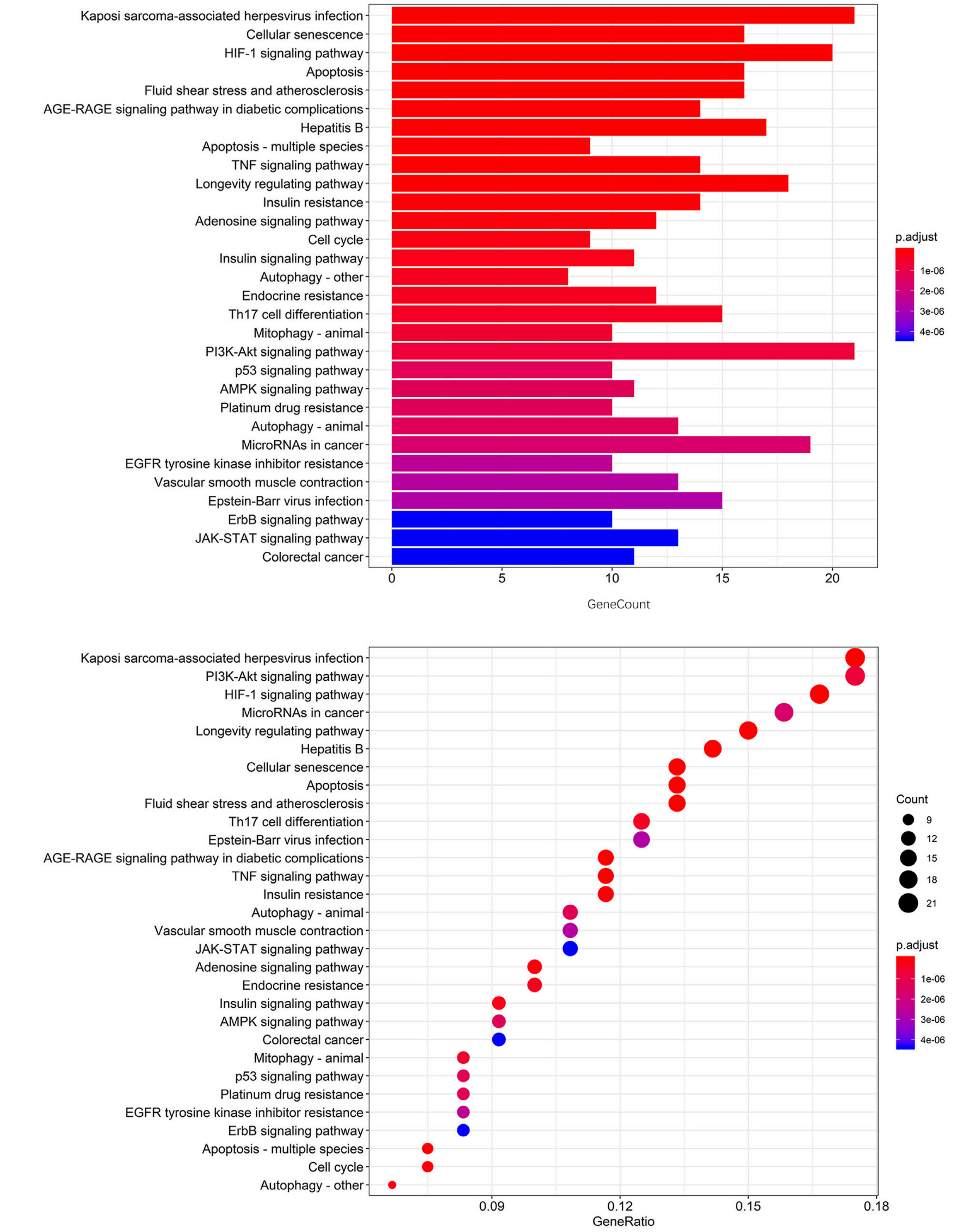


FIGURE 4 | The top 30 signaling pathways from KEGG analysis.

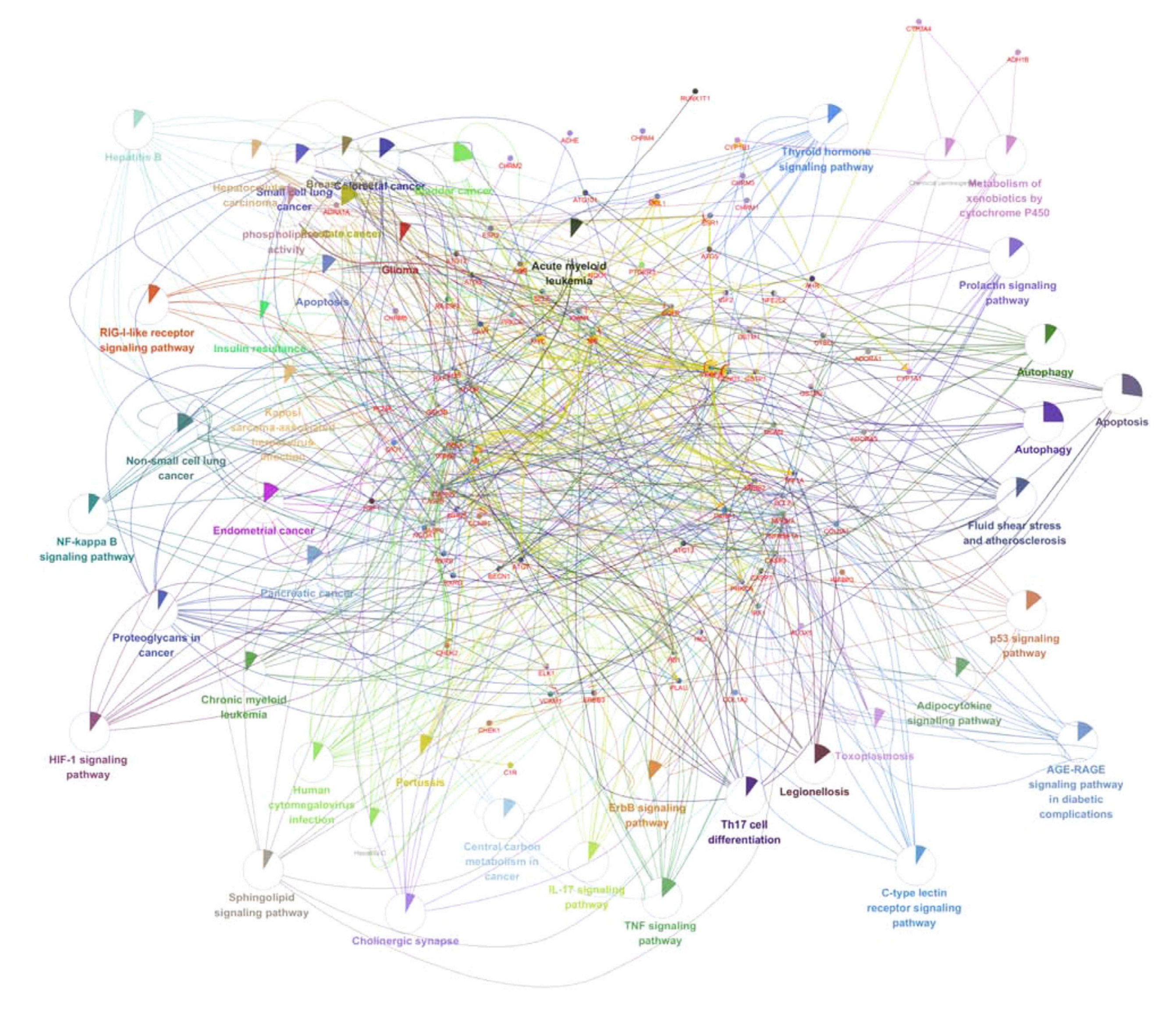

CYP3A4
ADH1B
RUNX1T1
Hepatitis B
Thyroid hormone signaling pathway
Hepatocellular carcinoma
Small cell lung cancer
Breast cancer
Colorectal cancer
Bladder cancer
Metabolism of xenobiotics by cytochrome P450
phospholipase activity
Prostate cancer
Glioma
Acute myeloid leukemia
Prolactin signaling pathway
Apoptosis
RIG-I-like receptor signaling pathway
Insulin resistance
Autophagy
Apoptosis
Kaposi sarcoma-associated herpesvirus infection
Autophagy
Non-small cell lung cancer
Fluid shear stress and atherosclerosis
NF-kappa B signaling pathway
Endometrial cancer
Pancreatic cancer
Proteoglycans in cancer
p53 signaling pathway
Chronic myeloid leukemia
Adipocytokine signaling pathway
HIF-1 signaling pathway
AGE-RAGE signaling pathway in diabetic complications
Toxoplasmosis
Human cytomegalovirus infection
Pertussis
Legionellosis
ErbB signaling pathway
Th17 cell differentiation
Central carbon metabolism in cancer
Sphingolipid signaling pathway
IL-17 signaling pathway
C-type lectin receptor signaling pathway
TNF signaling pathway
Cholinergic synapse

FIGURE 5 | Mechanism network of CCYSD in the treatment of CKD.

OMIM databases. As shown in **Figure 1**, 130 intersecting targets were obtained after matching.

Regulatory Network Construction and Key Targets

To identify the core proteins of CCYSD intervention for CKD, a regulatory network was constructed by matching the ingredients of CCYSD with CKD targets and lines were drawn between them. At the same time, the importance of components and targets was evaluated according to the degree of connection between the components and targets in the regulatory network. As shown in **Figure 2**, quercetin, foxglove flavonoids, 7-O-methylthiamine, isorhamnetin, catechuic acid, aloe-emodin, kumatakenin, and cordycepin are at key positions in the regulatory network of CCYSD, which are closely related to multiple gene targets. Among the gene targets, PRSS1, CHRM3, ATG16l1, AR, PTGS1, MTOR, NFKBIA, ALOX5, ESR1, CRP, GABRA1, ADORA2B, COL3A1, ATG101, HIF1A, CASP9, ADORA2A, MAPK8, ATG 3. ATG5, ATG7, BECN1, ESR1, IL6, BCL2, and other gene targets are important in the regulatory network, indicating that these genes are involved in the occurrence and development of CKD and are regulated by CCYSD in different degrees, which may be central targets for the therapeutic effect of CCYSD.

Enrichment Analysis of GO and KEGG

To further explore the potential signaling pathways regulated by CCYSD, the functional enrichment analyses of GO and KEGG were carried out based on the potential targets of CCYSD *via* ClueGO plug-in from Cytoscape 3.7.1. Filtering criteria was set as *p* value cutoff = 0.05 and *q* value cutoff = 0.05. Finally, 114

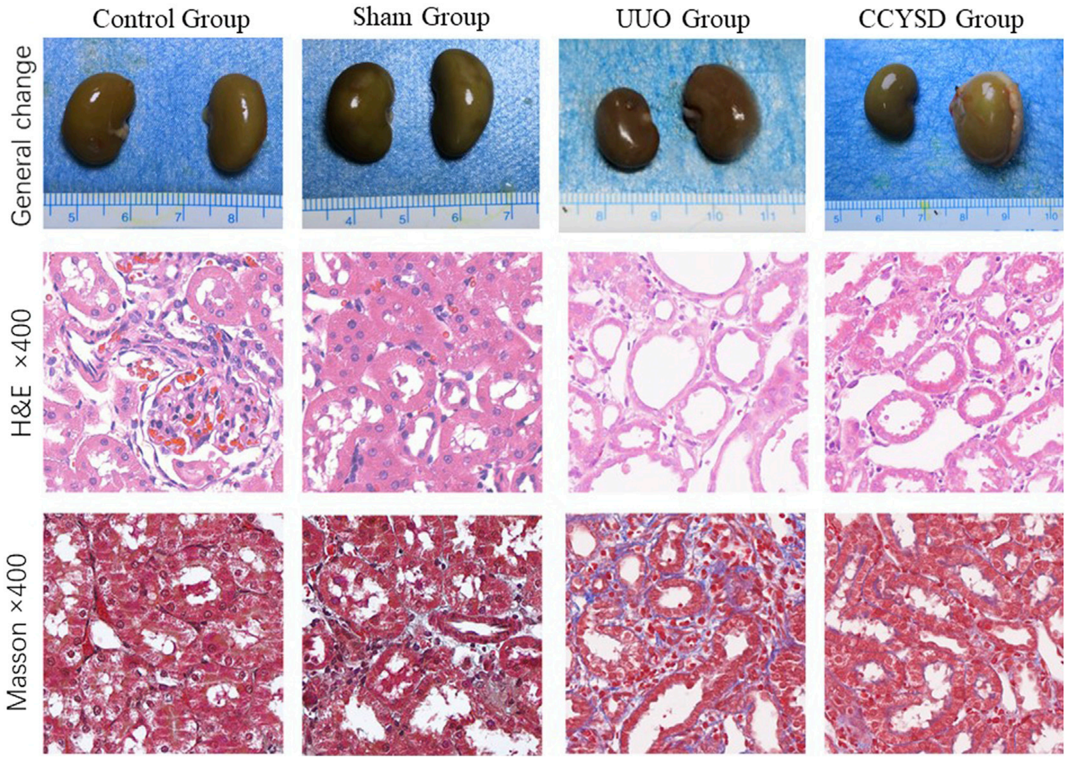


FIGURE 6 | Pathological changes of renal tissue.

biological progresses were obtained by GO functional enrichment analysis, and 112 related cell signaling pathways were obtained by KEGG pathway enrichment analysis (**Figures 3**, **4**).

Subsequently, 45 signaling pathways and functional activities with kappa >0.74 and $p < 0.01$ were labeled in the network by analyzing kappa statistics (**Figure 5**). As showcased in **Figure 5**, the pharmacological mechanisms of CCYSD in the treatment of chronic kidney disease may be involved with autophagy, apoptosis, the p53 signaling pathway, Th17-cell differentiation, adipocytokine signaling, C-type lectin receptor signaling, the ErbB signaling pathway, the TNF signaling pathway, the IL-17 signaling pathway, cholinergic protrusion signaling, the sheath ester signaling pathway, the NF-κB signaling pathway and HIF-1 signaling pathway, the RIG-1-like receptor signaling pathway, the phosphokinase C signaling pathway, and the cytochrome P450 system phagocytic metabolism of heterologous substances and other mechanisms.

Both GO and KEGG enrichment analysis suggested that the therapeutic effect of CCYSD on CKD was mostly related to the regulation of oxidative stress injury, inflammatory response, apoptosis, and autophagy in renal tissues. Further exploration of the abovementioned mechanisms in treating CKD and delaying renal tubulointerstitial fibrosis would be conducted in the subsequent experiments *in vivo*.

Pathological Changes of Renal Tissue

To further verify the key pharmacological mechanism of CCYSD in the treatment of CKD as predicted previously, UUO rats were constructed and administered intragastrically with CCYSD *in vivo*. According to **Figure 6**, the obstructed kidney tissues of UUO animals were significantly larger than that of the normal ones. Compared with the control group, the histological changes of the renal interstitium were examined by H&E staining, which exhibited the widened renal interstitium, dilated renal tubule, and a wide range of exfoliation in the brush border of renal tubular epithelial cells in the UUO group. More infiltration of inflammatory cells and interstitial hemorrhage were also seen in some fields. The results of Masson staining showed increased extracellular matrix and fibrous proliferation in the renal interstitium. Although obvious tubule damage and renal interstitial fibrosis were also observed in the CCYSD group, these pathological changes mentioned above can be reversed by CCYSD to some extent.

Degree of Renal Tissue Fibrosis

To evaluate the TIF, we detected fibrotic molecular marker (α-SMA and COL-III) expression levels and calculated the area of TIF with Masson staining. Compared with the control group, the expression levels of α-SMA and Col-III and the area of TIF in UUO renal tissues were significantly higher ($p < 0.05$). However, compared with the UUO group, the administration of CCYSD significantly meliorated fibrosis ($p < 0.05$) by lowering the expression levels of α-SMA and COL-III and shrinking the area of fibrosis ($p < 0.05$) (**Figure 7**).

Oxidative Stress Level in the Body

In order to evaluate the levels of oxidative stress in rats, we invalidated the expression changes of MDA, SOD, and GSH in serum. It was

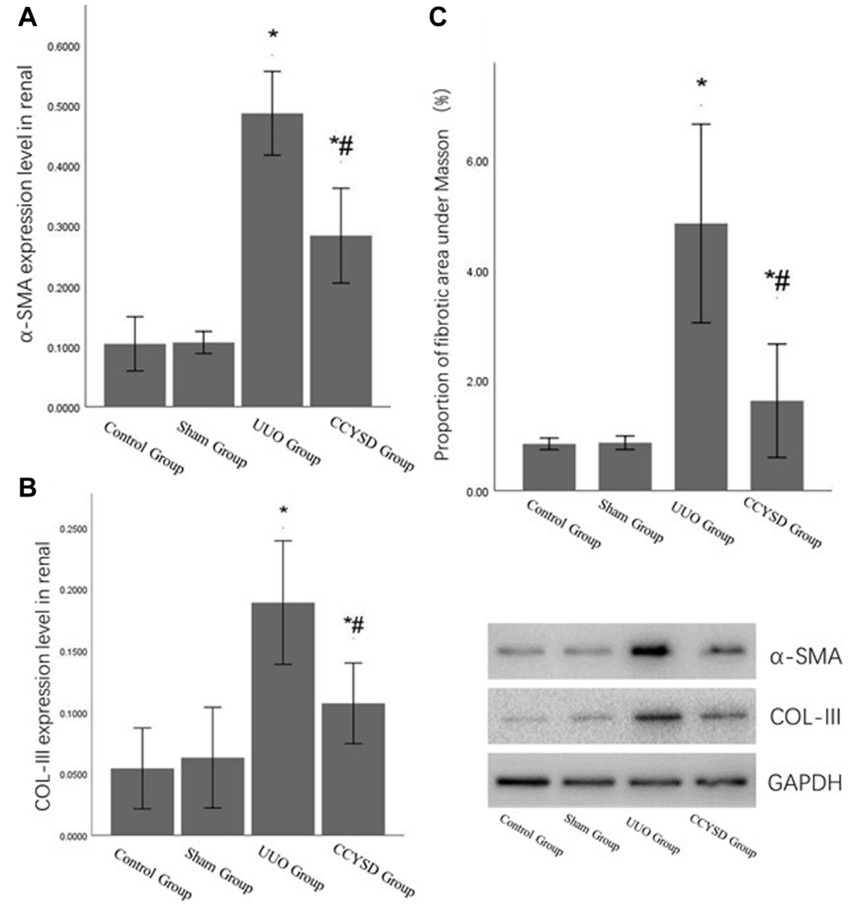


FIGURE 7 | TIF levels in each group. **(A)**: α-SMA relative expression level in renal tissue; **(B)**: col-III relative expression level in renal tissue; **(C)**: the area of TIF with Masson staining; *compared with the sham group, $p < 0.05$; and $^{\#}$compared with the UUO group, $p < 0.05$.

found that compared with the sham group, the serum MDA level in the UUO group was significantly increased, while the levels of SOD and GSH were markedly decreased ($p < 0.05$) (**Table 1**). This kind of fluctuation from serum largely reflected the oxidative stress injury levels in kidney tissues (Hruska et al., 2017; Nordholm et al., 2018). However, CCYSD treatment significantly reversed these alterations in UUO rats, which indicates the increase in SOD and GSH and decrease in MDA ($p < 0.05$). All these findings suggested that CCYSD administration significantly attenuated the oxidative stress injury of CKD, which was consistent with the prediction results of network pharmacology analysis given above. Oxidative stress injury is likely to be the crucial target of CCYSD in treating CKD.

Autophagy Level in Renal Tissue

To examine the effects of CCYSD on autophagy in renal tissue, the number and morphology of autophagosomes and autophagy-lysosomes were determined by transmission electron microscopy (TEM) (**Figure 8A**). A small amount of autophagosomes existed in normal renal tissues to maintain the circulation of substances in cells. On the contrary, the number of autophagosomes and autophagy-lysosomes was dramatically increased in renal tubular epithelial cells from the UUO group and further increased in the CCYSD group. Meanwhile, the level of Atg5 and LC3II/LC3I ratio in renal tissues verified the findings we discovered under TEM. The distinct autophagy was induced in renal tissues in the UUO group, and the activity of autophagy can be largely enhanced after the intervention of CCYSD ($p < 0.05$) (**Figure 8C**).

In addition, we also observed significant differences in the degree of mitochondrial damage in the renal tubule epithelial cells of each group under TEM. We found that severe mitochondrial damage observed in the renal tubular epithelial cells of the UUO group was relieved by treatment with CCYSD (**Figure 8B**). At the same time, we also discovered many subcellular structures such as mitochondria in some autophagosomes in CCYSD groups. Therefore, we speculated that the mitochondrial protective effect of CCYSD might be related to the activation of autophagy.

DISCUSSION

In recent years, as the prevalence of chronic diseases such as diabetes and hypertension increased with each passing year, chronic kidney disease secondary to that mentioned above has increased rapidly (Webster et al., 2017), which provokes

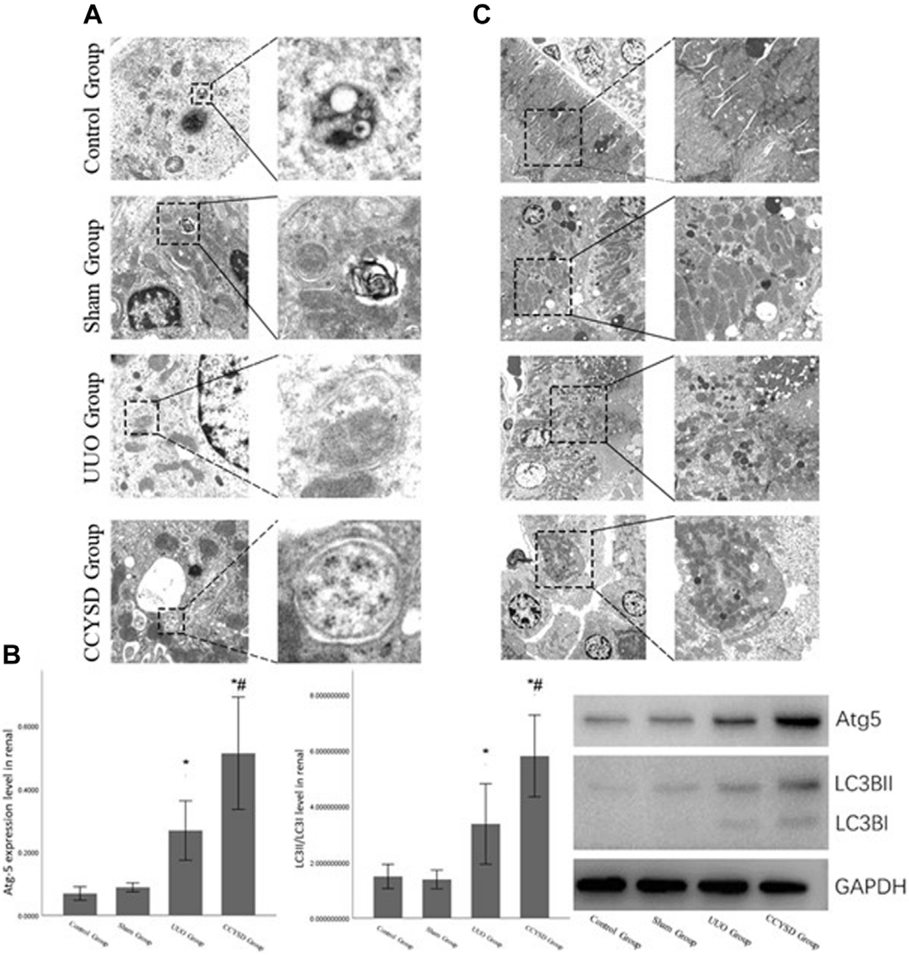


FIGURE 8 | Autophagy and mitochondrial damage in the renal tissue. **(A)**: autophagosomes and autophagy-lysosomes in the renal tubular epithelial cells of each group under TEM; **(B)**: degree of mitochondrial injury in renal tubular epithelial cells in each group under TEM; **(C)**:the level of Atg5 and LC3II/LC3I ratio in renal tissue; *compared with the sham group, $p < 0.05$; and $^{\#}$compared with the UUO group, $p < 0.05$).

enormous burdens on the public healthcare system around the world. Although growing importance has been attached to the prevention and treatment of CKD, a majority of patients will eventually receive dialysis or kidney transplantation because current treatment can only delay the progression of CKD to a certain extent. Abundant doctors are trying to apply Chinese medicine to the treatment of CKD, and massive research achievements have been obtained from clinical observation.

CCYSD, the renowned traditional Chinese herbal decoction, has been proven to be therapeutically effective and widely used in treating CKD for more than 10 years. What we had proved before was not only the clinical curative effect but also the potential mechanisms of CCYSD in treating CKD (Zhang, 2016; Mo et al., 2019a; Mo et al., 2019b). This study aims to detect the mechanism of CCYSD in treating CKD based on network pharmacology analysis, providing a supplement therapy strategy of TCM for CKD, and to further explore the specific mechanisms of CCYSD in treating CKD coupled with subsequent experimental validation. The network pharmacology systematically detected that the active ingredients such as cordycepin, quercetin, luteolin, and kaempferol play an important role in the treatment of CKD. 114 kinds of cellular functional activities and 112 related cellular signaling pathways were identified by GO and KEGG enrichment analysis mainly including apoptosis, autophagy, ubiquitin protein ligase system, protein phosphorylation, G protein-coupled receptor activation and serine/threonine kinase system, purine receptor family, regulation and control of nuclear transcription factor, hypoxia-inducing factor, inflammation, cell cycle regulation, hemodynamic regulation, and vascular endothelial cell injury, as well as a variety of cell functions and signaling pathways.

Notably, several signaling pathways and cell functions are closely related to autophagy, which were directly involved in autophagy regulation appearing in the drug regulatory network of CCYSD against CKD. Due to the vital role of autophagy regulation in CCYSD found through network pharmacological analysis, we carried out further exploration and verification in subsequent *in vivo* rats UUO model validation. Previous studies have confirmed that damaged mitochondria are the main sources of ROS in cells and induce oxidative stress damage in tissues (Shen et al., 2018). Autophagy can specifically degrade damaged mitochondria in cells and avoid the massive release of ROS when mitochondria rupture in order to reduce the oxidative stress damage caused by it. In this study, we found that tubulointerstitial fibrosis is directly related to oxidative stress injury, and CCYSD can delay this kind of fibrosis and significantly reduce oxidative stress injury *in vivo*. Transmission electron microscopy further showed that CCYSD can activate autophagy in epithelial cells and decrease mitochondrial damage. All these findings demonstrated that mitochondrial damage in the process of TIF can compensatively activate autophagy and, thus, play a self-protective role to some extent. Multiple bioactive components in CCYSD activate autophagy to delay TIF and treat CKD by degrading damaged mitochondria and ameliorating oxidative stress injury of tissues. Nevertheless, there were some limitations in the study. First, we only focused on the top 30 compounds and targets in the network pharmacology analysis, while ignoring these ranking after 30, which may attribute to a slight deviation of the results. Second, the validation of potential targets and signaling pathways is limited. Other predicted important targets and pathways not mentioned above require further experimental verification in the coming future.

CONCLUSION

In summary, the pharmacological mechanism of CCYSD on chronic kidney disease may be mainly related to its autophagy activation. CCYSD can promote orderly degradation of damaged mitochondria and avoid mitochondrial rupture and ROS release by activating autophagy, thereby delaying the progression of TIF and CKD. However, 114 kinds of cellular functional activities and 112 related cellular signaling pathways were involved in this network pharmacological analysis. Except for the autophagy and oxidative stress injury, the pharmacological mechanism of CCYSD against CKD may also relate to inflammatory injury, cell cycle regulation, apoptosis, and other mechanisms. We only chose to verify the crucial role of the autophagy activity in the treatment of CKD with CCYSD, while other predicted vital targets and signaling pathways require further experimental verification in the future. Therefore, the secret of “multi-ingredients, multitargets, and multi-pathways mode” in the treatment of CCYSD against CKD needs further exploration.

AUTHOR CONTRIBUTIONS

ZF was responsible for the study design, network pharmacology data analysis, animal experiments and draft writing. JC was responsible for the research design and the draft writing. QY is responsible for the manuscript writing and revision. JH is responsible for network pharmacology data analysis, animal experiment development, draft writing and revision.

FUNDING

This project is supported by the National Science Foundation of China (No. 82104756) and Science Foundation of Zhejiang province (No. LQ22H270002).

REFERENCES

Aggarwal, D., and Singh, G. (2020). Effects of Single and Dual RAAS Blockade Therapy on Progressive Kidney Disease Transition to CKD in Rats. *Naunyn-schmiedeberg's Arch. Pharmacol.* 393 (4), 615–627. doi:10.1007/s00210-019-01759-3

Collins, A. J., Foley, R. N., Chavers, B., Gilbertson, D., Herzog, C., Johansen, K., et al. (2012). 'United States Renal Data System 2011 Annual Data Report: Atlas of Chronic Kidney Disease & End-Stage Renal Disease in the United States. *Am. J. Kidney Dis.* 59A7 (1 Suppl. 1), A7–A420. doi:10.1053/j.ajkd.2011.11.015

Couser, W. G., Remuzzi, G., Mendis, S., and Tonelli, M. (2011). The Contribution of Chronic Kidney Disease to the Global burden of Major Noncommunicable Diseases. *Kidney Int.* 80 (12), 1258–1270. doi:10.1038/ki.2011.368

Djudjaj, S., and Boor, P. (2019). Cellular and Molecular Mechanisms of Kidney Fibrosis. *Mol. aspects Med.* 65 (1), 16–36. doi:10.1016/j.mam.2018.06.002

Fan, Z. L., Fang, Y., Jing, X. F., Xiang, Y., Wang, H. Y., and Yun, J. (2019). Effects of Cordycepin and Hirudin on Inflammatory Response, EMT and RIF in UUO Rats [J]. *Shi Zhen Traditional Chin. Med.* 30 (10), 2305–2310. doi:10.3969/j.issn.1008-0805.2019.10.001

Gewin, L. S. (2018). Renal Fibrosis: Primacy of the Proximal Tubule. *Matrix Biol.* 68-69 (1), 248–262. doi:10.1016/j.matbio.2018.02.006

Hruska, K. A., Sugatani, T., Agapova, O., and Fang, Y. (2017). The Chronic Kidney Disease - Mineral Bone Disorder (CKD-MBD): Advances in Pathophysiology. *Bone* 100 (7), 80–86. doi:10.1016/j.bone.2017.01.023

Humphreys, B. D. (2018). Mechanisms of Renal Fibrosis. *Annu. Rev. Physiol.* 80 (1), 309–326. doi:10.1146/annurev-physiol-022516-034227

Jie, Y. (2015). *Using Differential Protein Group to Study the Effect of Cordyceps Sinensis Yishi Formula on Renal Tissue in Unilateral Ureteral Obstruction Rats [D]*. Harbin: Heilongjiang University of Chinese Medicine.

Ma, X. P., Song, L. Q., Yang, X. L., Zhou, Q., Li, J. Q., and Fan, Z. L. (2020). Effect of Cordyceps Yishen Prescription on Wnt/β -binding Signaling Pathway in Unilateral Ureteral Obstruction Rats [J]. *Med. Rev.* 26 (03), 590–596. doi:10.3969/j.issn.1006-2084.2020.03.034

Mo, T. R., Song, L. Q., Fu, Q., Yun, J., Shen, Z. Y., and Fan, Z. L. (2019a). Effects of Cordyceps Yishen Prescription on Notch1 and Related microRNA Expression in UUO Rats [J]. *J. Liaoning Univ. Traditional Chin. Med.* 21 (07), 42–46. doi:10.13194/j.issn.1673-842x.2019.07.011

Mo, T. R., Yun, J., Shen, Z. Y., Fan, Z. L., and Song, L. Q. (2019b). Effects of Cordyceps Sinensis Yishen Formula on Notch Signaling Pathway in Unilateral Ureteral Obstruction Rats [J]. *Sichuan Traditional Chin. Med.* 37 (06), 39–43. doi:CNKI:SUN:SCZY.0.2019-06-015

National Kidney Foundation (2012). KDOQI Clinical Practice Guideline for Diabetes and CKD: 2012 Update. *Am. J. Kidney Dis.* 60 (5), 850–886. doi:10.1053/j.ajkd.2012.07.005

Nordholm, A., Mace, M. L., gravesen, E., Hofman-Bang, J., Morevati, M., Olgaard, K., et al. (2018). Klotho and Activin A in Kidney Injury: Plasma Klotho Is Maintained in Unilateral Obstruction Despite No Upregulation of Klotho Biosynthesis in the Contralateral Kidney. *Am. J. Physiology-Renal Physiol.* 314 (5), F753–F762. doi:10.1152/ajprenal.00528.2017

Shen, Q., Bi, X., Ling, L., and Ding, W. (2018). 1,25-Dihydroxyvitamin D3 Attenuates Angiotensin II-Induced Renal Injury by Inhibiting Mitochondrial Dysfunction and Autophagy. *Cell Physiol. Biochem.* 51 (4), 1751–1762. doi:10.1159/000495678

Webster, A. C., Nagler, E. V., Morton, R. L., and Masson, P. (2017). Chronic Kidney Disease. *The Lancet* 389 (10075), 1238–1252. doi:10.1016/s0140-6736(16)32064-5

Zhang, H. J. (2016). *Study on the Effect of Cordyceps Yishen Granule on Megsin and p38MAPK Signaling Pathway in Human Glomerular Mesangial Cells [D]*. Harbin: Heilongjiang University of Chinese Medicine.

Ziyang, S. (2019). *Effect of Cordyceps Yishen Prescription on Sonic Hedgehog Signaling Pathway in Renal Tissue of Unilateral Ureteral Obstruction Rats [D]*. Harbin: Heilongjiang University of Traditional Chinese Medicine.

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: 25 April 2022
doi: 10.3389/fmolb.2022.816939

The Value of H2BC12 for Predicting Poor Survival Outcomes in Patients With WHO Grade II and III Gliomas

Jie Zhou[1†], *Zhaoquan Xing*[2†], *Yilei Xiao*[3], *Mengyou Li*[3], *Xin Li*[3], *Ding Wang*[4] and *Zhaogang Dong*[4*]

[1]Department of Nursing, Liaocheng Vocational and Technical College, Liaocheng, China, [2]Department of Urology, Qilu Hospital of Shandong University, Jinan, China, [3]Department of Neurosurgery, Liaocheng People's Hospital, Liaocheng, China, [4]Department of Clinical Laboratory, Qilu Hospital of Shandong University, Jinan, China

Purpose: Glioma is a common primary malignant brain tumor. Grade II (GII) gliomas are prone to develop into anaplastic grade III (GIII) gliomas, which indicate a higher malignancy and poorer survival outcome. This study aimed to satisfy the increasing demand for novel sensitive biomarkers and potential therapeutic targets in the treatment of GII and GIII gliomas.

Methods: A TCGA dataset was used to investigate the expression of H2BC12 mRNA in GII and GIII gliomas and its relation to clinical pathologic characteristics. Glioma tissues were collected to verify results from the TCGA dataset, and H2BC12 mRNA was detected by RT-qPCR. ROC analysis was employed to evaluate the classification power for GII and GIII. The significance of H2BC12 mRNA GII and GIII gliomas was also investigated. In addition, H2BC12 expression-related pathways were enriched by gene set enrichment analysis (GSEA). DNA methylation level and mutation of H2BC12 were analyzed by the UALCAN and CBioPortal databases, respectively.

Results: Based on the sample data from multiple databases and RT-qPCR, higher expression of H2BC12 mRNA was found in GII and GIII glioma tissue compared to normal tissue, which was consistent with a trend with our clinical specimen. H2BC12 mRNA had a better power in distinguishing between GII and GIII and yielded an AUC of 0.706 with a sensitivity of 76.9% and specificity of 81.8%. Meanwhile, high H2BC12 levels were associated with IDH status, 1p/19q codeletion, primary therapy outcome, and the histological type of gliomas. Moreover, the overall survival (OS), disease-specific survival (DSS), and progress-free interval (PFI) of GII glioma patients with higher levels of H2BC12 were shorter than those of patients with lower levels as well as GIII patients. In the multivariate analysis, a high H2BC12 level was an independent predictor for poor survival outcomes of gliomas. The Wnt or PI3K-AKT signaling pathways, DNA repair, cellular senescence, and DNA double-strand break repair were differentially activated in

Edited by:
Sheng Li,
Zhongnan Hospital, Wuhan University, China

Reviewed by:
Jingzhe Han,
Harrison International Peace Hospital, China
Fangang Meng,
Capital Medical University, China

****Correspondence:***
Zhaogang Dong
zhaogang.dong@email.sdu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics,
a section of the journal Frontiers in Molecular Biosciences

Received: *17 November 2021*
Accepted: *23 March 2022*
Published: *25 April 2022*

Citation:
Zhou J, Xing Z, Xiao Y, Li M, Li X, Wang D and Dong Z (2022) The Value of H2BC12 for Predicting Poor Survival Outcomes in Patients With WHO Grade II and III Gliomas. Front. Mol. Biosci. 9:816939. doi: 10.3389/fmolb.2022.816939

Abbreviations: WHO, World Health Organization; GSEA, gene set enrichment analysis; ssGSEA, single sample gene set enrichment analysis; H2Bub1, H2B monoubiquitination; H2BC12, H2B Clustered Histone 12; MSigDB, Molecular Signatures Database; ssGSEA, single-sample GSEA; ROC, receiver operating characteristic; tROC, time-dependent ROC; OS, overall survival; AUC, area under the curve; DSS, disease-specific survival; PFI, progress-free interval; FDR, False Discovery Rate; WT, wild type; MUT, mutant type.

phenotypes that were positively associated with H2BC12. H2BC12 DNA methylation was high in TP53 nonmutant patients, and no H2BC12 mutation was observed in gliomas patients.

Conclusion: H2BC12 is a promising biomarker for the diagnosis and prognosis of patients with WHO grade II and III gliomas.

Keywords: H2BC12, TCGA, diagnosis, prognosis, gliomas

INTRODUCTION

Gliomas are tumors that occur at glial cells, which are important for cerebral nerve cells. They constitute the most prevalent primary brain cancer malignancy (Kiran et al., 2019). According to World Health Organization (WHO) classifications, histologically confirmed gliomas can be categorized into four grades: I, II, III, and IV. This is crucial for appropriate therapeutic strategies or clinical outcomes. Low-grade gliomas (LGG) show highly variable clinical behaviors (Cancer Genome Atlas Research Network et al., 2015) and correlate to a more favorable survival outcome. However, they still carry a 70% risk of disease progression within 10 years (Kiran et al., 2019). Aggressive and proliferative high-grade gliomas (HGG) show an unfavorable course, even when treated by surgical resection, radiotherapy, or chemotherapy that could prolong survival (Stupp et al., 2005; Huang et al., 2017). Treatment and prognosis also differ substantially among the four grades of glioma. It is worth noting that grade II (GII) gliomas are traditionally considered to have a low degree of malignancy, and they are prone to developing into anaplastic grade III (GIII) gliomas, indicating a higher malignancy with huge social, and medical burdens. Unfortunately, GIII exhibits invasive growth and complex pathological processes due to the lack of biomarkers for diagnosis and individualized treatment. GIII is associated with very poor survival outcomes in comparison to GII, and this has important therapeutic implications (Beppu et al., 2011; Suzuki et al., 2015). Thus, discriminating between GII and GIII gliomas is very important. However, the clinical reality is that clinicians often face difficulty when determining whether a patient has a GII or GIII glioma even if they have the patient's histopathology results. Much scientific research combines GII and GIII as low-grade gliomas, while fewer studies have investigated the difference between GII and GIII, such as differences in survival outcome, key drivers of survival, and biomarkers, etc. The various clinical biomarkers currently used, such as O6-methylguanine-DNA methyltransferase (MGMT), have insufficient sensitivity, and specificity when it comes to gliomas (Wick et al., 2014). Several novel biomarkers for the diagnosis and prognosis of gliomas have been explored, including YPEL1 (Li et al., 2022) and ELK3 (Liu et al., 2021). However, these biomarkers are still not available for clinical use. Therefore, we must find novel biomarkers with high sensitivity and specificity urgently to improve the early diagnosis and molecular-targeted therapy of patients with gliomas.

It is known that a genetic predisposition for tumorigenesis is always accompanied by epigenetic alterations. Genome instability is characterized by the accumulation of genetic alterations such as point mutations, copy number alterations, or changes in chromosome numbers, and structures (Hanahan and Weinberg, 2011). For example, aberrant histone modifications can potentially enhance the oncogenic drivers in disease progression, metastatic potential, and resistance to therapy (Müller and Almouzni, 2017). Structurally, histone modification-related proteins are responsible for the compact chromatin in nucleosomes and can be modified via diverse enzymes, including histone family genes (H2A, H2B, H3, and H4), two heterodimers (H2A and H2B), and one DNA-associated H3/H4 tetramer (Sansó et al., 2012). Heterodimers H2A and H2B are important in chromatin-related processes including transcription, DNA replication, and repair (Moyal et al., 2011; Chen et al., 2012). It has been established that H2B monoubiquitination (H2Bub1) at lysine 120 is vitally significant in proper DNA repair, and lacking H2Bub1 is associated with abnormal H2AX phosphorylation, resulting in durable DNA damage response (Kari et al., 2011; Sadeghi et al., 2014). Notably, RNF20/40, ubiquitin ligases indispensable for H2Bub1 were also a part of tumorigenesis (Sethi et al., 2018; Zhou et al., 2021). Recent research indicated that low H2Bub1 expression was prognostic for disease progression, which supported the role of H2Bub1 as a tumor suppressor (Tarcic et al., 2017). Furthermore, loss of H2Bub1 was associated with poor differentiation, cancer stemness, and enhanced malignancy of non-small cell lung cancer (Zhang et al., 2019). Based on the above findings, histone genes might play a crucial role in tumorigenesis, and progression.

Here, H2B Clustered Histone 12 (H2BC12) was investigated in GII and GIII glioma tissue, assessing biomarkers for gliomas, associations with clinical characteristics, prediction survival outcome values, and the involved biological pathways. The methylation level and mutation of H2BC12 were also analyzed. Our findings suggested that H2BC12 might be recognized as a promising biomarker for the prognosis of GII and GIII gliomas.

MATERIALS AND METHODS

Data Acquisition

Target RNA-seq data in TPM format, which were documented in TCGA and GTEx databases, were jointly processed by Toil workflow software (Vivian et al., 2017) and then downloaded from UCSC XENA (https://xenabrocwser.net/datapages/). TCGA database was searched for GII and GIII gliomas tissue (n = 528) and GTEx database was consulted to obtain matched normal tissue (n = 1,152). RNA-seq data were log2 transformed.

Corresponding clinical data were also obtained. The inclusion criteria were defined as WHO GII or GIII classified patients with complete prognostic information.

Inclusive and Exclusive Criteria of Enrolled Patients for the Construction of Risk Signature

The inclusive criteria of patients with gliomas for model construction were as follows: 1) patients with primary gliomas; 2) pathologic types of WHO II or III grade; 3) complete clinicopathological parameters; 4) only samples with RNA-sequencing data; 5) overall survival (OS) as the primary endpoint; 6) minimum follow-up of 90 days. The exclusion criteria included 1) patients with recurrent gliomas and 2) incomplete survival status and clinical information.

GSEA Analysis

Hallmark gene set collections, including C2. cp.v7.2. symbols.gmt [Curated] and h. all.v7.2. symbols.gmt [Hallmarks], were retrieved from the Molecular Signatures Database (MSigDB) and chosen as target sets. Correlations between H2BC12 expression and all genes were characterized by R (v.3.6.3), followed by GSEA analysis using R package clusterProfiler (Yu et al., 2012). The significance threshold was set to $|ES|>1$, p. adjust<0.05, and FDR<0.25.

Analysis of Immune Infiltration and Immune Regulatory Factor

From Bindea's investigation (Bindea et al., 2013), the marker gene of 24 immune cells was retrieved. Based on mRNA TPM data, single-sample GSEA (ssGSEA) (Finotello and Trajanoski, 2018) was utilized to quantify the number of tumor-infiltrating immune cells. Spearman correlation was used to determine the relationship between H2BC12 and 24 cells. The ggplot2 package was used to create the figures. Moreover, the correlation between H2BC12 and immune regulatory factors, such as immune inhibitors, immune stimulators, and the MHC molecule from the TISIDB databases (http://cis.hku.hk/TISIDB/), was also analyzed.

DNA Methylation Level and Mutation Analysis of H2BC12

The UALCAN database (Chandrashekar et al., 2017) (http://ualcan.path.uab.edu/index.html) was used to analyze the correlation between the DNA methylation level of the H2BC12 promoter region and the clinical characterization of gliomas. The CBioPortal database (Gao et al., 2013) (http://www.cbioportal.org/) was used to analyze H2BC12 mutation in patients with gliomas.

RNA Extraction and Quantitative Real-Time RT-qPCR

Glioma tissues were collected from the Department of Neurosurgery, Liaocheng People's Hospital (Shandong, China), and they included tissues from 22 GII and 26 GIII gliomas. Tissue RNAs were extracted using the RNAprep pure FFPE kit [cat. no. DP439, TIANGEN Biotech (Beijing) Co., Ltd.] according to instructions. The All-in-one™ First-Strand cDNA Synthesis kit (cat. no. QP006, GeneCopoeia, Inc.) was used to reverse-transcribe an equal amount of total RNA from each sample to cDNA. H2BC12 was detected using the CFX96 qPCR instrument (Bio-Rad Laboratories, Inc.) with the All-in-one™ qPCR Mix (cat. no. QP001, GeneCopoeia, Inc.). The primers for H2BC12 were as follows: forward 5′-AGA AGGGCTCGAAGAAAGCC-3′, reverse 5′-ATGGTCGAGCGC TTGTTGTA-3'. The size was 235 bp. The primers for GAPDH were as follows: forward 5′-GAAGGTGAAGGTCGGAGTC-3′, reverse 5′-GAAGATGGTGATGGGATTTC-3'. The size was 225 bp. The conditions were as follows: following initial denaturation at 95°C 10 min, then 40 cycles of 95°C for 15 s, 62°C for 20 s, and 72°C for 10 s. The amplification specificity was determined by melting curve analysis. Data were normalized to GAPDH, and relative expression levels were evaluated using the $2^{-\Delta\Delta CT}$ method.

Statistical Analysis

R (v.3.6.3) was run to complete all statistical analyses. The diagnostic receiver operating characteristic (ROC) curve was generated using package pROC, while the time-dependent ROC (tROC) curve was plotted with assistance from package timeROC. Differential expression of H2BC12 in gliomas versus normal was statistically analyzed via Wilcoxon rank-sum tests. For correlational analysis between H2BC12 mRNA and clinicopathologic characteristics, tumor samples were assigned to two cohorts representative of high and low H2BC12 expression, respectively, with the cutoff value being the median H2BC12 expression of all samples. A Chi-square test was implemented to identify significance. Comparisons between two sets of data were completed by a Wilcoxon rank-sum test for two groups or the Kruskal–Wallis test when there were three groups or more. Prognostic significance of H2BC12 mRNA expression and clinicopathologic characteristics for overall survival (OS) of gliomas patients were identified by univariate and multivariate Cox regression analysis. The survival significance of H2BC12 mRNA expression in subgroups of clinicopathologic characteristics was investigated by stratification and Kaplan-Meier analysis. p value < 0.05 was considered statistically significant.

RESULTS

Clinical Characteristics

The expression of H2BC12 mRNA and the corresponding clinicopathologic characteristics of 528 primary tumors were obtained from the glioma dataset; of these, 523 RNA-seq datasets were available. Matched clinical data were retrieved: WHO grade II and III, IDH status, 1p/19q codeletion, primary therapy outcome, gender, race, age, histological type, laterality, and OS event (**Table 1**).

High Expression of H2BC12 mRNA in Grade II and III Gliomas Tissue

Apart from gliomas samples acquired, matched normal samples (n = 1,152) were obtained from the GTEx database. H2BC12

TABLE 1 | Characteristics of patients with gliomas based on TCGA.

Characteristic	Levels	Overall
n		528
WHO grade, n (%)	GII	224 (48%)
	GIII	243 (52%)
IDH status, n (%)	WT	97 (18.5%)
	Mut	428 (81.5%)
1p/19q codeletion, n (%)	codel	171 (32.4%)
	non-codel	357 (67.6%)
Primary therapy outcome, n (%)	PD	110 (24%)
	SD	146 (31.9%)
	PR	64 (14%)
	CR	138 (30.1%)
Gender, n (%)	Female	239 (45.3%)
	Male	289 (54.7%)
Race, n (%)	Asian	8 (1.5%)
	Black or African American	22 (4.3%)
	White	487 (94.2%)
Age, n (%)	≤40	264 (50%)
	>40	264 (50%)
Histological type, n (%)	Astrocytoma	195 (36.9%)
	Oligoastrocytoma	134 (25.4%)
	Oligodendroglioma	199 (37.7%)
Laterality, n (%)	Left	256 (48.9%)
	Midline	6 (1.1%)
	Right	261 (49.9%)
OS event, n (%)	Alive	392 (74.2%)
	Dead	136 (25.8%)
DSS event, n (%)	Alive	397 (76.3%)
	Dead	123 (23.7%)
PFI event, n (%)	Alive	318 (60.2%)
	Dead	210 (39.8%)

mRNA was examined in two cohorts, showing a significant upward trend in primary tumor tissue. Furthermore, the level of H2BC12 mRNA in GIII gliomas was higher than that of GII gliomas (**Figure 1A**, $p < 0.001$). And H2BC12 of GIII and GII were both higher than normal. Results of our clinical specimen showed a similar trend between GII and GIII (**Supplementary Figure S1A**, $p < 0.05$). These results revealed that H2BC12 might be an oncogene in gliomas.

ROC Analysis for H2BC12 as a Biomarker of Grade II and III Gliomas

ROC curve was plotted to evaluate the diagnostic significance of H2BC12 mRNA for gliomas. The area under the curve (AUC) was 0.823 with 83.0% sensitivity and 68.4% specificity (**Figure 1B**), indicating significance in distinguishing between normal and tumor samples with certain accuracy. Furthermore, ROC analysis was also performed to compare GII and GIII gliomas. As shown in **Figure 1C**, AUC was 0.632, and the corresponding sensitivity and specificity were 56.5 and 72.5%, achieving a classification power for GIII and GII. The results of our clinical specimen also revealed that the AUC was 0.706 with a sensitivity of 76.9% and specificity of 81.8% (**Supplementary Figure S1C**). It seems that the results from our clinical specimen were better than from the dataset. This indicated that H2BC12 mRNA might be a more reliable biomarker.

Correlations Between H2BC12 mRNA and Clinicopathologic Characteristics of Gliomas

The correlational analysis demonstrated that there were significant associations between the H2BC12 mRNA and clinicopathologic characteristics, including IDH status, 1p/19q codeletion, primary therapy outcome, and histological type (**Figures 2A–D**). Our clinical results showed that H2BC12 mRNA was significantly correlated with IDH status, which was consistent with the conclusions drawn from the TCGA database (**Supplementary Figure S1B**, $p < 0.05$). In addition, tumor samples of each clinicopathologic subgroup were divided into two groups according to the median H2BC12 mRNA. Further analysis revealed that high H2BC12 mRNA expression was significantly associated with WHO grade, IDH status, 1p/19q codeletion, primary therapy outcome, histological type, OS event, disease-specific survival (DSS) event, and progress-free interval (PFI) event (**Table 2**, $p < 0.001$). Collectively, H2BC12 mRNA expression is intimately correlated with clinicopathologic

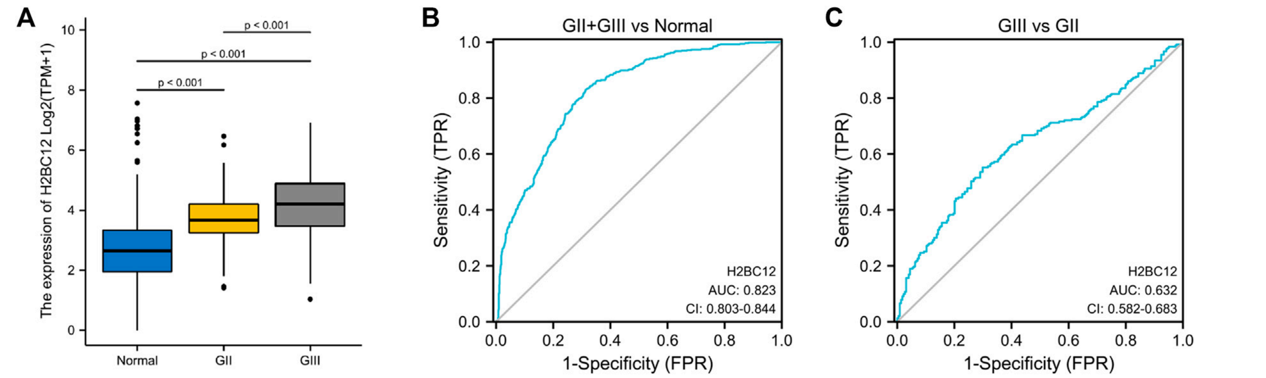


FIGURE 1 | The expression of H2BC12 mRNA in normal, GII, and GIII glioma tissue and its clinical value as a biomarker for distinguishing between GII and GIII gliomas. H2BC12 showed significantly higher expression in GII or GIII tissue versus normal tissue **(A)**, p < 0.001. The diagnostic ROC curve showed the accurate discriminative capability of H2BC12 in distinguishing between normal and GII + GIII (AUC = 0.823). **(B)** ROC analysis of H2BC12 in classification power for GII and GIII (AUC = 0.632).

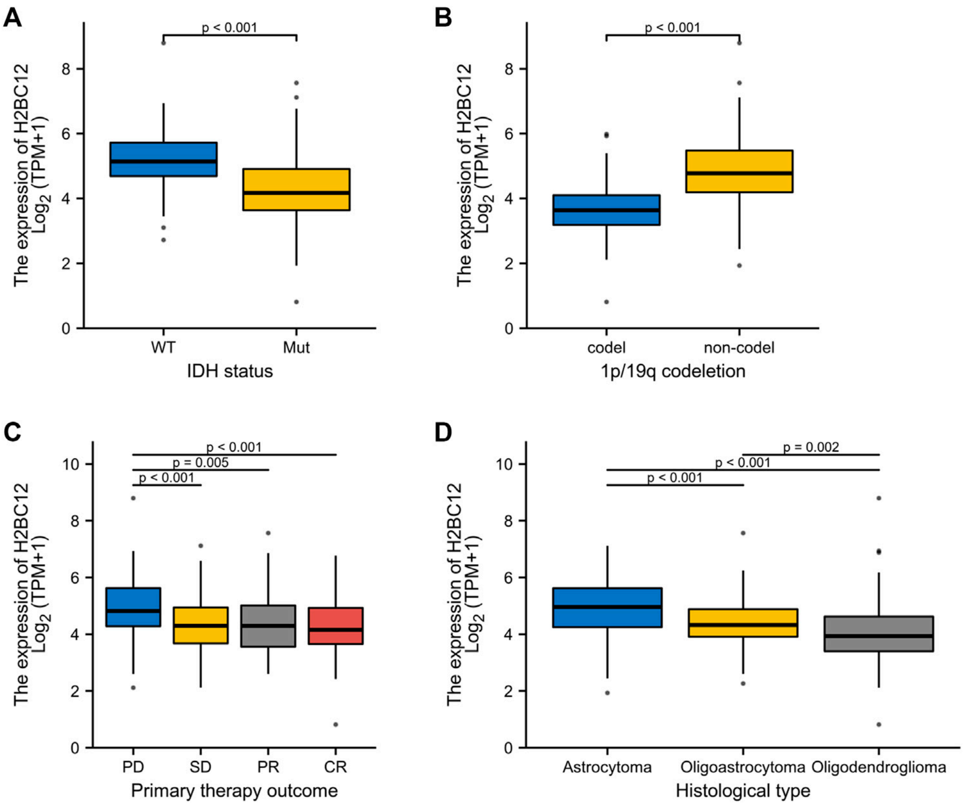


FIGURE 2 | Association between H2BC12 expression and clinicopathologic characteristics. H2BC12 expression correlated significantly with IDH status **(A)**, p < 0.001, 1p/19q codeletion **(B)**, p < 0.001, primary therapy outcome **(C)**, p < 0.01, and histological type **(D)**, p < 0.01.

features, suggesting that H2BC12 might be involved in glioma progression.

Role of H2BC12 in Grade II and III Glioma Patient Survival

Gliomas were considered to have different degrees of malignancy and survival outcomes. However, few studies investigated the relationship between gene expression and survival outcomes for GII and GIII separately. First, we explored the role of H2BC12 in survival outcomes, and **Figure 3A** shows that the OS of GII + GIII patients with high H2BC12 expression was much poorer compared to those with low H2BC12 expression (p < 0.001). Similar results were also observed as regards DSS and PFI (**Figures 3B,C**, p < 0.001). The prognostic value of H2BC12 in GII or GIII was further evaluated. **Figures 3D–F** shows that OS, DSS, and PFI of GII gliomas with higher levels of H2BC12 were shorter than those with lower levels [HR = 3.28 (1.68–6.37) for OS, HR = 3.51 (1.73–7.12) for DSS, and HR = 2.16 (1.38–3.38) for PFI]. A similar trend was also shown in GIII patients, and HR was 2.76 (1.78–4.26) for OS, 3.32 (2.10–5.26) for DSS, and 2.62 (1.78–3.85) for PFI (**Figures 3G–I**, *p* < 0.001). Besides, the tROC curves were drawn to identify the predictive ability of H2BC12 mRNA for OS of GII and/or GIII patients. The AUC values for 1-, 2-, and 3-years OS of GII + GIII were 0.766, 0.702, and 0.677, respectively (**Figure 3J**). The AUC values for 1-, 2- and 3-years GII were 0.492, 0.664, and 0.714 (**Figure 3K**). The AUC values for 1-, 2-, and 3-years GIII were 0.760, 0.675, and 0.6499 (**Figure 3L**). To identify the prognostic factors for OS of gliomas patients, univariate regression analysis was performed using a Cox model, demonstrating significant prognostic significance of H2BC12 mRNA, WHO grade, 1p/19q codeletion, TP53, IDH status, age, and histological type for OS (**Table 3**, p < 0.01). Additionally, a further multivariate model was established and revealed that H2BC12 mRNA, WHO grade, IDH status, age, and histological type had independent prognostic significance for gliomas OS (**Table 3**, p < 0.05). It was suggested that H2BC12 was equipped with a good prognostic performance.

Clinical Stratification

As proven in multivariate Cox regression analysis, primary therapy outcome, IDH status, age, and histological type were independent prognostic factors for glioma OS. Then, clinical stratification was conducted based on the glioma dataset; in subgroups of primary therapy outcomes PD&SD, primary therapy outcome PR&CR, IDH status: Mut, age < = 40, and age >40, patients with low H2BC12 expression had better survival outcomes than those with highly expressing H2BC12 (**Figures 4A–F**, p < 0.001). This reflected that H2BC12 had independent prognostic significance for glioma OS, and increased H2BC12 was associated with poorer OS.

TABLE 2 | Relationship between H2BC12 mRNA expression and clinical characteristics in gliomas.

Characteristic	Levels	Low expression of H2BC12	High expression of H2BC12	p
n		264	264	<0.001
WHO grade, n (%)	GII	138 (29.6%)	86 (18.4%)	
	GIII	95 (20.3%)	148 (31.7%)	
IDH status, n (%)	WT	13 (2.5%)	84 (16%)	<0.001
	Mut	250 (47.6%)	178 (33.9%)	
1p/19q codeletion, n (%)	codel	148 (28%)	23 (4.4%)	<0.001
	non-codel	116 (22%)	241 (45.6%)	
Primary therapy outcome, n (%)	PD	33 (7.2%)	77 (16.8%)	<0.001
	SD	76 (16.6%)	70 (15.3%)	
	PR	36 (7.9%)	28 (6.1%)	
	CR	84 (18.3%)	54 (11.8%)	
Gender, n (%)	Female	117 (22.2%)	122 (23.1%)	0.727
	Male	147 (27.8%)	142 (26.9%)	
Race, n (%)	Asian	4 (0.8%)	4 (0.8%)	0.230
	Black or African American	7 (1.4%)	15 (2.9%)	
	White	245 (47.4%)	242 (46.8%)	
Age, n (%)	≤40	131 (24.8%)	133 (25.2%)	0.931
	>40	133 (25.2%)	131 (24.8%)	
Histological type, n (%)	Astrocytoma	61 (11.6%)	134 (25.4%)	<0.001
	Oligoastrocytoma	69 (13.1%)	65 (12.3%)	
	Oligodendroglioma	134 (25.4%)	65 (12.3%)	
Laterality, n (%)	Left	122 (23.3%)	134 (25.6%)	0.412
	Midline	2 (0.4%)	4 (0.8%)	
	Right	137 (26.2%)	124 (23.7%)	
OS event, n (%)	Alive	237 (44.9%)	155 (29.4%)	<0.001
	Dead	27 (5.1%)	109 (20.6%)	
DSS event, n (%)	Alive	240 (46.2%)	157 (30.2%)	<0.001
	Dead	23 (4.4%)	100 (19.2%)	
PFI event, n (%)	Alive	196 (37.1%)	122 (23.1%)	<0.001
	Dead	68 (12.9%)	142 (26.9%)	

H2BC12-Related Signaling Pathways Based on GSEA

GSEA was performed to find the activated signaling pathways related to H2BC12 in gliomas. Based on the curated collection, there were six signaling pathways activated in H2BC12 overexpressed phenotype, including pathways in cancer, Wnt or the PI3K-AKT signaling pathway, DNA repair, cellular senescence, and DNA double-strand break repair. Based on the Hallmarks collection defined by MSigDB, other than the above six pathways, the KRAS signaling up, TNFA signaling via NFKB, G2M checkpoint, glycolysis, hypoxia, and p53 pathways also presented with significant enrichment in H2BC12 overexpressed phenotype (**Figure 5**; **Table 4**). Collectively, H2BC12 mRNA might serve as an important player in the initiation and development of gliomas.

H2BC12 Expression Was Linked to the Level of Immune Infiltration and Immune Regulatory Factor

Tumor-infiltrating lymphocytes are independent indicators of cancer survival. As result, we evaluated whether H2BC12 was related to immune infiltrate in gliomas. According to our findings, H2BC12 showed a strong positive correlation with macrophages, eosinophils, neutrophils, and T cells; H2BC12 exhibited a strong inverse relationship with pDC, NK CD56bright cells, TReg, and DC (**Figure 6A**). Further analysis showed that compared with the low-H2BC12 group, the infiltration of Neutrophils and T cells in the high-H2BC12 group was significantly increased (**Figure 6B**). The infiltration levels of pDC, NK CD56bright cells, Treg, and DC were significantly reduced in the high-H2BC12 group (**Figure 6C**). Moreover, Results of the relationship of H2BC12 with immune regulatory factors showed that H2BC12 was positively correlated with immun inhibitors, including PDCD1LG2, LGALS9, and L10RB (**Figure 7A**), as well as immune stimulators, including CD40, CD86, and MICB (**Figure 7B**), and MHC molecules, including HLA-DMA, HLA-DMB, and HLA-DOA (**Figure 7C**).

H2BC12 Promoter Methylation Level and Mutation Analysis

The level of DNA methylation in the H2BC12 promoter region in patients with TP53 nonmutant was significantly higher than that in patients with TP53 mutant (**Supplementary Figure S2A**). Moreover, the levels in those aged between 41 and 60 years were significantly higher than in those aged between 21 and 40 years (**Supplementary Figure S2B**). There have not been any significant differences in terms of gender or race yet (**Supplementary Figure S2C, D**). In addition, the H2BC12 mutation was not investigated in glioma patients and was very low in most brain tumors (**Supplementary Figure S2E**).

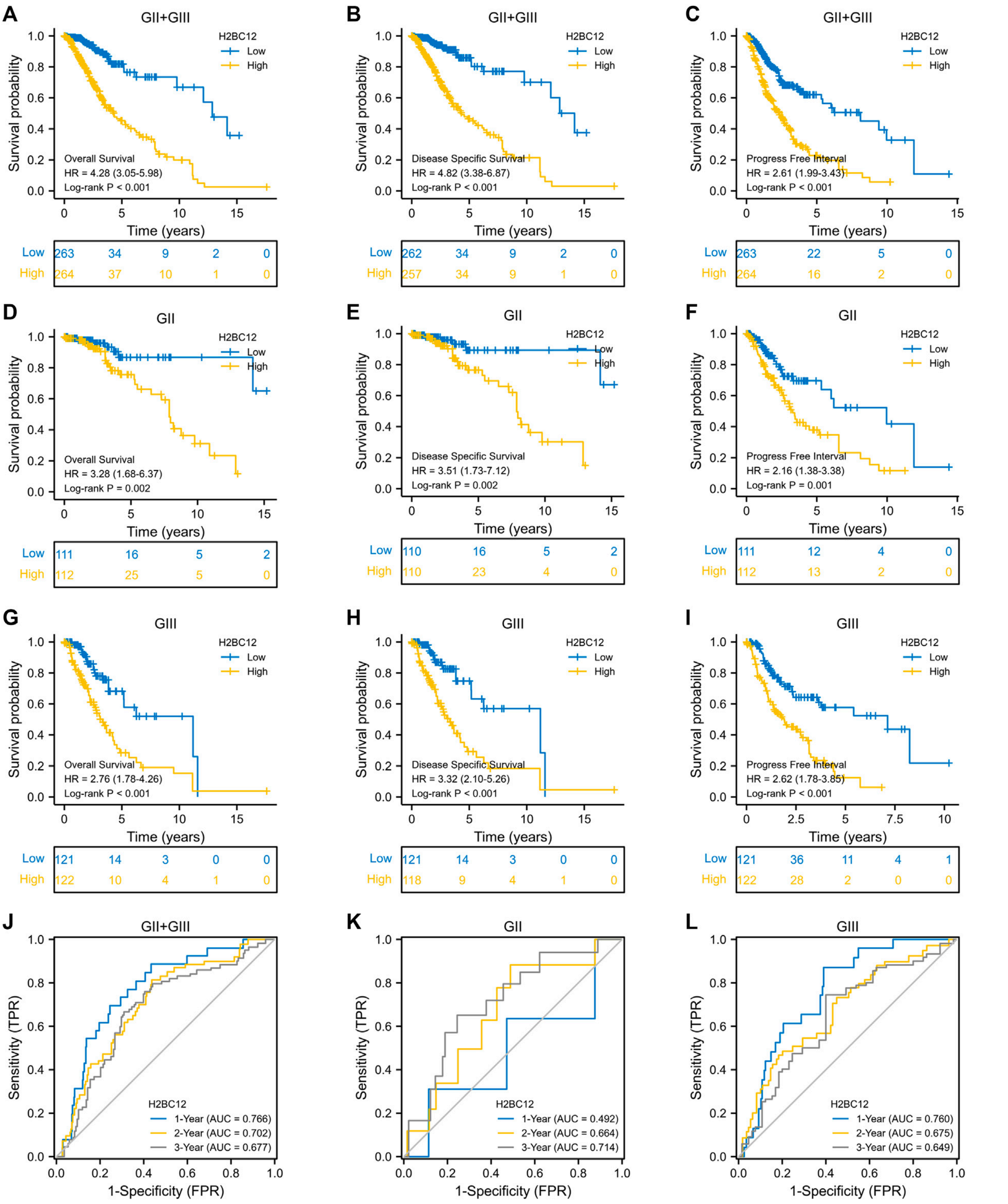


FIGURE 3 | High expression of H2BC12 is associated with poor OS, DSS, and PFI in patients with GII and/or GIII. OS **(A)**, p < 0.001, DSS **(B)**, p < 0.001, and PFI **(C)**, p < 0.001 were significantly poorer in GII + GIII patients with high H2BC12 expression than those with low H2BC12 expression. Furthermore, OS, DSS, and PFI of GII **(D–F)** and GIII **(G–I)** were analyzed respectively. OS, Overall Survival; DSS, Disease-Specific Survival; PFI, Progress-Free Interval. **(J)** tROC curve demonstrated AUC values for 1-, 2-, and 3-years survival in GII + GIII as 0.766, 0.702, and 0.677, respectively. The 1-, 2-, and 3-years AOC values in GII were 0.492, 0.664, and 0.714 **(K).** The 1-, 2-, and 3-years AOC values in GIII were 0.760, 0.675, and 0.6499 **(L)**.

TABLE 3 | Correlations between overall survival and mRNA expression of H2BC12 analyzed by univariate and multivariate Cox regression.

Characteristics	Total(N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	*p* value	Hazard ratio (95% CI)	*p* value
WHO grade (GIII vs. GII)	466	3.059 (2.046–4.573)	**<0.001**	1.845 (1.147–2.967)	**0.012**
1p/19q codeletion (non-codel vs. codel)	527	2.493 (1.590–3.910)	**<0.001**	1.293 (0.670–2.496)	0.443
TP53 (High vs. Low)	527	1.689 (1.189–2.400)	**0.003**	1.352 (0.874–2.091)	0.175
IDH status (Mut vs. WT)	524	0.186 (0.130–0.265)	**<0.001**	0.455 (0.281–0.735)	**0.001**
Gender (Male vs. Female)	527	1.124 (0.800–1.580)	0.499		
Age (>40 vs. ≤40)	527	2.889 (2.009–4.155)	**<0.001**	3.491 (2.191–5.561)	**<0.001**
Histological type (Oligoastrocytoma&Oligodendroglioma vs. Astrocytoma)	527	0.606 (0.430–0.853)	**0.004**	1.018 (0.642–1.615)	0.939
H2BC12 (High vs. Low)	527	4.415 (2.885–6.756)	**<0.001**	2.267 (1.252–4.104)	**0.007**

Bold values indicates that the significant values (p ≤ 0.05).

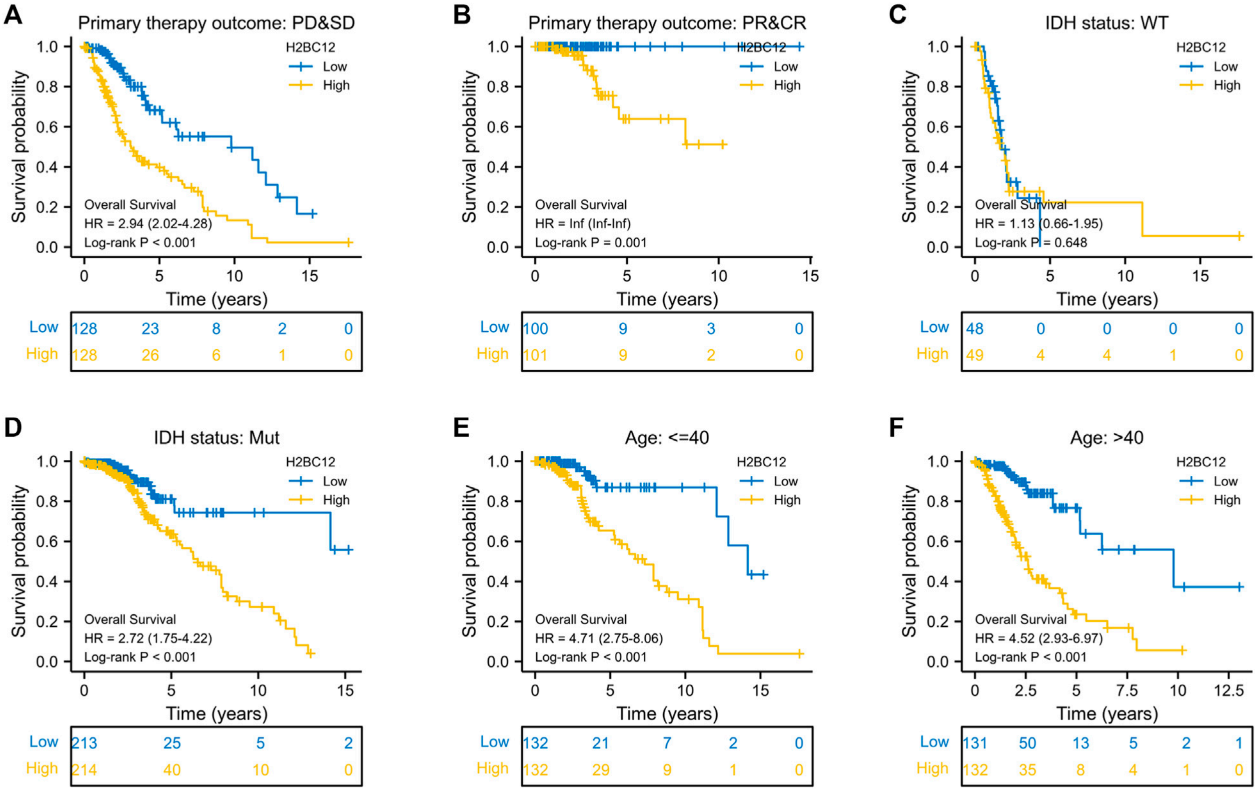


FIGURE 4 | Clinical stratification analysis of the survival difference in the high- and low-H2BC12 groups by primary therapy outcome, IDH status, and age. Kaplan-Meier survival curves of patients in the high- and low-H2BC12 groups within eight clinically stratified subgroups, including primary therapy outcome: PD&SD **(A)**, primary therapy outcome: PR&CR **(B)**, IDH status: WT **(C)**, IDH status: Mut **(D)**, age<=40 **(E)** and age>40 **(F)**, respectively. Patients in the low-H2BC12 group had better survival outcomes than those in the high-H2BC12 group across all clinically stratified subgroups except the IDH status of WT (p < 0.01).

DISCUSSION

Gliomas are fatal tumors most prevalent in the central nervous system (CNS), and they are among the most devastating forms of cancer. Low-grade tumors grow slowly with lesser malignant properties than high-grade tumors (Perez and Huse, 2021; Zhao et al., 2021). However, there is a high risk of disease progression to advanced gliomas in most low-grade glioma patients (Kiran et al., 2019). It is well known that GII gliomas can easily develop into GIII gliomas, which leads to a poor survival outcome after receiving chemotherapy (Xiao et al., 2020). There are no suitable biomarkers to discriminate between GII and GIII gliomas. The role of survival outcome, key drivers of survival, etc. remains to be further explored. According to bioinformatics, the WHO included several molecular markers, such as IDH mutation status and chromosome 1p or 19q codeletion (1p/19q codeletion) status, into the guidelines for the diagnosis of gliomas to increase the accuracy in disease diagnosis and further treatment (Louis et al., 2016). In this context, the demand for biomarkers with prognostic and diagnostic values is increasing,

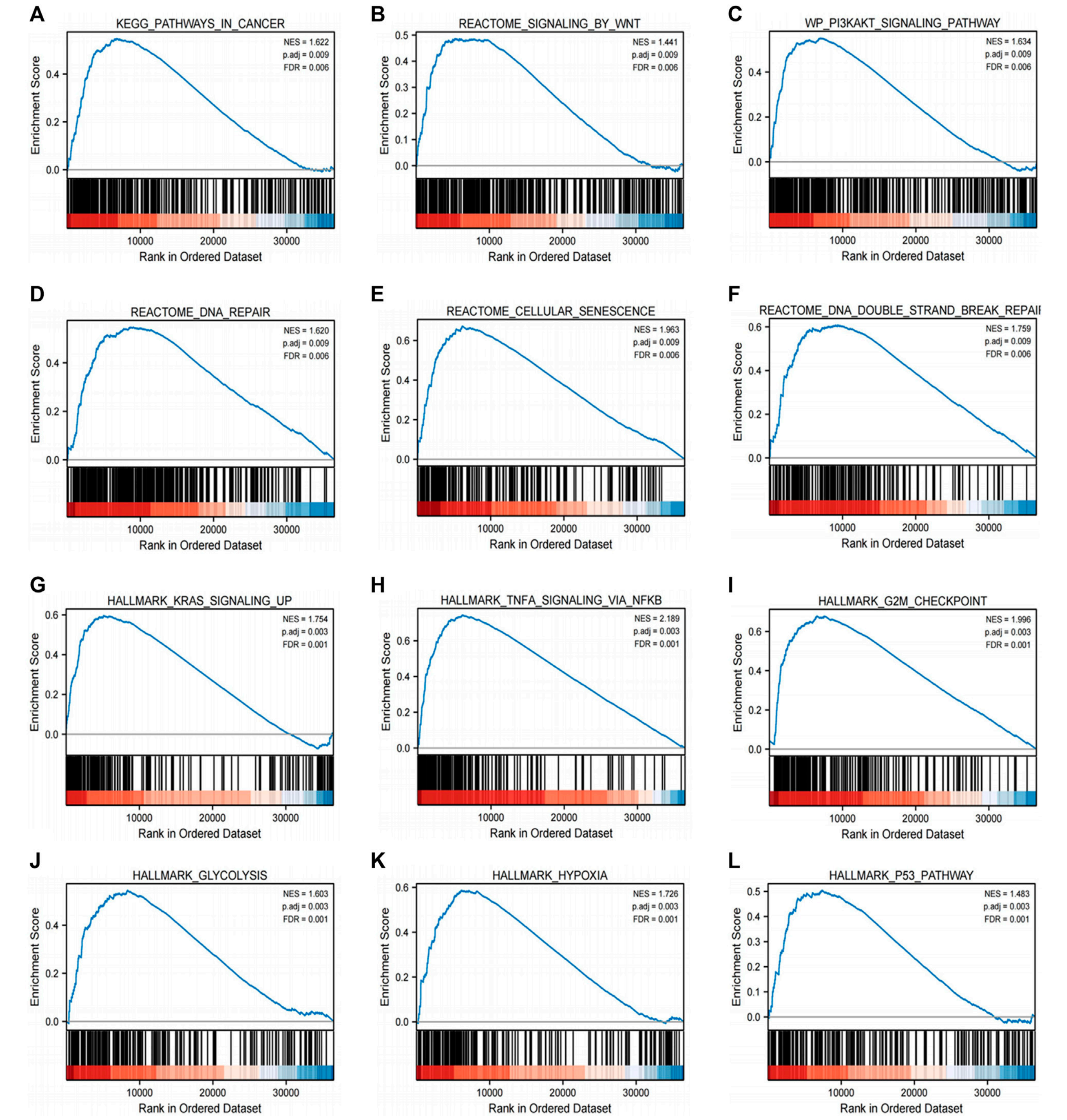


FIGURE 5 | Enrichment plots from GSEA. GSEA results showing pathways in cancer **(A)**, signaling by wnt **(B)**, the PI3K-AKT signaling pathway **(C)**, DNA repair **(D)**, cellular senescence **(E)**, DNA double-strand break repair **(F)**, KRAS signaling up **(G)**, TNFA signaling via NFKB **(H)**, G2M checkpoint **(I)**, glycolysis **(J)**, hypoxia **(K)**, and the p53 pathway **(L)**, which are differentially enriched in H2BC12-high expression phenotype. NES, normalized ES; p. adj, p. adjust; FDR, False Discovery Rate.

which will be of vital significance for the treatment and prognosis of patients with GII and GIII gliomas.

In this study, we firstly obtained RNA-seq data documented in TCGA and matched normal samples from GTEx in the UCSC XENA database, demonstrating that H2BC12 mRNA significantly increased in tumor tissue compared to normal control. A similar trend was observed between GII and GIII and was also confirmed by the clinical specimen. These suggested that H2BC12 might be active in promoting glioma initiation. H2BC12 encoded a replication-dependent histone that was a member of the histone H2B family.

TABLE 4 | Gene sets enriched in positively correlated with H2BC12 mRNA expression phenotype high.

MSigDB collection	Gene set name	NES	p.adj	FDR
c2.cp.v7.2.symbols.gmt [Curated]	KEGG_PATHWAYS_IN_CANCER	1.622	0.009	0.006
	REACTOME_SIGNALING_BY_WNT	1.441	0.009	0.006
	WP_PI3KAKT_SIGNALING_PATHWAY	1.634	0.009	0.006
	REACTOME_DNA_REPAIR	1.620	0.009	0.006
	REACTOME_CELLULAR_SENESCENCE	1.963	0.009	0.006
	REACTOME_DNA_DOUBLE_STRAND_BREAK_REPAIR	1.759	0.009	0.006
h.all.v7.2.symbols.gmt [Hallmarks]	HALLMARK_KRAS_SIGNALING_UP	1.754	0.003	0.001
	HALLMARK_TNFA_SIGNALING_VIA_NFKB	2.189	0.003	0.001
	HALLMARK_G2M_CHECKPOINT	1.996	0.003	0.001
	HALLMARK_GLYCOLYSIS	1.603	0.003	0.001
	HALLMARK_HYPOXIA	1.726	0.003	0.001
	HALLMARK_P53_PATHWAY	1.483	0.003	0.001

NES, normalized enrichment score; p.adj, adjust p *value; FDR, false discovery rate.*

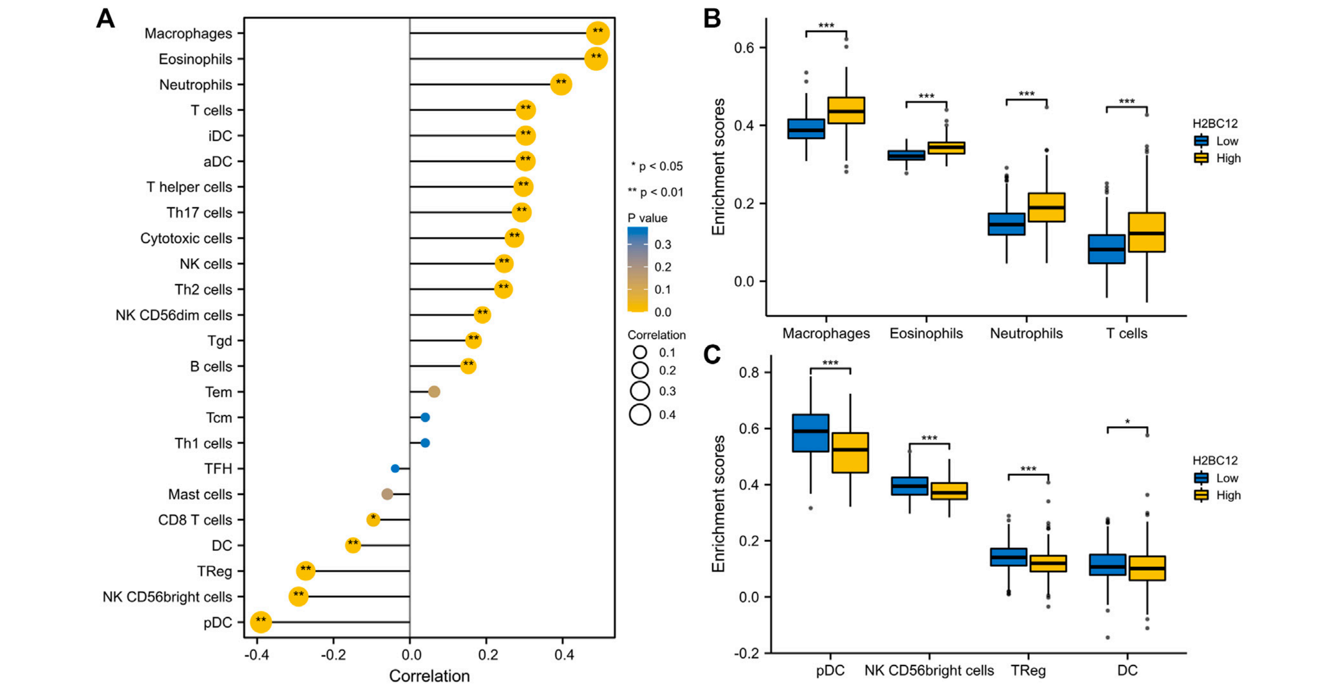


FIGURE 6 | Correlation analysis between H2BC12 and immune infiltration. **(A)** Association analysis between H2BC12 expression and immune cells. **(B, C)** Differences in immune cell infiltration levels between high and low H2BC12 expression groups.

H2B played a crucial role in chromatin-related processes involved in transcription, DNA replication, and repair. Kim et al. (Kim et al., 2012) reported the top six most highly expressed genes in breast cancer, including STAT3, CTSD, SREBF1, IGFBP5, and DDR1, from 49 signature genes of tumor dormancy based on cancer cell line data and microarray data, which further verified the role of H2BC12 as a potential tumor dormancy marker (Kim et al., 2012). Dormant cells are highly adaptable in chemotherapy since they can rapidly target proliferating cells. Meanwhile, they can still survive for a long time and even reproduce after chemotherapy is terminated. Han et al. (Han et al., 2019) reported that H2BC12 displayed increased expression in drug-resistant cell MDA-MB-231 in breast cancer, showing a close relationship between the H2BC12 and drug resistance. Here, we found that H2BC12 mRNA presented with high expression in gliomas compared with normal tissues, and its expression in GIII was also higher than in GII. This implied that H2BC12 might be a therapeutic target or biomarker and that it is involved in promoting glioma progression.

Research revealed that H2A and H2B are important participants in chromatin transcription, DNA replication, and repair (Li et al., 2017). Similarly, we noted the good diagnostic performance of H2BC12 for GII and GIII, characterized by an AUC of 0.823. Meanwhile, H2BC12 could distinguish GIII gliomas from GII gliomas with 76.9% sensitivity and 81.8%

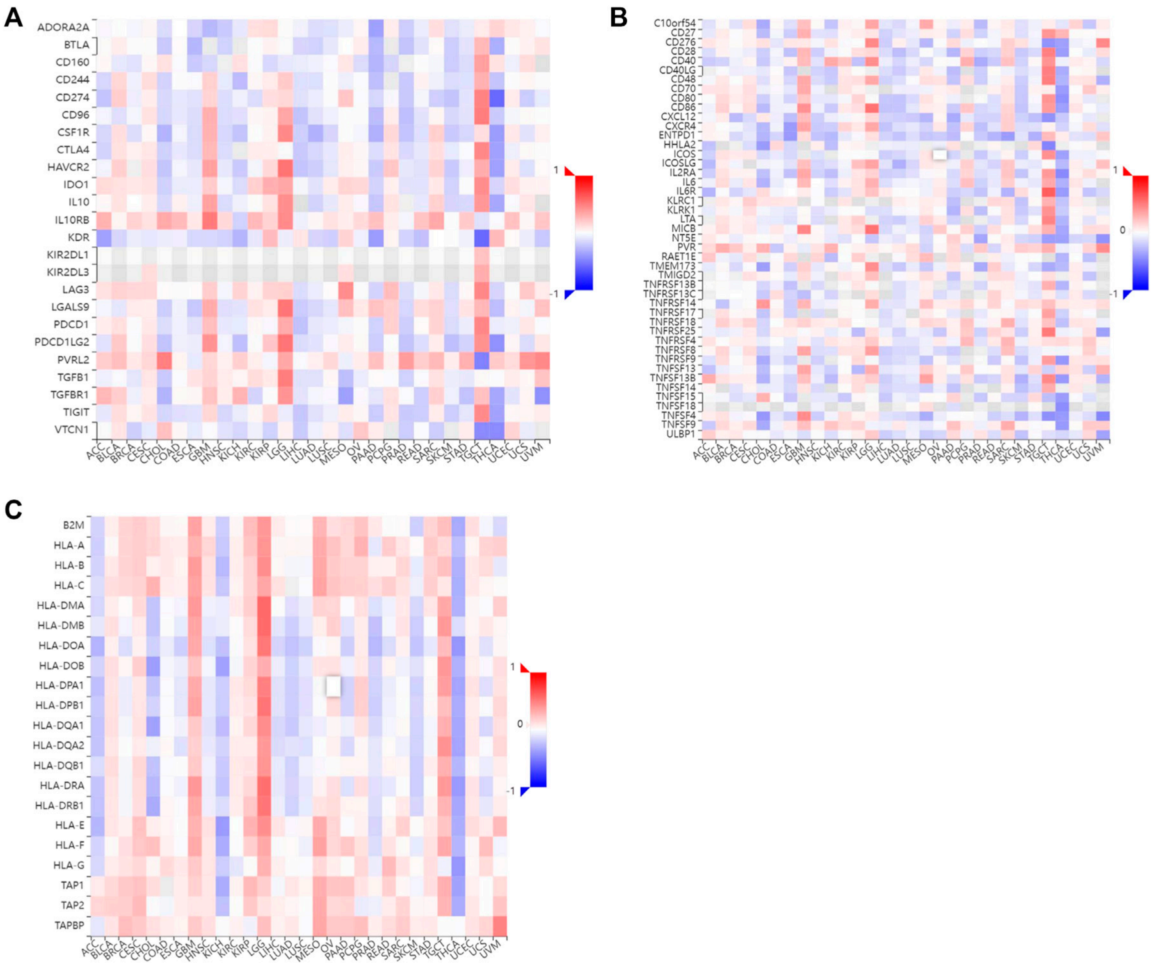


FIGURE 7 | The relationship between H2BC12 and immune regulatory factor. The level of H2BC12 mRNA was positively correlated with immune inhibitors **(A)**, immune stimulators **(B),** and MHC molecules **(C)**. Red indicated a significant positive correlation, and blue indicated a significant negative correlation.

specificity, which might improve the diagnosis and therapy of gliomas. We then profiled the association between H2BC12 and clinicopathologic characteristics of gliomas. Notably, the increased H2BC12 was correlated significantly with IDH status, 1p/19q codeletion, primary therapy outcome, and histological type. This demonstrated that H2BC12 mRNA is closely related to the clinicopathologic characteristics of gliomas, and H2BC12 might be involved in disease progression.

The tROC curve also validated the moderate prognostic value of H2BC12 for OS of GII and/or GIII in 1, 2, and 3 years. This indicated that H2BC12 might predict the survival outcome of gliomas, which was consistent with a previous study that showed that signatures based on histone gene family are potentially good indicators for the outcome of cervical cancer patients (Li et al., 2017). It was worth noting that the AUC was different between GII and GIII. In GII, the AUC of 3 years was more than that of 2 years and then 1 year. However, the opposite trend was observed in GIII, and the AUC of 1 year was better than those of 2 or 3 years. This gave us a hint that H2BC12 had a different value for predicting survival outcomes in patients with GII and GIII. However, its predictive power was different for different years in GII and GIII, indicating H2BC12 might play an important role in gliomas progression. No previous studies have reported a link between H2BC12 and gliomas. Further survival analysis was conducted to validate the association of high H2BC12 expression with adverse survival outcomes of GII and GIII patients. Interestingly, the higher H2BC12, the shorter OS, DSS, and PFI of GII patients. A similar trend was also observed in GIII patients. We thus believe that H2BC12 serves as a high-risk factor for GII and GIII. Previous bioinformatics analysis identified that high H2BC12 predicted adverse outcomes of breast, pancreatic, and ovarian cancers (Li et al., 2018; Li and Zhan, 2019; Yu et al., 2020). We next performed univariate and multivariate analyses to identify factors predicting OS with Cox regression models. Results showed that H2BC12, WHO grade, IDH status, age, and histological type could all be prognostic factors for gliomas. Given this, a further clinical stratification analysis was designed to identify whether H2BC12 was an independent predictor.

Finally, we conducted GSEA to uncover the H2BC12-related pathways in gliomas. Results showed that there were six pathways, including pathways in cancer, the Wnt or PI3K-AKT signaling pathway, DNA repair, cellular senescence, and DNA double-strand break repair, which demonstrated differential enrichment in higher H2BC12. Research reveals that activated PI3K-AKT could facilitate the invasiveness of glioma cells (Li et al., 2019). DNA repair genes are associated with gliomas (Tang et al., 2018). DNA repair damage is the main cause of radio-resistance and chemo-resistance in gliomas (Zeng et al., 2019). A study suspected that targeting an H2Bub1 that regulates both transcription and DNA damage repair may inhibit an oncogenic transcriptional expression profile while simultaneously impairing the ability of the cell to effectively repair DNA damage, thereby increasing its sensitivity to a second drug that induces DNA damage (Jeusset and McManus, 2021). It has also been found that RNF20 (and RNF40) expression is increased in luminal B tumors, and er-positive tumors with high H2Bub1 abundance have poorer survival (Tarcic et al., 2017). All these findings indicate the potential important role of H2BC12 in gliomas progression. Moreover, as a new therapeutic strategy, immunotherapy, has drawn the attention of the field of gliomas. However, only a minority of glioma patients got responses due to a lacking of effective biomarkers (Chiocca et al., 2019). The current results showed that H2BC12 had a positive correlation to immune cells, including macrophages, NK cells, Treg, and T cells. These findings gave us a hint that H2BC12 might be involved in the immunoregulation of gliomas, which was consistent with a previous study that DNAJC10 was correlated with immune cell infiltrations and immune checkpoint genes (Liu et al., 2022) as well as the replication factor C2 (Zhao et al., 2022). Furthermore, our results also showed that H2BC12 was positively associated with immune regulatory factors, including immune inhibitor PDCD1LG2, immune stimulator CD40, and MHC molecule HLA-DMA. H2BC12 could be a potential prognostic marker and immunotherapy marker in gliomas.

In all, this study verified the significance of H2BC12 in the diagnosis and prognosis of GII and GIII gliomas. Inevitably, limitations still exist. First, the study was carried out only with bioinformatics analysis, requiring further validation in clinical samples. Second, there is a need to clarify the H2BC12-mechanism of action.

CONCLUSION

This study identified the differentially up-regulated expression of H2BC12 in GII and GIII glioma tissue and proved its significant ability in predicting the adverse overall survival of GII and GIII gliomas patients. H2BC12, therefore, has promising application for the diagnosis and prognosis of gliomas.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by JZ, ZX, YX and ZD. The first draft of the manuscript was written by JZ, ML, XL, DW and ZD and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Taishan Scholar Project of Shandong Province of China (Grant no. tsqn202103200), the Latitudinal Research Project of Shandong University (Grant no. 6010120011), and the Shandong traditional Chinese medicine science and technology project (Grant no. 2021M033).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.816939/full#supplementary-material

Supplementary Figure S1 | The expression and clinical value of H2BC12 mRNA in GII and GIII gliomas tissues. H2BC12 showed significantly higher expression in GIII tissue than in GII tissue **(A)**, p<0.05 and was related to IDH status **(B)**, p<0.05. **(C)** ROC analysis of H2BC12 in classification power for GII and GIII (AUC=0.706).

Supplementary Figure S2 | H2BC12 promoter methylation level and mutation analysis. **(A–D)** The level of DNA methylation in the H2BC12 promoter region in patients with TP53 mutant status, age, gender, and race. **(E)** H2BC12 mutation in most brain tumors. ns: P>0.05; *P<0.05.

REFERENCES

Beppu, T., Sasaki, M., Kudo, K., Kurose, A., Takeda, M., Kashimura, H., et al. (2011). Prediction of Malignancy Grading Using Computed Tomography Perfusion Imaging in Nonenhancing Supratentorial Gliomas. *J. Neurooncol.* 103 (3), 619–627. doi:10.1007/s11060-010-0433-0

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* 39 (4), 782–795. doi:10.1016/j.immuni.2013.10.003

Cancer Genome Atlas Research NetworkBrat, D. J., Verhaak, R. G., Aldape, K. D., Yung, W. K., Salama, S. R., et al. (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N. Engl. J. Med.* 372 (26), 2481–2498. doi:10.1056/NEJMoa1402121

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B. V. S. K., et al. (2017). UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* 19 (8), 649–658. doi:10.1016/j.neo.2017.05.002

Chen, S., Li, J., Wang, D.-L., and Sun, F.-L. (2012). Histone H2B Lysine 120 Monoubiquitination Is Required for Embryonic Stem Cell Differentiation. *Cell Res* 22 (9), 1402–1405. doi:10.1038/cr.2012.114

Chiocca, E. A., Nassiri, F., Wang, J., Peruzzi, P., and Zadeh, G. (2019). Viral and Other Therapies for Recurrent Glioblastoma: Is a 24-month Durable Response Unusual? *Neuro Oncol.* 21 (1), 14–25. doi:10.1093/neuonc/noy170

Finotello, F., and Trajanoski, Z. (2018). Quantifying Tumor-Infiltrating Immune Cells from Transcriptomics Data. *Cancer Immunol. Immunother.* 67 (7), 1031–1040. doi:10.1007/s00262-018-2150-z

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* 6 (269), pl1–28. doi:10.1126/scisignal.2004088

Han, J., Lim, W., You, D., Jeong, Y., Kim, S., Lee, J. E., et al. (2019). Chemoresistance in the Human Triple-Negative Breast Cancer Cell Line MDA-MB-231 Induced by Doxorubicin Gradient Is Associated with Epigenetic Alterations in Histone Deacetylase. *J. Oncol.* 2019, 1345026. doi:10.1155/2019/1345026

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Huang, J., Samson, P., Perkins, S. M., Ansstas, G., Chheda, M. G., DeWees, T. A., et al. (2017). Impact of Concurrent Chemotherapy with Radiation Therapy for Elderly Patients with Newly Diagnosed Glioblastoma: A Review of the National Cancer Data Base. *J. Neurooncol.* 131 (3), 593–601. doi:10.1007/s11060-016-2331-6

Jeusset, L. M., and McManus, K. J. (2021). Characterizing and Exploiting the Many Roles of Aberrant H2B Monoubiquitination in Cancer Pathogenesis. *Semin. Cancer Biol.* S1044-579X (21), 00300–X. doi:10.1016/j.semcancer.2021.12.007

Kari, V., Shchebet, A., Neumann, H., and Johnsen, S. A. (2011). The H2B Ubiquitin Ligase RNF40 Cooperates with SUPT16H to Induce Dynamic Changes in Chromatin Structure during DNA Double-Strand Break Repair. *Cell Cycle* 10 (20), 3495–3504. doi:10.4161/cc.10.20.17769

Kim, R. S., Avivar-Valderas, A., Estrada, Y., Bragado, P., Sosa, M. S., Aguirre-Ghiso, J. A., et al. (2012). Dormancy Signatures and Metastasis in Estrogen Receptor Positive and Negative Breast Cancer. *PLoS One* 7 (4), e35569. doi:10.1371/journal.pone.0035569

Kiran, M., Chatrath, A., Tang, X., Keenan, D. M., and Dutta, A. (2019). A Prognostic Signature for Lower Grade Gliomas Based on Expression of Long Non-Coding RNAs. *Mol. Neurobiol.* 56 (7), 4786–4798. doi:10.1007/s12035-018-1416-y

Li, C., Luo, L., Wei, S., and Wang, X. (2018). Identification of the Potential Crucial Genes in Invasive Ductal Carcinoma Using Bioinformatics Analysis. *Oncotarget* 9 (6), 6800–6813. doi:10.18632/oncotarget.23239

Li, J., Xu, H., Wang, Q., Wang, S., and Xiong, N. (2019). 14-3-3ζ Promotes Gliomas Cells Invasion by Regulating Snail through the PI3K/AKT Signaling. *Cancer Med.* 8 (2), 783–794. doi:10.1002/cam4.1950

Li, N., and Zhan, X. (2019). Signaling Pathway Network Alterations in Human Ovarian Cancers Identified with Quantitative Mitochondrial Proteomics. *EPMA J.* 10 (2), 153–172. doi:10.1007/s13167-019-00170-5

Li, W., Huang, W., Wu, K., and Long, Y. (2022). Yippee Like 1 Suppresses Glioma Progression and Serves as a Novel Prognostic Factor. *Tohoku J. Exp. Med.* 256 (2), 141–150. doi:10.1620/tjem.256.141

Li, X., Tian, R., Gao, H., Yang, Y., Williams, B. R. G., Gantier, M. P., et al. (2017). Identification of a Histone Family Gene Signature for Predicting the Prognosis of Cervical Cancer Patients. *Sci. Rep.* 7 (1), 16495. doi:10.1038/s41598-017-16472-5

Liu, F., Tu, Z., Liu, J., Long, X., Xiao, B., Fang, H., et al. (2022). DNAJC10 Correlates with Tumor Immune Characteristics and Predicts the Prognosis of Glioma Patients. *Biosci. Rep.* 42 (1), BSR20212378. doi:10.1042/BSR20212378

Liu, Z., Ren, Z., Zhang, C., Qian, R., Wang, H., Wang, J., et al. (2021). ELK3: A New Molecular Marker for the Diagnosis and Prognosis of Glioma. *Front. Oncol.* 11, 608748. doi:10.3389/fonc.2021.608748

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 (6), 803–820. doi:10.1007/s00401-016-1545-1

Moyal, L., Lerenthal, Y., Gana-Weisz, M., Mass, G., So, S., Wang, S.-Y., et al. (2011). Requirement of ATM-dependent Monoubiquitylation of Histone H2B for Timely Repair of DNA Double-Strand Breaks. *Mol. Cel* 41 (5), 529–542. doi:10.1016/j.molcel.2011.02.015

Müller, S., and Almouzni, G. (2017). Chromatin Dynamics during the Cell Cycle at Centromeres. *Nat. Rev. Genet.* 18 (3), 192–208. doi:10.1038/nrg.2016.157

Perez, A., and Huse, J. T. (2021). The Evolving Classification of Diffuse Gliomas: World Health Organization Updates for 2021. *Curr. Neurol. Neurosci. Rep.* 21 (12), 67–77. doi:10.1007/s11910-021-01153-8

Sadeghi, L., Siggens, L., Svensson, J. P., and Ekwall, K. (2014). Centromeric Histone H2B Monoubiquitination Promotes Noncoding Transcription and Chromatin Integrity. *Nat. Struct. Mol. Biol.* 21 (3), 236–243. doi:10.1038/nsmb.2776

Sansó, M., Lee, K. M., Viladevall, L., Jacques, P.-É., Pagé, V., Nagy, S., et al. (2012). A Positive Feedback Loop Links Opposing Functions of P-TEFb/Cdk9 and Histone H2B Ubiquitylation to Regulate Transcript Elongation in Fission Yeast. *Plos Genet.* 8 (8), e1002822. doi:10.1371/journal.pgen.1002822

Sethi, G., Shanmugam, M. K., Arfuso, F., and Kumar, A. P. (2018). Role of RNF20 in Cancer Development and Progression - a Comprehensive Review. *Biosci. Rep.* 38 (4), BSR20171287. doi:10.1042/BSR20171287

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J. B., et al. (2005). Radiotherapy Plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *N. Engl. J. Med.* 352 (10), 987–996. doi:10.1056/nejmoa043330

Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., et al. (2015). Mutational Landscape and Clonal Architecture in Grade II and III Gliomas. *Nat. Genet.* 47 (5), 458–468. doi:10.1038/ng.3273

Tang, L., Deng, L., Bai, H. X., Sun, J., Neale, N., Wu, J., et al. (2018). Reduced Expression of DNA Repair Genes and Chemosensitivity in 1p19q Codeleted Lower-Grade Gliomas. *J. Neurooncol.* 139 (3), 563–571. doi:10.1007/s11060-018-2915-4

Tarcic, O., Granit, R. Z., Pateras, I. S., Masury, H., Maly, B., Zwang, Y., et al. (2017). RNF20 and Histone H2B Ubiquitylation Exert Opposing Effects in Basal-Like versus Luminal Breast Cancer. *Cell Death Differ* 24 (4), 694–704. doi:10.1038/cdd.2016.126

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., et al. (2017). Toil Enables Reproducible, Open Source, Big Biomedical Data Analyses. *Nat. Biotechnol.* 35 (4), 314–316. doi:10.1038/nbt.3772

Wick, W., Weller, M., van den Bent, M., Sanson, M., Weiler, M., von Deimling, A., et al. (2014). MGMT Testing-The Challenges for Biomarker-Based Glioma Treatment. *Nat. Rev. Neurol.* 10 (7), 372–385. doi:10.1038/nrneurol.2014.100

Xiao, Y., Zhu, Z., Li, J., Yao, J., Jiang, H., Ran, R., et al. (2020). Expression and Prognostic Value of Long Non-Coding RNA H19 in Glioma via Integrated Bioinformatics Analyses. *Aging* 12 (4), 3407–3430. doi:10.18632/aging.102819

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 (5), 284–287. doi:10.1089/omi.2011.0118

Yu, S., Li, Y., Liao, Z., Wang, Z., Wang, Z., Li, Y., et al. (2020). Plasma Extracellular Vesicle Long RNA Profiling Identifies a Diagnostic Signature for the Detection of Pancreatic Ductal Adenocarcinoma. *Gut* 69 (3), 540–550. doi:10.1136/gutjnl-2019-318860

Zeng, F., Liu, X., Wang, K., Zhao, Z., and Li, G. (2019). Transcriptomic Profiling Identifies a DNA Repair-Related Signature as a Novel Prognostic Marker in Lower Grade Gliomas. *Cancer Epidemiol. Biomarkers Prev.* 28 (12), 2079–2086. doi:10.1158/1055-9965.epi-19-0740

Zhang, K., Yang, L., Wang, J., Sun, T., Guo, Y., Nelson, R., et al. (2019). Ubiquitin-Specific Protease 22 Is Critical to *In Vivo* Angiogenesis, Growth and Metastasis of Non-Small Cell Lung Cancer. *Cell Commun Signal* 17 (1), 167–192. doi:10.1186/s12964-019-0480-x

Zhao, J., Liu, Z., Zheng, X., Gao, H., and Li, L. (2021). Prognostic Model and Nomogram Construction Based on a Novel Ferroptosis-Related Gene Signature in Lower-Grade Glioma. *Front. Genet.* 12, 753680. doi:10.3389/fgene.2021.753680

Zhao, X., Wang, Y., Li, J., Qu, F., Fu, X., Liu, S., et al. (2022). RFC2: A Prognosis Biomarker Correlated with the Immune Signature in Diffuse Lower-Grade Gliomas. *Sci. Rep.* 12 (1), 3122–3145. doi:10.1038/s41598-022-06197-5

Zhou, S., Cai, Y., Liu, X., Jin, L., Wang, X., Ma, W., et al. (2021). Role of H2B Mono-Ubiquitination in the Initiation and Progression of Cancer. *Bull. du Cancer* 108 (4), 385–398. doi:10.1016/j.bulcan.2020.12.007

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 | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 25 May 2022
doi: 10.3389/fmolb.2022.878073

Identification of Four Novel Prognostic Biomarkers and Construction of Two Nomograms in Adrenocortical Carcinoma: A Multi-Omics Data Study *via* Bioinformatics and Machine Learning Methods

Xiaochun Yi, Yueming Wan, Weiwei Cao, Keliang Peng*, Xin Li and Wangchun Liao

Department of Urology, Yueyang People's Hospital, Hunan Normal University, Yueyang, China

Background: Adrenocortical carcinoma (ACC) is an orphan tumor which has poor prognoses. Therefore, it is of urgent need for us to find candidate prognostic biomarkers and provide clinicians with an accurate method for survival prediction of ACC via bioinformatics and machine learning methods.

Methods: Eight different methods including differentially expressed gene (DEG) analysis, weighted correlation network analysis (WGCNA), protein-protein interaction (PPI) network construction, survival analysis, expression level comparison, receiver operating characteristic (ROC) analysis, and decision curve analysis (DCA) were used to identify potential prognostic biomarkers for ACC via seven independent datasets. Linear discriminant analysis (LDA), K-nearest neighbor (KNN), support vector machine (SVM), and time-dependent ROC were performed to further identify meaningful prognostic biomarkers (MPBs). Cox regression analyses were performed to screen factors for nomogram construction.

Results: We identified nine hub genes correlated to prognosis of patients with ACC. Furthermore, four MPBs (ASPM, BIRC5, CCNB2, and CDK1) with high accuracy of survival prediction were screened out, which were enriched in the cell cycle. We also found that mutations and copy number variants of these MPBs were associated with overall survival (OS) of ACC patients. Moreover, MPB expressions were associated with immune infiltration level. Two nomograms [OS-nomogram and disease-free survival (DFS)-nomogram] were established, which could provide clinicians with an accurate, quick, and visualized method for survival prediction.

Conclusion: Four novel MPBs were identified and two nomograms were constructed, which might constitute a breakthrough in treatment and prognosis prediction of patients with ACC.

Keywords: adrenocortical carcinoma, WGCNA, hub genes, nomogram, prognosis, immune microenvironment, copy number variations

OPEN ACCESS

Edited by:
Tianyi Zhao,
Harbin Institute of Technology, China

Reviewed by:
Zixin Guo,
Wuhan University, China
Xianfeng Shen,
Hubei University of Medicine, China

****Correspondence:***
Keliang Peng
yy19871818@163.com

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *17 February 2022*
Accepted: *08 April 2022*
Published: *25 May 2022*

Citation:
Yi X, Wan Y, Cao W, Peng K, Li X and Liao W (2022) Identification of Four Novel Prognostic Biomarkers and Construction of Two Nomograms in Adrenocortical Carcinoma: A Multi-Omics Data Study via Bioinformatics and Machine Learning Methods. Front. Mol. Biosci. 9:878073. doi: 10.3389/fmolb.2022.878073

INTRODUCTION

Though adrenocortical carcinoma (ACC) is an uncommon malignancy, the prognosis of patients with this malignancy is poor (Jasim and Habra, 2019). The disease tends to occur in 3 to 4-year-old children and 40 to 50-year-old adults (Libé, 2018). The incidence of ACCs in children is reported to be as low as 0.2% of pediatric cancers (Libé, 2018). As a recent study reported, there were 15,800 new cases of ACC worldwide in 2018 (Bray et al., 2018). However, the incidence of ACCs varies from place to place around the world (Bray et al., 2018). In some countries, such as southern Brazil, the incidence is 10–15 times what it is in America (Bray et al., 2018). But the consensus is that this malignancy heavily endangers health and is very difficult to cure (Lo et al., 2019). According to the previous studies, some researchers tried to diagnose ACC earlier to grasp the optimal treatment opportunity (Lo et al., 2019). What is worse, about 40 percent of ACCs had distant metastasis when they were diagnosed (Guillaume et al., 2014). Nowadays, the discoveries of new small biomarkers greatly aid diagnosis of malignant tumors by using the methods of molecular biology and bioinformatics (He et al., 2017). In order to better diagnosis ACCs and improve the prognosis of patients, the objective of the research is to screen several effective prognostic biomarkers of ACC. Also, we attempted to provide clinicians with several choices for ACC therapy. The CMap analysis demonstrated that five small molecule drugs including chlorpromazine, trifluoperazine, alpha-estradiol, 15-delta prostaglandin J2, and vorinostat might be novel drugs for ACC treatment. These MPBs were also significantly enriched in the cell cycle. As for the enriched drugs, ASPM was significantly enriched in 6 drugs, BIRC5 was associated with 6 drugs, CCNB2 was related to 11 drugs, and CDK1 was enriched in 6 drugs. Moreover, we devoted ourselves to provide clinicians with an accurate, individual, and visualized method to predict overall survival (OS) or disease-free survival (DFS) of patients with ACC. To do this, we thought it could help clinicians to understand and master the illness and better formulate the treatment scheme.

METHODS AND MATERIALS

ACC Microarray Studies Identification

All the GEO datasets were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). For differentially expressed gene (DEG) screening, datasets with related control tissues were collected and used. Then two datasets including GSE75415 (West et al., 2007) and GSE12368 (Soon et al., 2009) were included. Then four datasets including GSE76021 (Pinto et al., 2016), GSE19750 (Demeure et al., 2013), GSE10927 (Giordano et al., 2009), and GSE76019 (Surakhy et al., 2020) from this database were collected and used in the present study, because of the complete clinical and survival information they contained. Moreover, we retrieved microarray data of ACC (TCGA-ACC data) and the related clinical information via The Cancer Genome Atlas (TCGA) database (https://genome-cancer. ucsc.edu/). All in all, GSE12368 and GSE75415 were included for DEG identification because they included normal tissues. GSE76021, GSE19750, GSE10927, GSE76019, and TCGA-ACC data were included because they had related clinical information (stage, grade, etc.) and survival information. The details of all the datasets are shown in **Supplementary Table S1**.

Data Preprocessing and DEG Identification

For the TCGA data, we firstly downloaded RNA sequencing data (FPKM value) of gene expressions from the TCGA database using R package "TCGAbiolinks" (Colaprico et al., 2016). In order to compare and validate the results with GEO datasets, these data were further transformed into a transcripts per kilobase million (TPM) profile. For the datasets from the GEO database, the robust multichip average algorithm (Irizarry et al., 2003) was used because the data were displayed as RAW series. Moreover, log2 transformation and normalization were conducted based on R package "affy" (Gautier et al., 2004).

How we validated the results among this study is shown in **Supplementary Figure S1**. A total of 29 ACCs included in GSE76021 were used for WGCNA. We sorted genes according to their variance across all samples, all genes were selected for WGCNA. Moreover, differentially expressed genes (DEGs) between ACCs and normal tissues were filtered out by the criterion (p value < 0.05, |log2 fold change (FC) | ≥ 1.5) via R package "limma" (Ritchie et al., 2015) for further study. Then DEGs overlapped between GSE75415 and GSE12368 were screened for subsequent analysis.

Co-Expression Network Construction

Before conducting WGCNA, the expression matrix of the transcript level was checked via two approaches (goodSamplesGenes and sample network methods) in R package "WGCNA" (Zhang and Horvath, 2005). Only samples of Z.Ku ≥ -2.5 were included for co-expression network construction. By means of the scale free topology criterion, β (soft threshold power beta) was chosen. We subsequently transformed adjacency into TOM. Then based on the TOM, genes were classified into modules via the branch cutting approach. Some important parameters set in the present study were shown as below: minClusterSize = 30, deepSplit = 2. In addition, by selecting a cut line reckoned dissimilarity of module eigengenes (MEs), modules showing high correlation with each other were merged.

Survival-Associated Module Identification

After determining modules composed of genes, two methods were applied on screening hub modules which were relevant to survival status (the aimed clinical trait). The correlation between module eigengenes and traits were quantized. Next, through evaluating gene significance (GS), the relationship between genes and traits was measured. In addition, the average GS of all the genes in a module was further worked out, which represented the module significance (MS). After finishing the above analyses, we identified the most related module as the key module.

Connectivity Map Analysis

As a convenient webtool, researchers can quickly locate molecule drugs which have potential against related diseases through CMap (https://portals.broadinstitute.org/cmap/) (Lamb et al., 2006). Therefore, CMap analysis was conducted via the screened DEGs, in order to explore potential drugs showing a strong relationship with ACC. Drugs meeting the requirement [number of instances (n) > 10, p value < 0.05] were considered significant. Furthermore, drugs with |mean| ≥ 0.40 were further screened out, which might be useful choices for treating ACC.

Candidate Hub Gene Construction

After choosing the key module, genes of |cor.geneModuleMembership| >0.8 and |cor.geneTraitSignificance| >0.2 were regarded as hub genes in WGCNA. Then we constructed a protein-protein interaction (PPI) network of these genes via the Search Tool for the Retrieval of Interacting Genes (STRING) (Szklarczyk et al., 2015). The following parameters were important and listed: network scoring: degree cutoff = 2; cluster finding: node score cutoff = 0.2, k-core = 2, and max. depth = 100. A vehicle named network analyzer in Cytoscape (Shannon et al., 2003) was used for the gene degree of connectivity calculation. In this research, we regarded a gene as a hub gene in the PPI network when its degree ≥4. We also constructed a PPI network for DEGs to screen hub genes in DEGs by using the same standard. Finally, genes overlapping between hub genes in WGCNA and hub genes in DEGs were considered as candidate hub genes, which were included for further analysis. Gene ontology (GO) (Ashburner et al., 2000) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) pathway analysis were conducted via R package “clusterProfiler” (Yu et al., 2012) for functional annotation of candidate hub genes. We selected p < 0.05 as the standards to define significant BPs and KEGG pathway terms.

Hub Gene Identification

Hub genes related to survival and prognosis of ACC patients were screened through performing survival analysis among candidate hub genes based on R package “survival” (Therneau, 2015) for datasets with complete survival information (GSE19750, GSE76019, GSE76021, and TCGA-ACC data). For TCGA-ACC data, 79 samples with complete overall survival (OS) information were included for OS analysis, meanwhile 54 samples with complete disease-free survival (DFS) information were included for DFS analysis. According to the candidate hub gene expression levels, we split samples into two groups (high expression group and low expression group) in all the datasets (the median expression of each candidate hub gene in each dataset was set as the grouping standard). Genes of p < 0.05 in all survival analyses were considered as hub genes.

Hub Gene Validation

Based on datasets with complete stage information (GSE10927, GSE19750, GSE75415, GSE76019, GSE76021 and TCGA-ACC data), we plotted tumor stage (I, II, III and IV) boxplots using the “ggstatsplot” (Patil, 2018) R package. Moreover, tumor grade boxplots were also plotted based on GSE10927 (low grade and high grade) and GSE19750 (grade 1, grade 2, grade 3, and grade 4). A one-way analysis of variance (ANOVA) test was conducted to evaluate the results when samples were divided into more than two groups. We used unpaired t test to measure the statistical significance when samples were divided into two groups. Moreover, the difference of hub gene expression values in ACCs, ACAs, and normal adrenal samples were measured using GSE10927, GSE12368, GSE19750, GSE75415, and TCGA-BLCA data.

Receiver Operating Characteristic Analysis and Decision Curve Analysis

Through R package “plotROC” (Sachs, 2017), ROC curve analysis was performed. In GSE10927, GSE12368, GSE19750, and GSE75415, the AUC was calculated to differentiate ACC samples and normal tissues. In GSE10927, GSE19750, GSE75415, GSE76019, GSE76021, and TCGA-ACC data, we worked out the AUC to distinguish localized ACC and advanced ACC. In this study, we regarded ACC of stages I or II as localized ACC and ACC of stages III or IV as advanced ACC. In both GSE10927 and GSE19750, we worked out the AUC to distinguish ACC of low grade (grades 1 or 2) and ACC of high grade (grades 3 or 4). Moreover, we distinguished ACA and ACC in GSE10927, GSE12368, and GSE75415. In this study, we thought genes could distinguish ACC samples from normal tissues (localized ACC from advanced ACC or low grade ACC from high grade ACC) well when the AUC was more than 0.70. Furthermore, DCA (Vickers and Elkin, 2006) was performed for verifying the hub genes’ diagnostic potential by using GSE76021.

Linear Discriminant Analysis, K-Nearest Neighbor, and Support Vector Machine to Screen Genes With High Accuracy of Predicting OS Among Hub Genes

To validate hub genes’ prognostic potential, genes were taken as variables, relative mRNA expression values of which were taken as variable values. LDA, KNN, and SVM analyses were immediately conducted. LDA was conducted via R package “MASS” (Venables and Ripley, 2002). The cross validation approach was used to pick out the best K parameter via R package “caret” (Kuhn, 2015). Based on the best K parameter, R packages “class” (Venables and Ripley, 2002) and “kknn” were used for the KNN method. In addition, we performed four types of SVM methods via R package “e1071”. They were linear-SVM, polynomial-SVM, radial basis function (RBF) SVM, and sigmoid-kernel SVM, separately. The SVM factors setting was based on “kernlab” in R software. TCGA-ACC data were included in this part. We regarded a gene as a meaningful prognostic biomarker (MPB) with the average accuracy of classification in three analyses ≥0.80.

Time-Dependent ROC Analysis for MPBs

To verify the potential of the prognosis prediction of MPBs, based on TCGA-ACC data, time-independent (1-, 3-, 5-years) receiver operating characteristic (ROC) analysis was conducted via the "timeROC" (Heagerty et al., 2000) package. The AUC was worked out, we considered that MPBs showed good performance for prognosis prediction when the AUC was more than 0.70 (the same as we set in ROC analysis before).

MPB Mutations and Copy Number Variations

With the aim of screening out mutations and CNVs of genes with high accuracy of predicting OS, all the ACCs and their CNV data from the TCGA database were obtained. The genetic alterations of these genes were screened via the CBio Cancer Genomics Portal (http://www.cbioportal.org/). The correlation between CNVs and relative MPB expression was subsequently identified. The results were measured by ANOVA or Kruskal–Wallis methods. In addition, the relationship between mutations or CNVs of prognostic biomarkers and ACC patients' survival was screened via survival analysis.

Functional Exploration of MPBs

Gene set enrichment analysis (GSEA) might help researchers to comprehend the role of genes in biological behaviors. Therefore, we conducted GSEA for MPBs. A total of 79 ACCs were divided into a high-expression group ($n = 39$) and low-expression group ($n = 40$) according to the prognostic biomarkers' expression median. "c2.cp.kegg.v7.0.symbols.gmt" was chosen as the annotated gene set. We thought a biological pathway of nominal $p < 0.05$, $|ES| > 0.6$, gene size (n) ≥100, and FDR <25% to be significant. In addition, "DSigDBv1.0.gmt" was downloaded from the Drug SIGnatures DataBase (Yoo et al., 2015) (http://tanlab.ucdenver.edu/DSigDB/DSigDBv1.0/download.html) to explore drugs highly associated with prognostic biomarkers. Also, we set the same cut-off criteria as KEGG pathways identification.

Exploring the Relationship Between MPBs and Immune Microenvironment

In this part, the association between MPBs and immunocytes was explored via TIMER (Li et al., 2017) (https://cistrome.shinyapps.io/timer/). We thought an MPB with |correlation coefficient (cor) | ≥0.2 and p value < 0.05 strongly related to an immune cell infiltrating level as previously found. Furthermore, we explored MPB expressions in 33 different cancer types by using the gene module in TIMER.

Exploring the Difference of Immune Infiltration Levels Between a Low Expression and High Expression of MPBs

Based on TCGA-ACC data, ESTIMATE scores, immune scores, and stromal scores were firstly evaluated via applying the ESTIMATE algorithm based on R package "estimate" (Yoshihara et al., 2013). Then we divided ACCs into a high-(ESTIMATE, immune, stromal) score group and low-(ESTIMATE, immune, stromal) score group to perform survival analysis via R package "survival". Moreover, we conducted an unpaired t test to test the difference of score levels between a low expression and high expression of MPBs.

Cox Proportional Hazards Regression Analysis

With the aim of the prognostic value of MPB validation, MPBs and other essential clinical features (gender, age, stage, and laterality) from TCGA-ACC data were selected for OS and DFS univariable Cox analysis. A factor of p value < 0.05 was identified and further selected to conduct multivariate Cox analysis. This analysis could determine whether an MPB was independent from the rest of the clinical factors for predicting OS or DFS of ACCs.

Nomogram Construction

Moreover, with the aim of exploring a simple, quick, and visualized method to predict the possibility of OS or DFS of patients with ACC, two nomograms were constructed (one for OS, the other for DFS) via TCGA-ACC data by using package "rms" (Yizhou et al., 2013). Factors showed meaningful p value in Cox regression analysis (including MPBs and clinical features). Calibrate curves were drawn to test the nomogram, the 45° line was defined as the best prediction. In addition, we evaluated the consistency index (C-index) between the actual probability and predicted probability to further measure the prediction effectiveness of the nomogram (Michael and Ralph, 2010). With the aim of avoiding the over-fitting problem, we conducted cross-validation before nomogram construction. Two datasets (GSE10927 and GSE19750) including their OS information were obtained for external verification of the OS-nomogram by calculating C-index and AUC. Meanwhile, GSE76019 and GSE76021 with integral DFS information were included for DFS-nomogram verification.

RESULTS

DEG Screening

By using the "limma" package in R, we screened 511 DEGs in GSE75415 and 724 DEGs in GSE12368, separately. As shown in **Figures 1A,B**, 203 over-expressed and 308 low-expressed genes were screened via GSE75415. Furthermore, 258 genes with high expression and 466 genes with low expression were explored via GSE12368. The DEGs both belonged to GSE75415 and GSE12368, including 165 genes (59 upregulated and 106 downregulated) which were finally screened out (**Figures 1C,D**). All the DEGs we identified are available in **Supplementary Table S2**.

Weighted Co-Expression Network Construction and Key Module Identification

After weeding out the outlier samples, a total of 29 samples were used in WGCNA (**Supplementary Figures S2A,B**). After constructing a co-expression network, the soft-thresholding [beta (β) = 9 (scale free $R^2 = 0.84$)] was determined as shown in **Supplementary Figures S2C–F**. In WGCNA, soft-thresholding was used for further adjacencies evaluation.

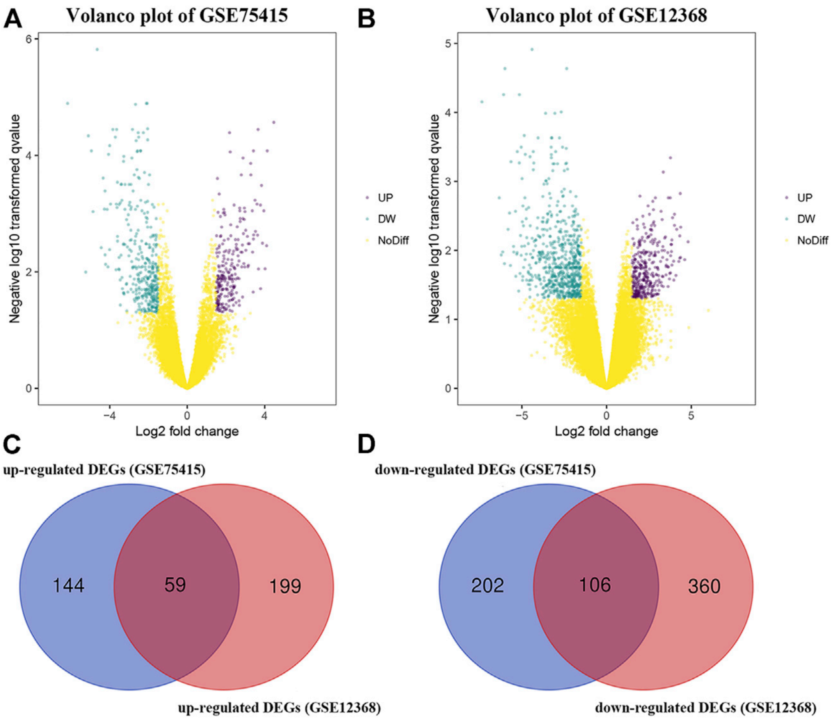


FIGURE 1 | (A) Volcano plot visualizing DEGs in GSE75415. **(B)** Volcano plot visualizing DEGs in GSE12368. **(C)** Identification of overlapped upregulated DEGs between GSE75415 and GSE12368. **(D)** Identification of overlapped downregulated DEGs between GSE75415 and GSE12368.

Immediately, genes were assigned to modules. Also, modules with pairwise correlation of > 0.75 were merged. Finally, 51 modules were screened out (**Supplementary Figure S2G**). Among them, the most relevant module was the blue module ($P = 2e\text{-}05$, $r = 0.80$) (**Figure 2A**). We also found that the MS of the blue module was the highest compared with the rest of the modules (**Figure 2B**). As shown in **Figure 2C**, MM and GS of the blue module showed a significant relationship ($P = 1e\text{-}200$, $cor = 0.73$). Thus, we regarded the blue module as the key module in the present study.

Five Small Molecule Drugs Showed Powerful Potential to Treat ACC

By performing CMap analysis, we could recommend some drugs to treat ACC. As shown in **Supplementary Table S3**, we screened out eight molecule drugs. Five small molecule drugs including chlorpromazine, trifluoperazine, alpha-estradiol, 15-delta prostaglandin J2, and vorinostat might be potential drugs to treat ACC. The detailed information of the five drugs is shown in **Supplementary Table S3**.

Candidate Hub Gene and Hub Gene Identification

Firstly, a PPI network of the 165 DEGs was built. We regarded 74 genes as hub biomarkers because of their high degrees of connectivity (degree ≥ 4, **Supplementary Figure S3A**). A total of 123 genes with |cor.geneModuleMembership| > 0.8; |cor.geneTraitSignificance| > 0.2 were screened, 99 of which were subsequently chosen via PPI network construction (degree ≥ 4, **Supplementary Figure S3B**). Finally, 29 genes overlapping between hub genes in DEGs (n = 74) and hub genes in the hub modules (n = 99) were identified, which were considered to be candidate hub genes.

As shown in **Supplementary Table S4**, the survival analysis indicated that 24 genes were associated with overall survival (OS) and diseases-free survival (DFS) in TCGA-ACC data. A total of 19 genes were associated with OS in GSE19750. Overall, 21 genes in GSE76019 and 29 genes in GSE76021 were associated with event-free survival (EFS). Genes that showed a significant p value ($p < 0.05$) in these survival analyses were considered to be hub genes related to survival and prognosis of patients with ACC. Finally, nine genes [ASPM (abnormal spindle microtubule assembly), BIRC5 (baculoviral IAP repeat containing 5), CCNB2 (cyclin B2), CDK1 (cyclin dependent kinase 1), DLGAP5 (DLG associated protein 5), FOXM1 (forkhead box M1), RACGAP1 (Rac GTPase activating protein 1), TOP2A (DNA topoisomerase II alpha), and TPX2 (TPX2 microtubule nucleation factor)] were screened out. The results of survival analyses of the hub genes are shown in **Figure 3** (OS, TCGA-ACC data), **Supplementary Figure S4** (DFS, TCGA-ACC data), **Supplementary Figure S5** (OS, GSE19750), **Supplementary Figure S6** (EFS, GSE76019), and **Supplementary Figure S7** (EFS, GSE76021). Also, we explored univariate Cox analysis for the nine genes based on TCGA-ACC data, GSE76019, and

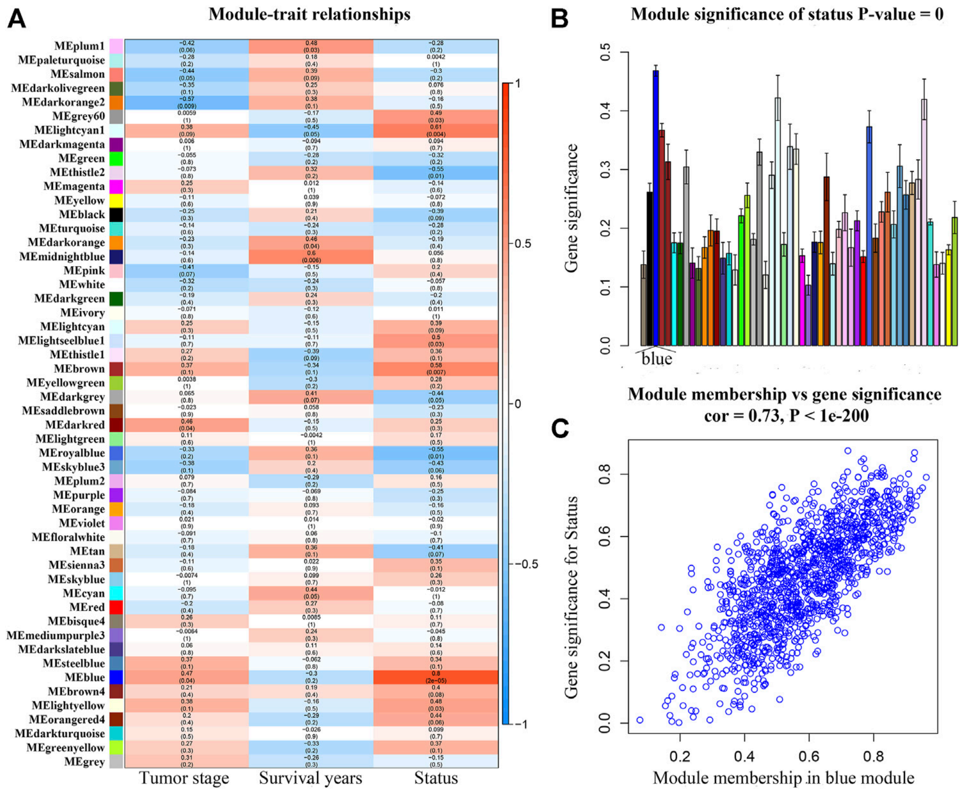


FIGURE 2 | (A) Heatmap of the correlation between module eigengenes (MEs) and different clinical information of ACC [tumor stage, survival years (survival time), and survival status]. **(B)** Distribution of average gene significances and errors in the modules associated with the survival status of ACC. **(C)** Scatter plot of module eigengenes related to survival status in the blue module.

GSE76021. As shown in **Supplementary Table S5**, the result was consistent with what we got for survival analysis.

Hub Gene Validation

Based on GSE10927, GSE19750, GSE75415, GSE76019, GSE76021, and TCGA-ACC data, the stage plots of hub genes were determined and these genes did not perform as well as we expected. In TCGA-ACC data, ASPM (F = 6.939, p = 0.001), BIRC5 (F = 3.368, p = 0.034), CCNB2 (F = 4.844, p = 0.009), CDK1 (F = 6.779, p = 0.001), DLGAP5 (F = 4.170, p = 0.014), FOXM1 (F = 7.569, p = 0.001), RACGAP1 (F = 4.717, p = 0.009), TOP2A (F = 4.687, p = 0.008), and TPX2 (F = 5.232, p = 0.005) were significantly associated with tumor stage (**Supplementary Table S6**). In GSE10927, only ASPM showed a significant p value (F = 4.254, p = 0.030) (**Supplementary Table S6**). In GSE19750, GSE75415, and GSE76019, unfortunately none of these hub genes were closely relevant to tumor stage (**Supplementary Table S6**). In GSE76021, only CCNB2 (F = 7.569, p = 0.001) was significantly related to tumor stage (**Supplementary Table S6**). As for grade plots, the results of the unpaired t test suggested that ASPM, BIRC5, CCNB2, CDK1, DLGAP5, FOXM1, RACGAP1, TOP2A, and TPX2 were closely related to tumor grade based on GSE10927 (the p values are shown in **Supplementary Table S7**). In GSE19750, only CCNB2 (F = 6.271, p = 0.013) was significantly associated with tumor grade (**Supplementary Table S6**). In bioinformatics analysis of each dataset (GSE10927, GSE12368, GSE19750, and GSE75415), all the hub genes were highly expressed in ACCs compared to normal tissue (**Supplementary Table S8**).

ROC and DCA

By using GSE10927, GSE12368, GSE19750, and GSE75415, ROC curve analysis was performed and the AUC was evaluated for distinguishing ACCs and normal samples. The AUC values of hub genes were greater than 0.84, which suggested that all of the hub genes could distinguish ACCs from normal tissues well (**Table 1**). Also, the AUC was calculated to distinguish localized ACC (stages I or II) and advanced ACC (stages III or IV) based on all the datasets we mentioned in this study. In TCGA-ACC data, all the hub genes could distinguish localized ACC and advanced ACC well (**Table 1**). In GSE19750, BIRC5 (AUC = 0.727) and TOP2A (AUC = 0.765) worked well (**Table 1**). In GSE76019, only ASPM (AUC = 0.713) could distinguish localized ACC and advanced ACC well (**Table 1**). In GSE10927, GSE75415, and GSE76021, none of these hub genes could distinguish localized ACC from advanced ACC well (**Table 1**), which is not what we expected. According to the results of distinguishing ACC of low grade and ACC of high grade, all these genes showed a significant p value (AUC > 0.80) based on GSE10927 (**Table 1**). But in GSE19750, BIRC5

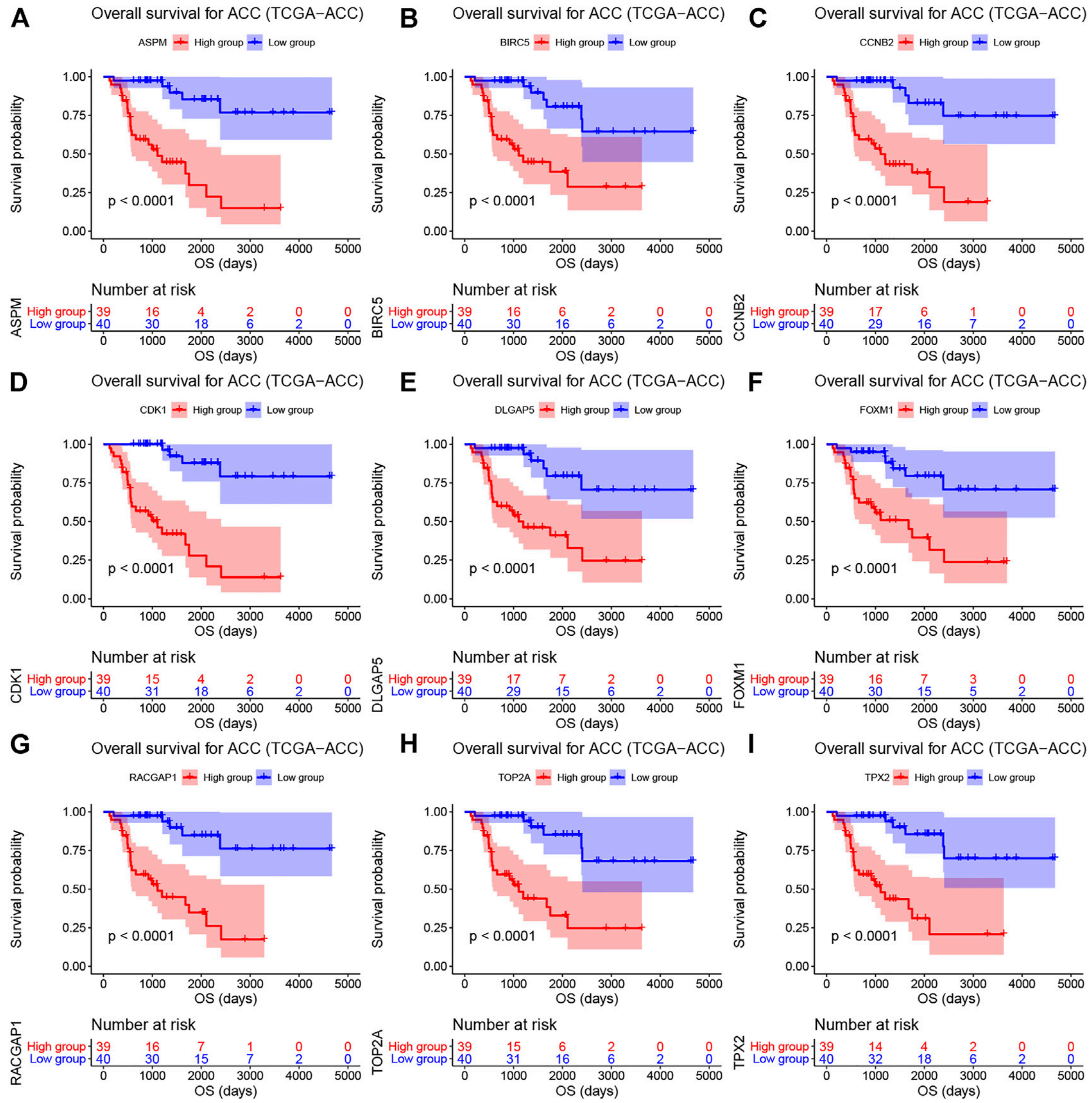


FIGURE 3 | Overall survival analyses on hub genes (ASPM (A), BIRC5 **(B)**, CCNB2 **(C)**, CDK1 **(D)**, DLGAP5 **(E)**, FOXM1 **(F)**, RACGAP1 **(G)**, TOP2A **(H)**, and TPX2 **(I)**) based on TCGA-ACC data. Survival curves for patients in different groups. Red lines represent high expression of hub genes, while blue lines represent low expression of hub genes.

(AUC = 0.495) could not distinguish ACC of low grade and ACC of high grade well (**Table 1**). As for results of AUC to distinguish ACA and ACC, the AUC values of hub genes were greater than 0.85 by using GSE10927 and GSE12368, which suggested that all the hub genes worked well (**Table 1**). But in GSE75415, only FOXM1 (AUC = 0.747), TOP2A (AUC = 0.726), and TPX2 (AUC = 0.726) could distinguish ACC and ACA well. All the results of this part are shown in **Table 1**. As for the DCA results, eight of the hub genes (ASPM, BIRC5, CDK1, DLGAP5, FOXM1, RACGAP1, TOP2A, and TPX2) expressed a strong potential for clinical practice (**Supplementary Figure S8**). Whatever the threshold probability (Pt) expressed, the eight genes displayed great potential. For CCNB2, it performed well, only Pt was approximately between 0.20 and 0.60. All in all, these results suggested that though these hub genes performed well in some datasets, they need to be tested by more in-depth study.

Hub Gene-Associated Biological Pathways

GO analysis indicated that candidate hub genes were involved in 10 biological processes (BPs), including nuclear division, organelle fission, mitotic nuclear division, chromosome segregation, nuclear chromosome segregation, sister chromatid segregation, mitotic sister chromatid segregation, cell cycle checkpoint, regulation of chromosome segregation, and microtubule cytoskeleton organization involved in mitosis (**Supplementary Figure S9A**). As for the KEGG pathways,

TABLE 1 | AUC of hub genes.

Genes	Normal tissues vs. ACC				Stages I/II vs. stages III/IV						Low grade vs. high grade		ACA vs. ACC		
	GSE10927	GSE12368	GSE19750	GSE75415	GSE10927	GSE19750	GSE75415	GSE76019	GSE76021	TCGA	GSE10927	GSE19750	GSE10927	GSE12368	GSE75415
ASPM	0.982	0.986	0.957	0.917	0.532	0.553	0.513	0.713	0.593	0.751	0.86	0.859	0.972	0.995	0.663
BIRC5	0.968	0.861	0.932	0.895	0.511	0.727	0.563	0.661	0.591	0.762	0.823	0.495	0.979	0.859	0.695
CCNB2	0.991	1	0.852	0.902	0.481	0.606	0.538	0.654	0.598	0.799	0.858	0.901	0.986	1	0.674
CDK1	0.988	0.858	0.966	0.91	0.474	0.58	0.6	0.671	0.529	0.799	0.877	0.82	0.977	0.99	0.674
DLGAP5	0.979	0.931	0.847	0.88	0.489	0.564	0.613	0.654	0.544	0.76	0.912	0.875	0.987	0.885	0.642
FOXM1	0.994	0.917	0.952	0.914	0.57	0.67	0.575	0.63	0.618	0.752	0.813	0.747	0.98	0.938	0.747
RACGAP1	1	0.944	0.989	0.925	0.53	0.561	0.475	0.637	0.544	0.745	0.896	0.823	1	0.979	0.589
TOP2A	0.97	0.858	0.969	0.955	0.538	0.765	0.463	0.682	0.578	0.763	0.827	0.685	0.979	0.901	0.726
TPX2	0.97	1	1	0.962	0.5	0.644	0.488	0.626	0.578	0.733	0.823	0.859	0.968	0.953	0.726

Note: AUC: area under curve.

candidate biomarkers were majorly associated with cell cycle, progesterone-mediated oocyte maturation, oocyte meiosis, cellular senescence, and p53 signaling pathway (**Supplementary Figure S9B**). To summarize, we found that candidate hub genes were majorly associated with cell cycle and DNA replication-related biological pathways.

Four MPBs Showed Powerful Potential to Predict OS

To pick out some genes with great value for predicting OS among the nine hub genes, three methods including LDA, KNN, and SVM from the machine learning field were included in this part. As **Table 2** describes, though all the sixteen biomarkers might perform well in recognizing ACCs from alive ACC samples (the average accuracy ≥ 0.70), four MPBs including ASPM (average accuracy = 0.8228), BIRC5 (average accuracy = 0.8059), CCNB2 (average accuracy = 0.8080), and CDK1 (average accuracy = 0.8080) were screened out for more accurate prediction of OS. Furthermore, time-dependent ROC analysis for the four genes was conducted. We concluded that all the four MPBs could predict OS of patients with ACC well (**Figure 4**). For ASPM, the AUCs of 1-, 3-, and 5-years OS were 0.816, 0.939, and 0.885, respectively (**Figure 4A**). For BIRC5, the AUCs of 1-, 3-, and 5-years OS were 0.816, 0.953, and 0.790, respectively (**Figure 4B**). For CCNB2, the AUCs of 1-, 3-, and 5-years OS were 0.762, 0.948, and 0.805, respectively (**Figure 4C**). For CDK1, the AUCs of 1-, 3-, and 5-years OS were 0.841, 0.925, and 0.863, respectively (**Figure 4D**).

Mutations and CNVs of MPBs Were Associated With OS of Patients With ACC

According to the result, four MPBs were altered in 15 (20%) of 76 ACC patients (**Figure 5B**). The most altered gene was ASPM (12%, **Figure 5A**). And we further concluded that mRNA high was the main type (**Figure 5A**). For exploring the relationship between gene expression and gene alteration, we found that genes with more alterations were more likely to be highly expressed. **Figure 5C** shows the network containing 54 nodes (including 4 MPBs and 50 most altered neighbor genes). In addition, this network also demonstrated that CDK1 and BIRC5 were the targets of some kinds of anticancer drugs, which suggested that ASPM and CCNB2 might be new therapeutic targets to treat ACC. Moreover, CNVs of ASPM (gains), BIRC5 (shallow deletions, gains), CCNB2 (shallow deletions, gains), and CDK1 (shallow deletions, gains) caused their higher expressions compared with samples without CNVs (diploids), which demonstrated that CNVs of MPBs were associated with their expression levels (**Figure 5D**).

As for the effect of CNVs and mutations of genes on OS, we concluded that ACCs with ASPM shallow deletions ($p = 0.0200$) had better OS compared to those affected by ASPM copy number gains. In addition, there was a contrary conclusion that ACCs with shallow deletions in CDK1 ($p = 0.0047$) had poor OS (**Figure 5E**). Moreover, ACCs of alterations in the four biomarkers had worse OS (total alterations: $p < 0.0001$; ASPM alterations: $p = 0.00015$; BIRC5 alterations: $p = 0.00055$; CCNB2 alterations: $p < 0.0001$; CDK1 alterations: $p < 0.0001$; **Figure 5F**).

TABLE 2 | The accuracy of classification of LDA-based classifier, KNN-based classifier, linear-SVM-based classifier, polynomial-SVM-based classifier, RBF-SVM-based classifier, and sigmoid-kernel-SVM based classifier.

TCGA-ACC	LDA	KNN	Linear-SVM	Polynomial-SVM	RBF-SVM	Sigmoid-kernel SVM	Average accuracy
ASPM	0.8354	0.8101	0.8101	0.8228	0.8354	0.8228	0.8228
BIRC5	0.8101	0.8228	0.8101	0.7722	0.8101	0.8101	0.8059
CCNB2	0.7975	0.8481	0.7975	0.8101	0.7975	0.7975	0.8080
CDK1	0.7848	0.8354	0.8101	0.8101	0.8101	0.7975	0.8080
DLGAP5	0.7595	0.7975	0.7595	0.7722	0.7722	0.7848	0.7743
FOXM1	0.7595	0.7848	0.7595	0.7595	0.7468	0.7722	0.7637
RACGAP1	0.7975	0.7848	0.7975	0.7975	0.7975	0.6329	0.7680
TOP2A	0.7848	0.7975	0.7342	0.7975	0.7975	0.7089	0.7701
TPX2	0.7342	0.7848	0.7342	0.7595	0.7468	0.7342	0.7490

Note: LDA: linear discriminant analysis; KNN: K-nearest neighbor; RBF: radial basis function; SVM: support vector machine.

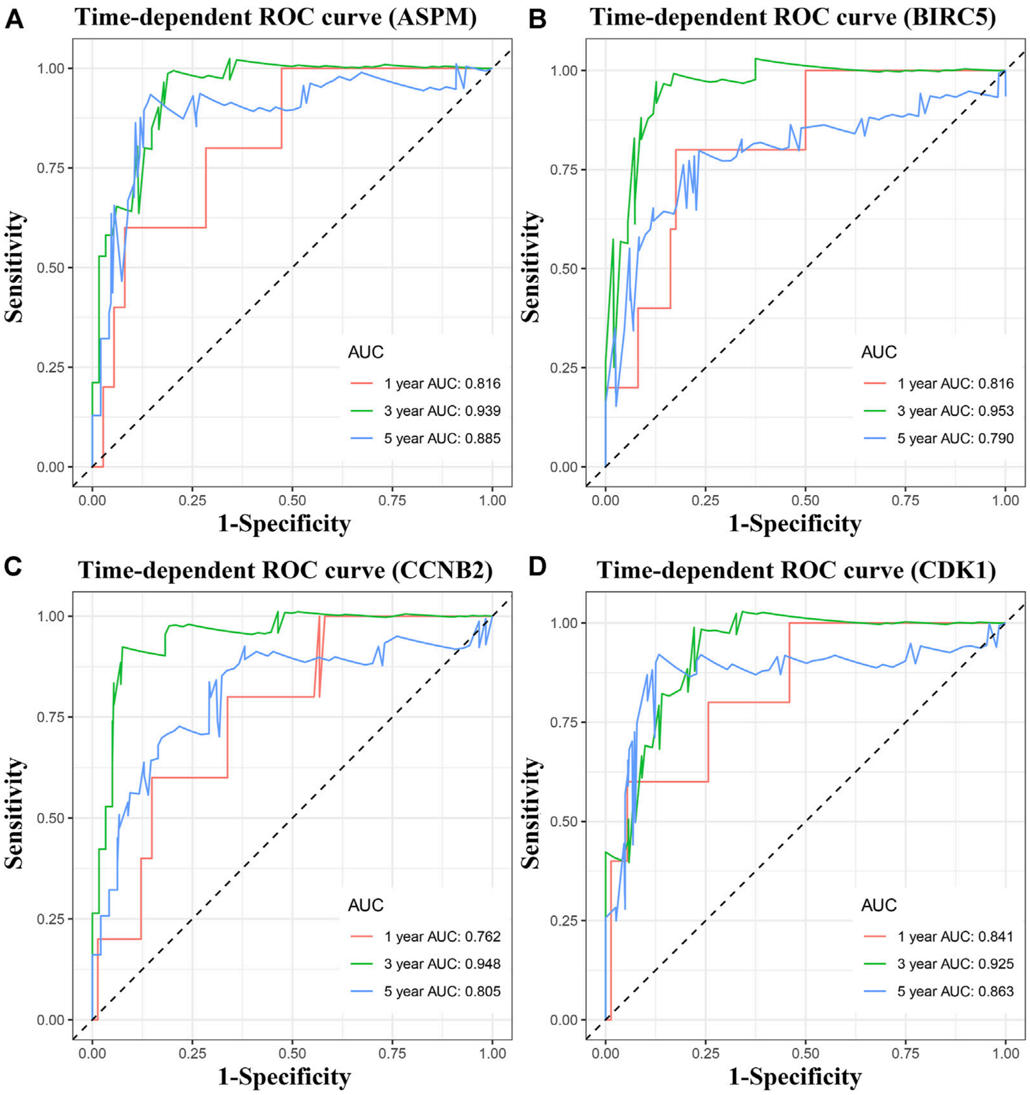


FIGURE 4 | Time dependent ROC analyses at 1, 3, and 5 years based on MPB expression. **(A)** ASPM, **(B)** BIRC5, **(C)** CCNB2, and **(D)** CDK1.

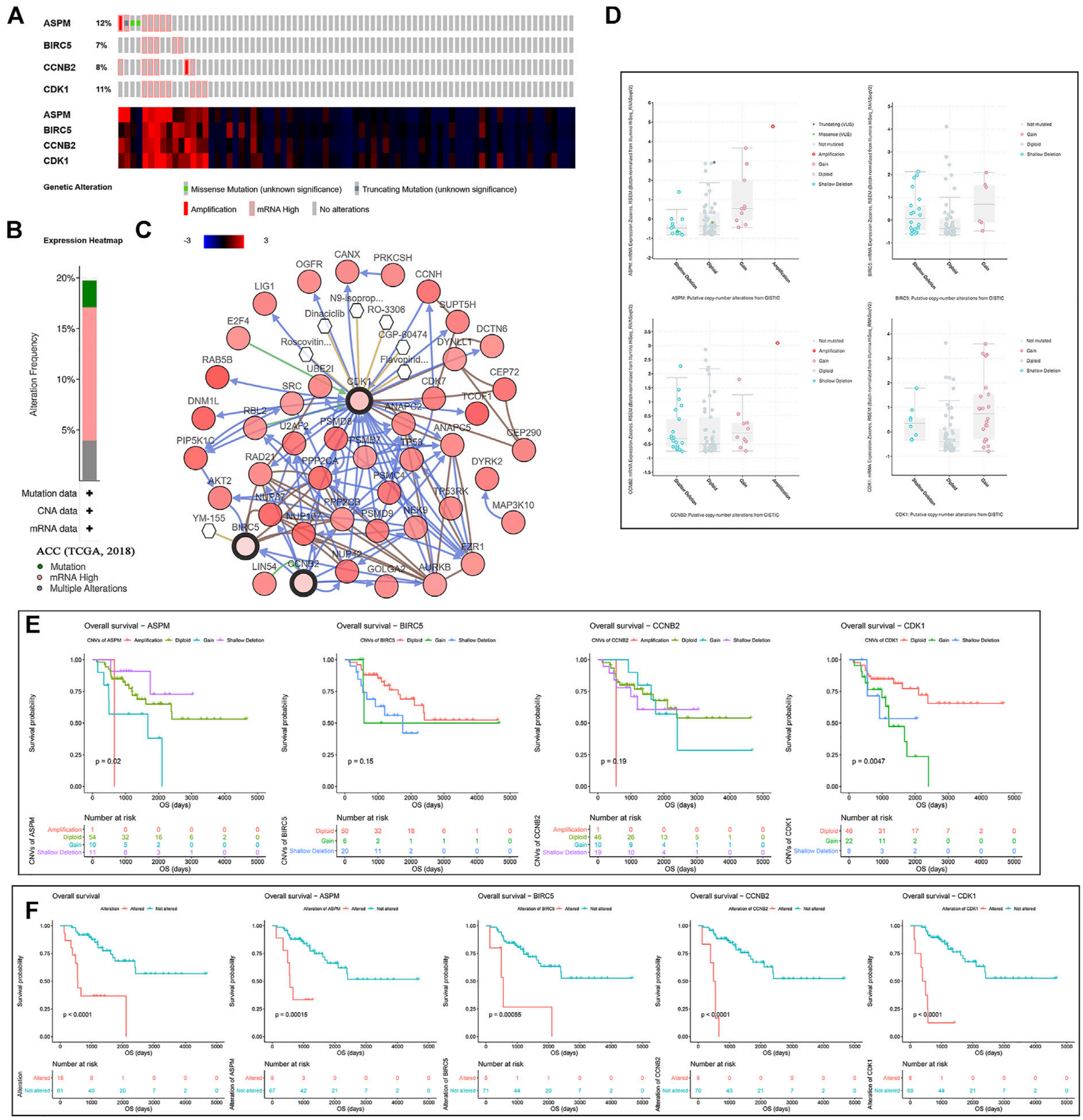


FIGURE 5 | A summary of mutations and CNVs of MPBs. **(A)** Genetic alterations associated with MPBs and expression heatmap of MPBs based on the data from TCGA. **(B)** The total alteration frequency of MPBs in TCGA-ACC is illustrated. **(C)** The network contains 54 nodes, including our 4 query genes and the 50 most frequently altered neighbor genes. The relationship between hub genes and tumor drugs is also illustrated. **(D)** Correlation between different CNV patterns and mRNA expression levels of MPBs respectively. **(E)** Survival analysis of ACC patients with CNVs of MPBs based on TCGA ACC data. **(F)** Survival analysis of ACC patients with mutations of MPBs based on TCGA ACC data.

Identification of MPB-Related Biological Pathways and Drugs

With the standards set before, interestingly, ASPM, BIRC5, CCNB2, and CDK1 were involved in just one KEGG signaling pathway called cell cycle as shown in **Supplementary Table S8** (*p* values are also described in **Supplementary Table S8**). As for the enriched drugs, ASPM was significantly enriched in 6 drugs, BIRC5 was associated with 6 drugs, CCNB2 was related to 11 drugs, and CDK1 was enriched in 6 drugs (**Table 3**).

MPB Expressions Were Related to Immune Infiltration Level in ACC

Immune infiltration level was an independent predictor of sentinel lymph node status and survival in tumors. Here we

TABLE 3 | Gene set enrichment analyses in four hub genes' (ASPM, BIRC5, CCNB2, CDK1) high-expression phenotype.

Gene symbol	Reference gene set	Name	Size	ES	NES	NOM *p*-val	FDR *q*-val
ASPM	c2.cp.kegg.v7.0.symbols.gmt	KEGG_CELL_CYCLE	122	−0.7048	−2.1129	0.0000	0.0029
	DSigDBv1.0.gmt	LUCANTHONE_CTD_00006227	202	−0.8190	−1.8734	0.0000	0.0357
		MONOBENZONE_PC3_DOWN	196	−0.6530	−2.1874	0.0000	0.0160
		8-AZAGUANINE_PC3_DOWN	192	−0.6831	−2.1539	0.0000	0.0160
		THIOGUANOSINE_MCF7_DOWN	145	−0.6482	−2.0313	0.0000	0.0392
		AZACYCLONOL_MCF7_UP	123	−0.6266	−1.5119	0.0481	0.2097
		RESVERATROL_MCF7_DOWN	100	−0.8126	−1.8743	0.0000	0.0377
BIRC5	c2.cp.kegg.v7.0.symbols.gmt	KEGG_CELL_CYCLE	122	−0.7182	−2.2001	0.0000	0.0000
	DSigDBv1.0.gmt	DASATINIB_CTD_00004330	474	−0.6359	−1.7919	0.0000	0.0585
		LUCANTHONE_CTD_00006227	202	−0.8058	−1.8166	0.0000	0.0538
		MONOBENZONE_PC3_DOWN	196	−0.6272	−2.0778	0.0020	0.0415
		8-AZAGUANINE_PC3_DOWN	192	−0.6720	−2.1018	0.0000	0.0488
		THIOGUANOSINE_MCF7_DOWN	145	−0.6460	−1.9901	0.0000	0.0248
		RESVERATROL_MCF7_DOWN	100	−0.8303	−1.9074	0.0000	0.0322
CCNB2	c2.cp.kegg.v7.0.symbols.gmt	KEGG_CELL_CYCLE	122	−0.6765	−2.0239	0.0000	0.0110
	DSigDBv1.0.gmt	DASATINIB_CTD_00004330	474	−0.6032	−1.7238	0.0041	0.0930
		LUCANTHONE_CTD_00006227	202	−0.7869	−1.7530	0.0000	0.0810
		MONOBENZONE_PC3_DOWN	196	−0.6328	−2.0912	0.0000	0.0598
		8-AZAGUANINE_PC3_DOWN	192	−0.6633	−1.9833	0.0000	0.0525
		THIOGUANOSINE_MCF7_DOWN	145	−0.6569	−1.9812	0.0000	0.0481
		PRENYLAMINE_MCF7_UP	140	−0.6343	−1.6213	0.0120	0.1535
		MEFLOQUINE_MCF7_UP	136	−0.6014	−1.5385	0.0373	0.2092
		AZACYCLONOL_MCF7_UP	123	−0.6562	−1.5766	0.0237	0.1748
		AZACITIDINE_PC3_UP	115	−0.6318	−1.4942	0.0339	0.2350
		FENDILINE_MCF7_UP	112	−0.6230	−1.5397	0.0354	0.2081
		RESVERATROL_MCF7_DOWN	100	−0.8189	−1.8592	0.0000	0.0580
CDK1	c2.cp.kegg.v7.0.symbols.gmt	KEGG_CELL_CYCLE	122	−0.7160	−2.1710	0.0000	0.0049
	DSigDBv1.0.gmt	DASATINIB_CTD_00004330	474	−0.6174	−1.7945	0.0020	0.0569
		LUCANTHONE_CTD_00006227	202	−0.8038	−1.8426	0.0000	0.0481
		MONOBENZONE_PC3_DOWN	196	−0.6407	−2.1553	0.0000	0.0338
		8-AZAGUANINE_PC3_DOWN	192	−0.6686	−2.0766	0.0000	0.0501
		THIOGUANOSINE_MCF7_DOWN	145	−0.6511	−2.0313	0.0000	0.0485
		RESVERATROL_MCF7_DOWN	100	−0.8223	−1.8556	0.0000	0.0474

Note: ES, enrichment score; NES, normalized enrichment score; NOM p*-val, nominal* p *value; FDR, false discovery rate q value.*

assessed the correlation of MPB expressions with immune infiltration level in ACC. The analysis concluded that ASPM expression was positively relevant to tumor purity (cor = 0.300, *p* = 0.009) and infiltrating levels of B cells (cor = 0.272, *p* = 0.020) and dendritic cells (cor = 0.236, *p* = 0.044) but had no significant correlations with infiltrating levels of CD8 + T cells, CD4 + cells, macrophages, and neutrophils (**Supplementary Figure S10A**). Unfortunately, BIRC5 expression was not related to tumor purity or infiltrating levels of immune cells (**Supplementary Figure S10B**). Moreover, there was a positive relationship between CCNB2 expression and tumor purity (cor = 0.268, *p* = 0.021) and infiltrating level of dendritic cells (cor = 0.238, *p* = 0.043) (**Supplementary Figure S10C**). As for CDK1, the expression of CDK1 only had a positive correlation with tumor purity (cor = 0.283, *p* = 0.015) (**Supplementary Figure S10D**). To summarize, we found that ASPM, CCNB2, and CDK1 expressions were significantly associated with tumor purity, which could be a sign that ASPM, CCNB2, and CDK1 played specific roles in immune infiltration in ACC.

In addition, as shown in **Supplementary Figure S13**, ASPM expression (**Supplementary Figure S11A**), BIRC5 expression (**Supplementary Figure S11B**), and CCNB2 expression (**Supplementary Figure S11C**) were significantly higher in 17 types of cancer compared with adjacent normal tissues. In 25 cancer types, CDK1 showed an upregulated trend when comparing to normal tissues (**Supplementary Figure S11D**). Unfortunately, there was a lack of adjacent normal tissues in ACC (based on TCGA data), and we could not compare the expressions between ACC and normal tissue of MPBs. But as per our previous results in this study, all the MPBs were over-expressed (based on GSE10927, GSE12368, GSE19750, and GSE75415).

Association of MPB Expressions With Immune Microenvironment Score Levels

Then we found that patients with ACC in the high ESTIMATE-score group had better OS and disease-free survival (DFS). Patients of low ASPM expression and low CDK1 expression had higher ESTIMATE scores (**Figure 6C**). As shown in **Figure 6D**, there was a trend where ACCs which had a high-immune score had superior OS compared to those with a low-immune score. Meanwhile, patients with a high-immune score had better disease-free survival (DFS) compared with patients with a low-immune score, significantly (**Figure 6E**). The unpaired

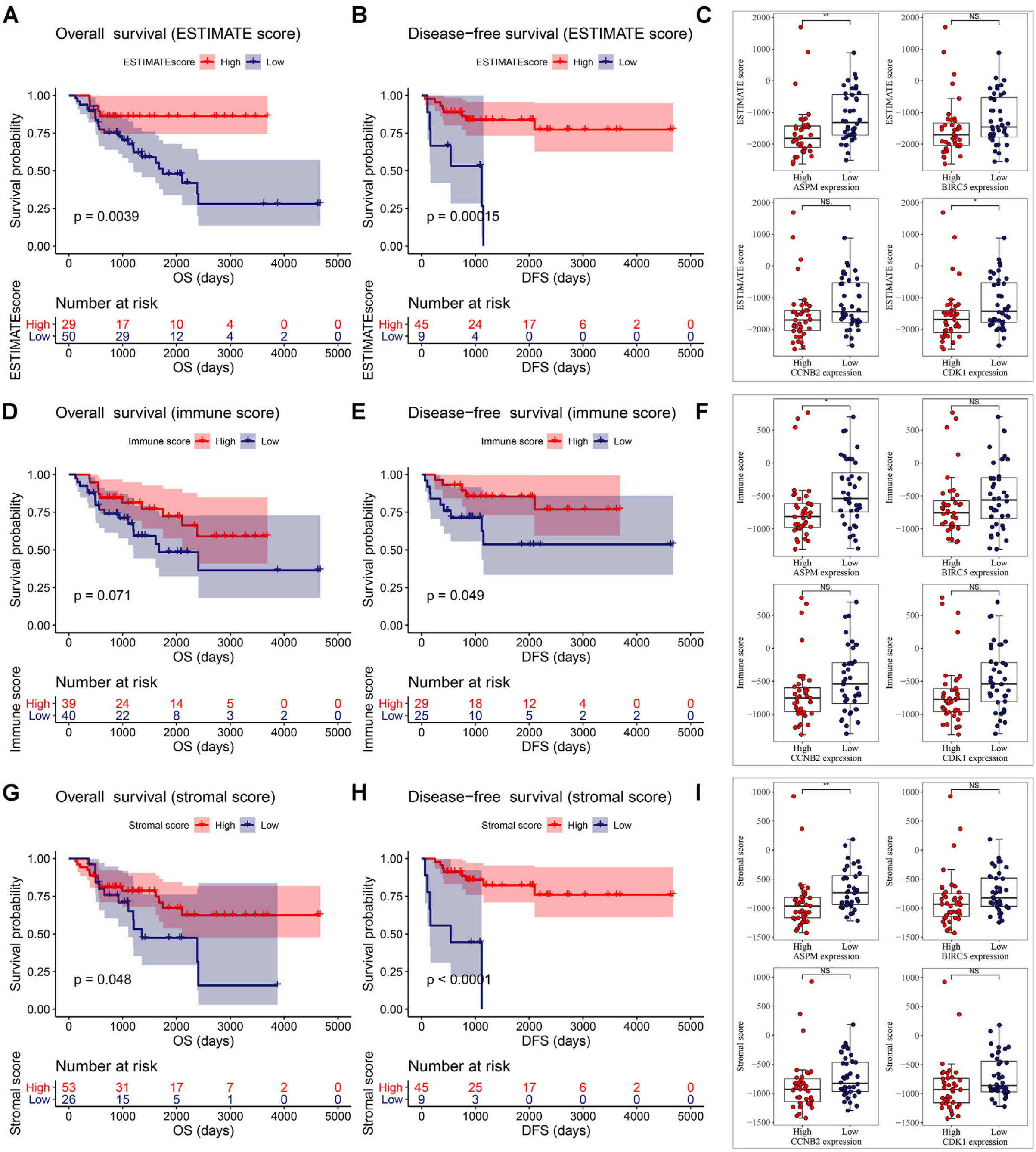


FIGURE 6 | ESTIMATE scores were associated with overall survival **(A)** and disease-free survival **(B)** of patients with ACC. Correlation of MPB **(C)** expression with ESTIMATE scores in ACC. Immune scores were associated with overall survival **(D)** and disease-free survival **(E)** of patients with ACC. Correlation of MPB **(F)** expression with immune scores in ACC. Stromal scores were associated with overall survival **(G)** and disease-free survival **(H)** of patients with ACC. Correlation of MPB **(I)** expression with stromal scores in ACC. *: $p < 0.05$; **: $p < 0.01$; NS: no significance.

t test indicated that there was a negative relationship between ASPM expression and immune score level (**Figure 6F**). As shown in **Figure 6G**, an ACC patient with a high-stromal score had better OS. Meanwhile, patients with a high-stromal score had better disease-free survival (DFS) compared with patients with a low-stromal score, significantly (**Figure 6H**). The unpaired *t* test also suggested that there was a negative relationship between ASPM expression and stromal score level (**Figure 6I**). These results demonstrated that high expressions of MPBs had worse OS and DFS in ACC patients indirectly.

TABLE 4 | Cox univariable and multivariable analyses of overall survival (OS) and disease-free survival (DFS).

	Variable	Univariate analysis				Multivariate analysis			
		HR	LCI	UCI	*p* value	HR	LCI	UCI	*p* value
Overall survival (OS)	ASPM	3.184	2.157	4.701	<0.001	3.262	1.138	9.350	0.028
	BIRC5	2.862	1.946	4.209	<0.001	1.941	0.811	4.646	0.137
	CCNB2	2.812	1.935	4.085	<0.001	0.989	0.443	2.208	0.978
	CDK1	3.873	2.342	6.405	<0.001	0.489	0.140	1.710	0.263
	Age	1.011	0.987	1.036	0.365				
	Gender	1.000	0.468	2.135	0.999				
	Laterality	0.841	0.394	1.796	0.654				
	Pathologic stage	2.912	1.858	4.562	<0.001	1.943	1.191	3.170	0.008
Disease-free survival (DFS)	ASPM	2.768	1.450	5.284	0.002	7.335	1.538	34.994	0.012
	BIRC5	1.674	0.979	2.865	0.060				
	CCNB2	2.441	1.422	4.189	0.001	2.297	0.717	7.362	0.162
	CDK1	1.928	1.040	3.576	0.037	0.114	0.021	0.617	0.012
	Age	0.998	0.963	1.034	0.919				
	Gender	2.386	0.665	8.563	0.182				
	Laterality	0.423	0.132	1.348	0.146				
	Pathologic stage	1.848	1.024	3.338	0.042	1.205	0.597	2.433	0.602

Prognostic Value of the Four Biomarkers

According to the result of univariate Cox analysis (**Table 4**), ASPM, BIRC5, CCNB2, CDK1, and pathologic stage were interfering factors of OS. *p* values are shown in **Table 4**. Subsequent multivariate Cox analysis confirmed that ASPM could predict the prognosis of ACC patients by individual. By using the Coxph function in R package “survival”, we conducted a Schoenfeld individual test for investigating the proportional hazards assumption. The global Schoenfeld test showed no significance ($p = 0.1936$, **Supplementary Figure S12A**). Also, each variable including age ($p = 0.6709$), gender ($p = 0.6919$), laterality ($p = 0.6219$), pathologic stage ($p = 0.1688$), ASPM ($p = 0.3394$), BIRC5 ($p = 0.1285$), CCNB2 ($p = 0.4813$), and CDK1 ($p = 0.0657$) was not statistically significant ($p > 0.05$, **Supplementary Figure S12A**). Thus, this Cox model conformed to the proportional hazards assumption. For DFS, ASPM (hazard ratio = 2.768, 95% CI of ratio: 1.450–5.284, $p = 0.002$), CCNB2 (hazard ratio = 2.441, 95% CI of ratio: 1.422–4.189, $p = 0.001$), CDK1 (hazard ratio = 1.928, 95% CI of ratio: 1.040–3.576, $p = 0.037$), and pathologic stage (hazard ratio = 1.848 95% CI of ratio: 1.024–3.338, $p = 0.042$) were interfering factors of DFS via univariate Cox analysis. ASPM must be the most important factor for DFS of ACC patients suggested by multivariate Cox analysis (hazard ratio = 7.335, $p = 0.012$). By using the Coxph function in R package “survival”, we conducted a Schoenfeld individual test for investigating the proportional hazards assumption. The global Schoenfeld test showed no significance ($p = 0.8934$, **Supplementary Figure S12B**). Also, each variable including age ($p = 0.3864$), gender ($p = 0.8702$), laterality ($p = 0.5409$), pathologic stage ($p = 0.5914$), ASPM ($p = 0.3790$), BIRC5 ($p = 0.4194$), CCNB2 ($p = 0.5385$), and CDK1 ($p = 0.5581$) was not statistically significant ($p > 0.05$, **Supplementary Figure S12B**). Thus, this Cox model conformed to the proportional hazards assumption.

Clinical Application of MPBs

Based on the factors which showed a significant *p* value in multivariate Cox analysis, we constructed two nomograms (one for OS, the other for DFS) to make better use of these prognostic biomarkers. Two features including ASPM and pathologic stage were used for construction of the OS-nomogram (**Figure 7A**) meanwhile the DFS-nomogram contained three factors including ASPM, CDK1, and pathologic stage (**Figure 8A**). By reviewing the C-index and AUC, we found that both the two nomograms performed well in survival prediction. The OS-nomogram could make accurate predictions about ACC patients’ OS via TGCA-ACC data (C-index: 0.875; AUC: 0.871; **Figure 7E**), GSE10927 (C-index: 0.748; AUC: 0.740; **Figure 7F**), and GSE19750 (C-index: 0.612; AUC: 0.844; **Figure 7G**). As for the predication performance of the DFS-nomogram, it was obvious that the DFS-nomogram showed accurate prediction potential of ACC patients’ DFS based on TCGA-ACC data (C-index: 0.834; AUC: 0.818; **Figure 8E**), GSE76019 (C-index: 0.694; AUC: 0.735; **Figure 8F**), and GSE76021 (C-index: 0.749; AUC: 0.783; **Figure 8G**). As the result of the calibration curve suggested, both the OS-nomogram (**Figures 7B–D**) and DFS-nomogram (**Figures 8B–D**) had good prediction effectiveness compared to the ideal model for a nomogram’s 1-, 3-, and 5-years OS estimates.

DISCUSSION

Though ACC is a relatively orphan malignant tumor, most ACCs are diagnosed in advanced stages (Guillaume et al., 2014). The 5-years survival rate of ACC is still not satisfactory (only 35% as reported) (Guillaume et al., 2014). In consideration of the poor prognosis of ACC patients, it was of urgent need to explore a few effective and novel biomarkers predicting the survival and prognosis of patients with ACC by integrative bioinformatics

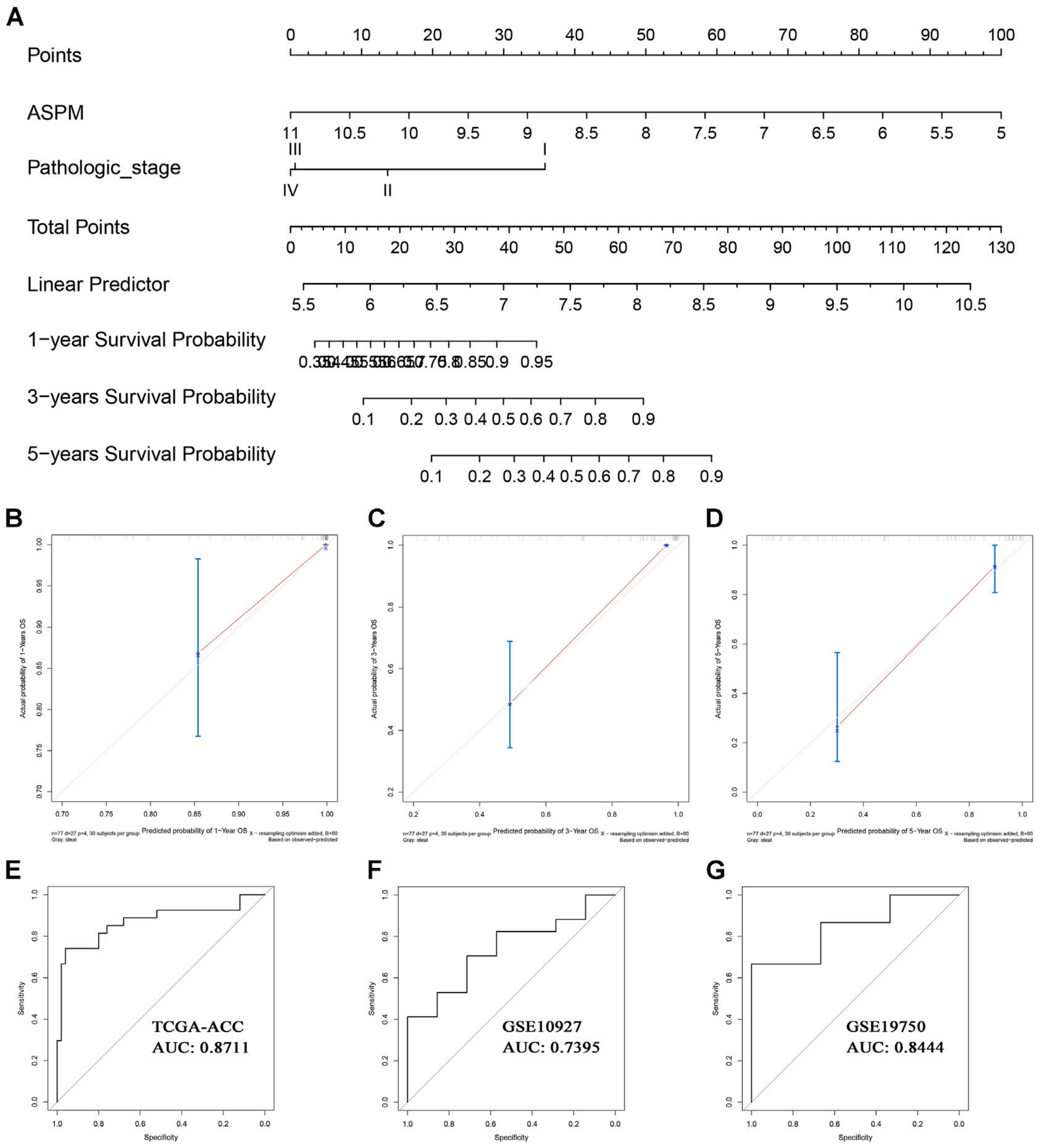


FIGURE 7 | The nomogram for predicting the proportion of ACC patients with 1-, 3-, or 5-years OS **(A)**. The calibration plots for predicting 1- **(B)**, 3- **(C)**, or 5- **(D)** year OS. Receiver operating characteristic (ROC) curves and area under the curve (AUC) statistics to evaluate the diagnostic efficiency of the nomogram in TCGA-ACC data **(E)**, GSE10927 **(F)**, and GSE19750 **(G)**.

analysis. Moreover, we attempted to provide clinicians a simple, quick, and accurate method for survival prediction by constructing nomograms.

Based on WGCNA, DEG, and PPI analysis, we identified nine genes which might be candidate biomarkers in ACC. We further explored the potential functions of these hub genes. The results of functional enrichment analysis suggested that the hub genes were majorly enriched in cell cycle and DNA replication-related pathways. Cell cycle is the basic process of cell proliferation (Kaistha et al., 2015). Interestingly, two previous studies demonstrated that most of the nine biomarkers were effectively involved in the cell cycle of renal cell carcinoma (Chen et al., 2018; Wang et al., 2018), which made us more confident in our findings. Yuan at al. confirmed that ASPM, FOXM1, RACGAP1, and TPX2 were significantly associated with not only tumor progression but also prognosis of ACC (Yuan et al., 2018). In the same datasets they used (TCGA-ACC data and GSE19750), we came to the same conclusion. But in other datasets (GSE10927, GSE19750, GSE75415, GSE76019, and GSE76021), these genes were not significantly related to tumor progression as we expected. Therefore, we thought there needs to be stronger evidence and more in-depth validation for exploring the correlation between the nine genes and tumor progression. Previous studies indicated that

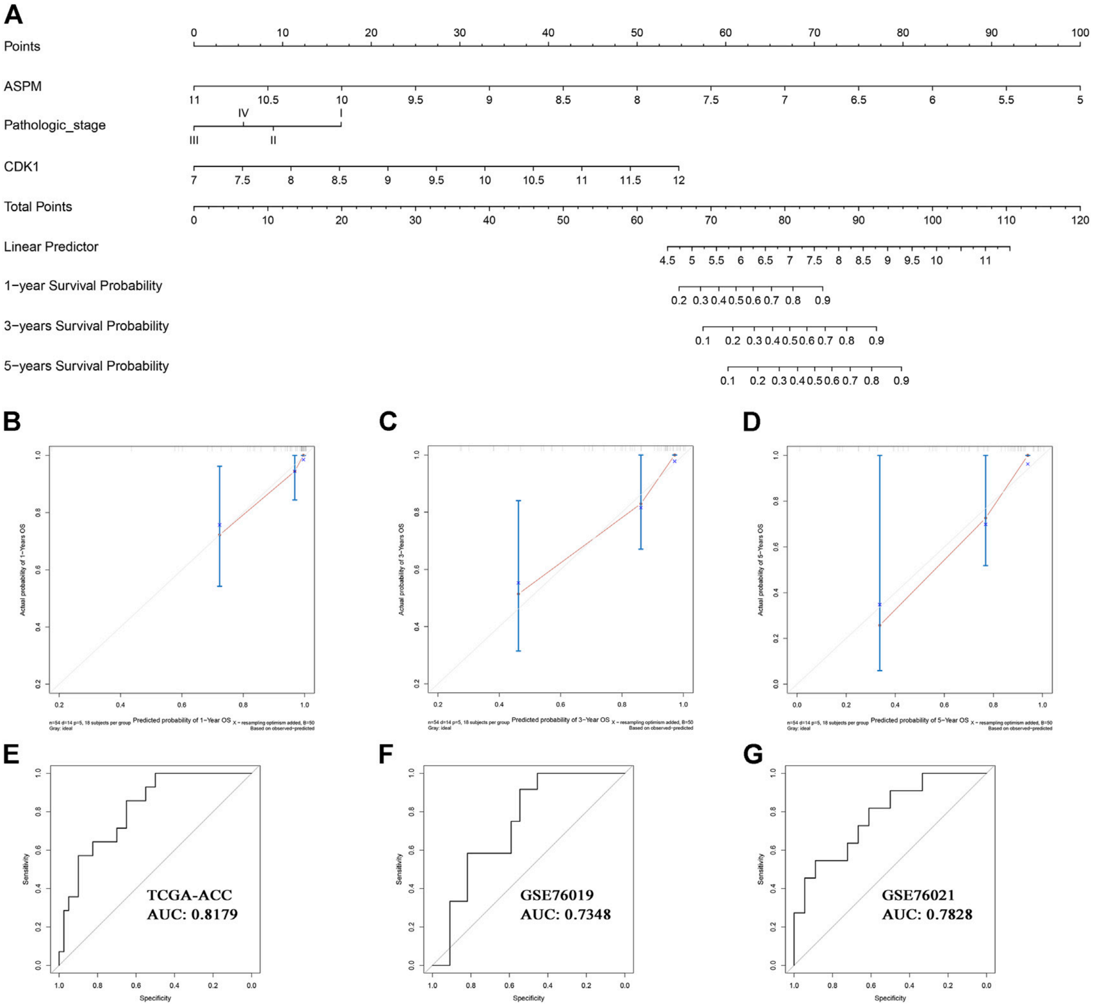


FIGURE 8 | The nomogram for predicting the proportion of ACC patients with 1-, 3-, or 5-years DFS **(A)**. The calibration plots for predicting 1- **(B)**, 3- **(C)**, or 5- **(D)** year DFS. Receiver operating characteristic (ROC) curves and area under the curve (AUC) statistics to evaluate the diagnostic efficiency of the nomogram in TCGA-ACC data **(E)**, GSE76019 **(F)**, and GSE76021**(G)**.

DNA replication regulation was one of the core events of cell cycle regulation (Fragkos et al., 2015). Cell cycle and DNA replication influenced each other and there existed a complicated relationship between them (Lin et al., 2017). To summarize, the conclusions of the above studies provided strong support for suggesting the sixteen genes as new prognostic biomarkers for ACC patients.

Then four MPBs (including ASPM, BIRC5, CCNB2, and CDK1) with higher accuracy in predicting survival were screened out among the nine genes by performing LDA, KNN, SVM, and time-dependent ROC. The effect of mutations and CNVs of MPBs were subsequently evaluated. These MPBs were altered in 15 (20%) patients with ACC. ASPM was altered most and mRNA high was the main type. The next-step process concluded that mutations and CNVs of MPBs were related to ACC patients' OS.

Considering that the tumor immune microenvironment showed a strong correlation with progression and treatment of tumors. We also attempted to explore the relationship in this study. The results suggested that MPB expressions were significantly correlated with immune infiltration level in ACC. Moreover, high expressions of MPBs were effectively associated with worse survival in patients with ACC.

In addition, the CMap analysis demonstrated that five small molecule drugs including chlorpromazine, trifluoperazine, alpha-estradiol, 15-delta prostaglandin J2, and vorinostat might be novel drugs for ACC treatment. These MPBs were also significantly enriched in cell cycle. As for the enriched drugs, ASPM was significantly enriched in 6 drugs, BIRC5 was associated with 6 drugs, CCNB2 was related to 11 drugs, and CDK1 was enriched in 6 drugs. All in all, these drugs might be potential choices for treating ACC.

A nomogram mainly assigns scores to each value level of each influencing factor through the contribution of each influencing factor to the outcome variable in the model, and then adds each score to obtain the total score. Finally, through the functional conversion relationship between the total score and the occurrence probability of the outcome event, the predicted value of the individual outcome event is calculated. In this manuscript, based on the factors which showed a significant *p* value in multivariate Cox analysis, we constructed two nomograms (one for OS, the other for DFS) to make better use of these prognostic biomarkers. Clinicians might realize the individualized and accurate prediction of ACC patients via the two nomograms.

We also have to discuss the deficiencies of our study. Firstly, there was a lack of validation by using *in vitro* or *in vivo* models. Therefore, we will verify the four genes by conducting histology or animal experiments in further research. Secondly, although we identified and validated the four MPBs which were related to prognosis of ACC patients by using several independent datasets, these datasets were of small size, and there was a lack of clinical trials by using samples from patients. Therefore, we need to verify our results by collecting large amounts of patient samples and relevant clinical data in a further study.

In conclusion, we performed eight independent methods to screen nine hub genes related to survival and prognosis of ACC by using seven independent datasets. Four MPBs among them were further screened out, which performed well in ACC survival and prognosis prediction. Furthermore, two nomograms including the OS-nomogram and DFS-nomogram were established, which provided clinicians with a quick, accurate, and visualized method for OS and DFS prediction of patients with ACC.

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

KP, and XY designed and presented the study, XY, YW, and WC developed the analysis procedures, XY, YW, WC, XL, and WL analyzed the results, KP, and XY utilized the analysis tools, and XY wrote the manuscript. All authors agreed to publish the manuscript.

FUNDING

This work was supported by the Science and Technology Planning Project of Yueyang City under Grant (2020-02).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.878073/full#supplementary-material

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J. Clin.* 68, 394–424. doi:10.3322/caac.21492

Chen, L., Yuan, L., Qian, K., Qian, G., Zhu, Y., Wu, C.-L., et al. (2018). Identification of Biomarkers Associated with Pathological Stage and Prognosis of Clear Cell Renal Cell Carcinoma by Co-expression Network Analysis. *Front. Physiol.* 9, 399. doi:10.3389/fphys.2018.00399

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

Demeure, M. J., Coan, K. E., Grant, C. S., Komorowski, R. A., Stephan, E., Sinari, S., et al. (2013). Pttg1 Overexpression in Adrenocortical Cancer Is Associated with Poor Survival and Represents a Potential Therapeutic Target. *Surgery* 154, 1405–1416. doi:10.1016/j.surg.2013.06.058

Fragkos, M., Ganier, O., Coulombe, P., and Méchali, M. (2015). Dna Replication Origin Activation in Space and Time. *Nat. Rev. Mol. Cell Biol.* 16, 360–374. doi:10.1038/nrm4002

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

Giordano, T. J., Kuick, R., Else, T., Gauger, P. G., Vinco, M., Bauersfeld, J., et al. (2009). Molecular Classification and Prognostication of Adrenocortical Tumors by Transcriptome Profiling. *Clin. Cancer Res.* 15, 668–676. doi:10.1158/1078-0432.CCR-08-1067

Guillaume, A., Eric, L., Martin, F., Anne, J., Windy, L., Hanin, O., et al. (2014). Integrated Genomic Characterization of Adrenocortical Carcinoma. *Nat. Genet.* 46, 607–612.

He, Z., Sun, M., Ke, Y., Lin, R., Xiao, Y., Zhou, S., et al. (2017). Identifying Biomarkers of Papillary Renal Cell Carcinoma Associated with Pathological Stage by Weighted Gene Co-expression Network Analysis. *Oncotarget* 8, 27904–27914. doi:10.18632/oncotarget.15842

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent Roc Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics* 56, 337–344. doi:10.1111/j.0006-341x.2000.00337.x

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix Genechip Probe Level Data. *Nucleic Acids Res.* 31, 15e–15. doi:10.1093/nar/gng015

Jasim, S., and Habra, M. A. (2019). Management of Adrenocortical Carcinoma. *Curr. Oncol. Rep.* 21, 20. doi:10.1007/s11912-019-0773-7

Kaistha, B. P., Lorenz, H., Schmidt, H., Sipos, B., Pawlak, M., Gierke, B., et al. (2015). Plac8 Localizes to the Inner Plasma Membrane of Pancreatic Cancer Cells and Regulates Cell Growth and Disease Progression through Critical Cell-Cycle Regulatory Pathways. *Cancer Res.* 76, 96–107. doi:10.1158/0008-5472.can-15-0216

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Kuhn, M. (2015). Caret: Classification and Regression Training. *Astrophys. Source Code Libr.* 129, 291–295.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* 313, 1929–1935. doi:10.1126/science.1132939

Li, 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

Libé, R. (2018). Clinical and Molecular Prognostic Factors in Adrenocortical Carcinoma. *Minerva Endocrinol.* 44, 58–69. doi:10.23736/S0391-1977.18.02900-0

Lin, A. B., McNeely, S. C., and Beckmann, R. P. (2017). Achieving Precision Death with Cell-Cycle Inhibitors that Target Dna Replication and Repair. *Clin. Cancer Res.* 23, 3232–3240. doi:10.1158/1078-0432.CCR-16-0083

Lo, W. M., Kariya, C. M., and Hernandez, J. M. (2019). Operative Management of Recurrent and Metastatic Adrenocortical Carcinoma: a Systematic Review. *Am. Surg.* 85, 23–28. doi:10.1177/000313481908500111

Michael, J. P., and Ralph, B. D. (2010). Overall C as a Measure of Discrimination in Survival Analysis: Model Specific Population Value and Confidence Interval Estimation. *Stat. Med.* 23, 2109–2123.

Patil, I. (2018). *ggstatsplot: 'ggplot2' Based Plots with Statistical Details*. Available at: https://cran.r-project.org/web/packages/ggstatsplot/index.html.

Pinto, E. M., Rodriguez-Galindo, C., Choi, J. K., Pounds, S., Liu, Z., Neale, G., et al. (2016). Prognostic Significance of Major Histocompatibility Complex Class Ii Expression in Pediatric Adrenocortical Tumors: a St. Jude and Children's Oncology Group Study. *Clin. Cancer Res.* 22, 6247–6255. doi:10.1158/1078-0432.CCR-15-2738

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

Sachs, M. C. (2017). Plotroc: A Tool for Plotting Roc Curves. *J. Stat. Softw.* 79, 2. doi:10.18637/jss.v079.c02

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

Soon, P. S. H., Gill, A. J., Benn, D. E., Clarkson, A., Robinson, B. G., McDonald, K. L., et al. (2009). Microarray Gene Expression and Immunohistochemistry Analyses of Adrenocortical Tumors Identify Igf2 and Ki-67 as Useful in Differentiating Carcinomas from Adenomas. *Endocr. Relat. Cancer* 16, 573–583. doi:10.1677/ERC-08-0237

Surakhy, M., Wallace, M., Bond, E., Grochola, L. F., Perez, H., Di Giovannantonio, M., et al. (2020). A Common Polymorphism in the Retinoic Acid Pathway Modifies Adrenocortical Carcinoma Age-dependent Incidence. *Br. J. Cancer* 122, 1231–1241. doi:10.1038/s41416-020-0764-3

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* 43, D447–D452. doi:10.1093/nar/gku1003

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

Venables, W. N., and Ripley, B. D. (2002). Modern Applied Statistics with S. *Statistics Comput.* 52, 704–705. doi:10.1007/978-0-387-21706-2

Vickers, A. J., and Elkin, E. B. (2006). Decision Curve Analysis: a Novel Method for Evaluating Prediction Models. *Med. Decis. Mak.* 26, 565–574. doi:10.1177/0272989x06295361

Wang, Y., Chen, L., Wang, G., Cheng, S., Qian, K., Liu, X., et al. (2018). Fifteen Hub Genes Associated with Progression and Prognosis of Clear Cell Renal Cell Carcinoma Identified by Coexpression Analysis. *J. Cell. Physiology* 234, 10225–10237. doi:10.1002/jcp.27692

West, A. N., Neale, G. A., Pounds, S., Figueredo, B. C., Rodriguez Galindo, C., Pianovski, M. A. D., et al. (2007). Gene Expression Profiling of Childhood Adrenocortical Tumors. *Cancer Res.* 67, 600–608. doi:10.1158/0008-5472.CAN-06-3767

Yizhou, W., Jun, L., Yong, X., Renyan, G., Kui, W., Zhenlin, Y., et al. (2013). Prognostic Nomogram for Intrahepatic Cholangiocarcinoma after Partial Hepatectomy. *J. Clin. Oncol.* 31, 1188–1195.

Yoo, M., Shin, J., Kim, J., Ryall, K. A., Lee, K., Lee, S., et al. (2015). DSigDB: Drug Signatures Database for Gene Set Analysis: Fig. 1. *Bioinformatics* 31, 3069–3071. doi:10.1093/bioinformatics/btv313

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). Clusterprofiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Yuan, L., Qian, G., Chen, L., Wu, C.-L., Dan, H. C., Xiao, Y., et al. (2018). Co-expression Network Analysis of Biomarkers for Adrenocortical Carcinoma. *Front. Genet.* 9, 328. doi:10.3389/fgene.2018.00328

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. doi:10.2202/1544-6115.1128

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: 02 June 2022
doi: 10.3389/fmolb.2022.858577

Identifying Topics and Evolutionary Trends of Literature on Brain Metastases Using Latent Dirichlet Allocation

Jiarong Chen[1,2,3]*, *Matt Williams*[3,4], *Yanming Huang*[1] *and Shijing Si*[5]*

[1]*Clinical Experimental Center, Jiangmen Key Laboratory of Clinical Biobanks and Translational Research, Jiangmen Central Hospital, Jiangmen, China,* [2]*Department of Oncology, Jiangmen Central Hospital, Jiangmen, China,* [3]*Computational Oncology Group, Department of Surgery and Cancer, Imperial College London, London, United Kingdom,* [4]*Department of Radiotherapy, Charing Cross Hospital, Imperial College Healthcare NHS Trust, London, United Kingdom,* [5]*Duke University, Durham, NC, United States*

OPEN ACCESS

Edited by:
Charles Hsu,
Qatar University, Qatar

Reviewed by:
Peng Wang,
Temple University, United States
Alexandra Menni,
University General Hospital of Thessaloniki AHEPA, Greece
Ahmad Fathan Hidayatullah,
Islamic University of Indonesia, Indonesia

***Correspondence:**
Jiarong Chen
garwingchan@163.com
Shijing Si
shijing.si@outlook.com

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics,
a section of the journal Frontiers in Molecular Biosciences

Received: 20 January 2022
Accepted: 01 April 2022
Published: 02 June 2022

Citation:
Chen J, Williams M, Huang Y and Si S (2022) Identifying Topics and Evolutionary Trends of Literature on Brain Metastases Using Latent Dirichlet Allocation.
Front. Mol. Biosci. 9:858577.
doi: 10.3389/fmolb.2022.858577

Research on brain metastases kept innovating. We aimed to illustrate what topics the research focused on and how it varied in different periods of all the studies on brain metastases with topic modelling. We used the latent Dirichlet allocation model to analyse the titles and abstracts of 50,176 articles on brain metastases retrieved from Web of Science, Embase and MEDLINE. We further stratified the articles to find out the topic trends of different periods. Our study identified that a rising number of studies on brain metastases were published in recent decades at a higher rate than all cancer articles. Overall, the major themes focused on treatment and histopathology. Radiotherapy took over the first and third places in the top 20 topics. Since the 2010’s, increasing attention concerned about gene mutations. Targeted therapy was a popular topic of brain metastases research after 2020.

Keywords: brain metastases, topic modelling, LDA, research trends, research topics

1 INTRODUCTION

Brain metastases are a common and devastating complication of cancer. It is estimated that brain metastases develop in 20% of patients with cancer (Nayak et al., 2012; Tabouret et al., 2012) although the true rate, as measured in autopsy studies may be as high as 40% (Percy et al., 1972; Tsukada et al., 1983; Achrol et al., 2019). The prognosis of patients who develop brain metastases is poor, with only 7% surviving more than 2 years (Hall et al., 2000).

Brain metastases are the result of haematogenous seeding of spread cells from primary tumours to the brain (Achrol et al., 2019). The most common primary tumours for patients with brain metastases are lung, breast, colorectal cancers, melanoma and renal cell carcinoma (Ostrom et al., 2018; Achrol et al., 2019). Established treatments for brain metastases include surgery, chemotherapy and radiotherapy, while newer approaches include immunotherapy and targeted therapies (Soliman et al., 2016; Niranjan et al., 2019; Galldiks et al., 2020). Prognostic factors,

Abbreviations: ALK, anaplastic lymphoma kinase; CT, computed tomography; EGFR, epidermal growth factor receptor; KPS, Karnofsky performance status; LDA, latent Dirichlet allocation; MRI, magnetic resonance imaging; PET, positron emission tomography; SRS, stereotactic radiotherapy; TKI, tyrosine kinase inhibitor; WBRT, whole brain radiotherapy.

including age, Karnofsky performance status (KPS) and control of primary tumour are well recognized, and predict median overall survival periods of between 2.3 and 7.1 months (Gaspar et al., 1997). Given their frequency and poor outcomes, there has been a substantial amount of research into identifying the mechanisms behind brain metastases and improving treatment strategies. Molecular analyses have revealed some genes specific to the risk of developing brain metastases, such as the tumour suppressor LKB1 and KRAS (Zhao et al., 2014). Gene expression profiling of brain metastases suggests metastases evolve from primary tumours in order to gain more neuronal cell characteristics and adapt to the microenvironment in the brain (Park et al., 2011; Brastianos et al., 2015).

An important element of conducting research is to understand the current literature. One way of doing this is through systematic reviews and meta-analysis. However, such approaches have very carefully defined inclusion criteria, and thus offer very detailed analysis, but only of a small portion of the literature. For example, our current systematic review and network meta-analysis of first-line treatment for brain metastases includes only randomized trials of different treatment approaches (Williams et al., 2018); it therefore explicitly excludes published work on risk factors, biology, prognosis, etc. As a consequence, such systematic reviews ignore much of the published literature (Kozlowski et al., 2021), and thus do not help us understand the literature as a whole.

One approach to obtaining a better overview of the total scope of the literature is topic modelling, and the commonest approach is latent Dirichlet allocation (LDA). LDA is a popular topic modelling algorithm that has been widely used in different areas such as marketing, economics and bioinformatics (Blei et al., 2003; Shirota et al., 2014; Liu et al., 2016; Amado et al., 2018; Kozlowski et al., 2021) and helps discover topics in large corpora of text through clustering. Rather than considering the meaning of the sentences, the LDA model breaks the input text into single words and looks at groups of words that then occur together (Delen and Crossland, 2008; Liu et al., 2016). Such an approach requires some degree of pre-processing, in terms of removing common, non-significant words, and aligning related words that may have different ending (lemmatization). LDA allows us to identify research topics across a large body of literature and, importantly, does not require us to define a target topic defined before the analyses, and thus offers a relatively unbiased view of the literature.

In this study, we retrieved articles on cancer in general and brain metastases specifically and analysed the number of articles published, extracted topics and themes using LDA, and examined trends in these over time.

2 MATERIALS AND METHODS

2.1 Publication Assessment

We used a previously developed website to identify studies published and indexed in PubMed between 1947 and 2021 (https://esperr.github.io/pubmed-by-year/). We carried out two separate searches with terms of cancer and brain metastases on the platform in June 2021 to identify relevant publications, and reported numbers in each category and proportions over time.

2.2 Study Cohort

We searched for relevant studies with keywords of "brain metastases" (**Supplementary Table S1**) without limits on time or language (Soon et al., 2014; Zheng et al., 2016). The search was conducted in three databases including Web of Science (1970–2021), Embase (1947–2021) and MEDLINE (1950–2021) in June 2021. The search results were then imported into Endnote 20 (Camelot United Kingdom Bidco Limited, United Kingdom). We identified and removed duplicates with Endnote by comparing title, author, year, journal, volume and issues.

2.3 Data Pre-Processing

We extracted the title and abstract of every study found in the search. We pre-processed the text using a standard approach, and in line with other work (Cheng et al., 2020; Min et al., 2020). All text were converted to lowercase and we removed double spaces, special characters, and numbers. Subsequently, we applied a list of general English stop words and general words in abstracts (such as introduction, aim, purpose, method, conclusion, and discussion) to the titles and abstracts to remove non–information-bearing words from the text. We lemmatized words using the Python package scispaCy.

2.4 Topic Modelling and Themes

LDA, first proposed by Blei et al. (2003) in 2003, has been widely used in the biomedical literature analysis. The process for LDA is shown as follows:

First, the Dirichlet distribution η and θ in the selection process are defined: θ with parameter α for word selection and η with parameter β for topic section. Second, the general process for each document W is described in the following two steps:

1) Choose $\theta \sim \mathrm{Dir}(\beta)$.
2) For each of the n words ω_n:
   a) Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$.
   b) Choose a word ω_n from $p(\omega_n|z_n; \beta)$, a multinomial probability conditioned on the topic z_n.

Every cleaned, lemmatized abstract and title were treated as a single entry for the model. Topic analyses were conducted using the LDA model imported from the Gensim package in Python (3.0). The LDA model ignores the order of occurrence of terms and sentence structure and so regards each entry as a "bag-of-terms." Topics were defined as co-occurrence probability of individual words from the bag-of-terms. Following Cheng's study (Cheng and Hung, 2018), we chose the number of topics that yields the largest perplexity score. We identified the 20 topics, and manually grouped topics into related themes.

Finally, we collected all the textual data collected from every article (i.e., title and abstracts with stop words removed) into individual words to develop the text corpus for the whole data set and subsequently analyzed the word frequency using CountVectorizer in the Python package scikit-learn.

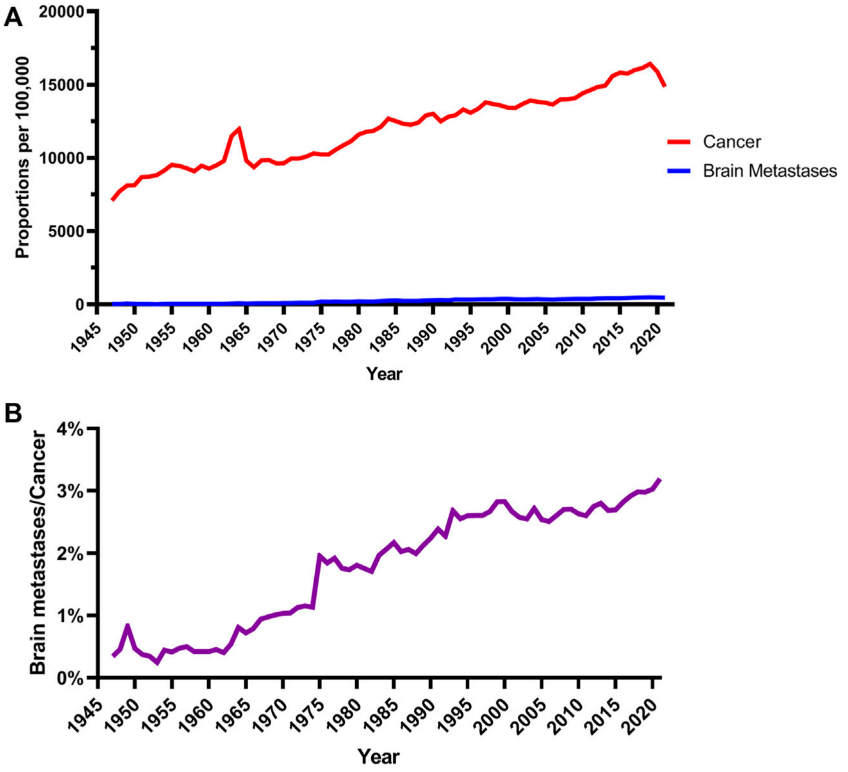


FIGURE 1 | Publication trends on cancer and brain metastases. **(A)** Annual PubMed proportion for cancer and brain metastases (source: https://esperr.github.io/pubmed-by-year/). **(B)** Percentage of articles on brain metastases in all cancer studies.

2.5 Visualization of Topics

Word clouds of the top 20 topics were generated using the WordCloud package in Python. Figures were plotted independently for each topic based on the first 20 terms. Size variation of the terms indicated the probability of the term in that topic.

2.6 Topic Trends

To illustrate the change of research trends on brain metastases in different periods, we stratified the data into cohorts by decades of the publishing dates. We illustrated the top 10 topics with 20 terms in the analysis for each cohort.

3 RESULTS

3.1 Publication Trends

Besides the absolute numbers of studies on brain metastases, it is useful to look at relative proportions. As illustrated in **Figure 1A**, literature on both brain metastases and cancer took up increasing proportions of all publications on PubMed during 1947 and 2021. All cancer research formed 7% of the articles in 1947, but gradually increased to 14% by 2009. However, within that general increase in research on brain metastases, brain metastases comprised less than 1% of all cancer research before 1968, but was over 3% by 2020 (**Figure 1B**).

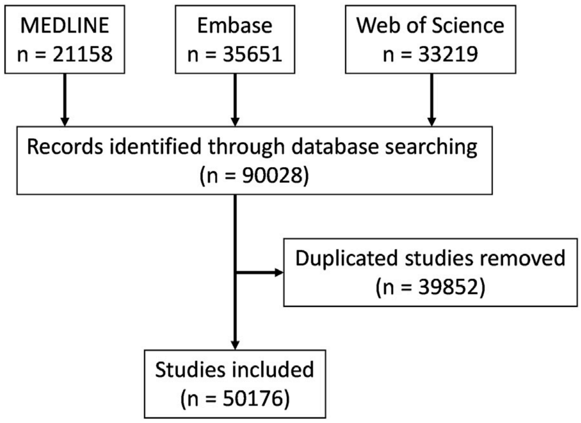


FIGURE 2 | Flow chart of study inclusion.

3.2 Data Inclusion

We retrieved 90,028 results, including 21,158 from MEDLINE, 35,651 from Embase and 33,219 from Web of Science. After duplicates removal, 50,176 results remained for further analysis (**Figure 2**). There were fewer than 50 articles on brain metastases per year before 1955 but the number kept increasing steadily between the 1960’s and 1980’s and reached 100 publications per year in 1974. The increase rate became even higher since 2000 and

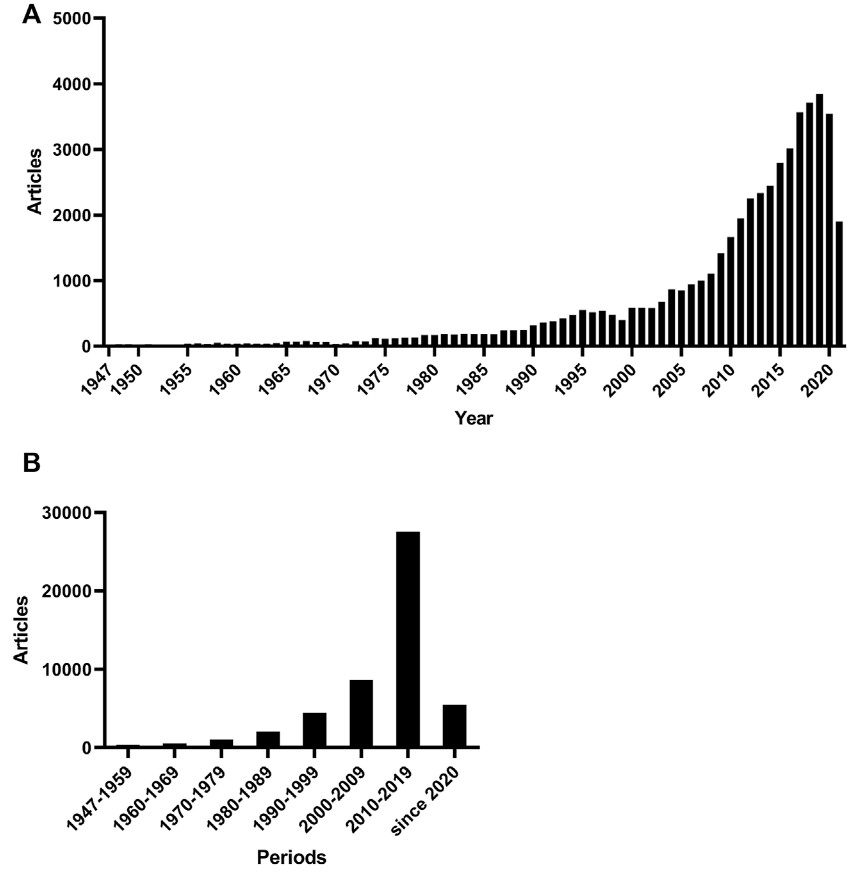


FIGURE 3 | The number of articles by year of publication.

publications stayed above 1,000 per year after 2007 (**Figure 3A**). The articles published in the 2010's overtook the total number between 1947 and 2009 (**Figure 3B**).

3.3 Topics on Brain Metastases

We identified 20 topics for brain metastases-related articles (**Table 1**; **Figure 4**). These topics correlated with different areas, including treatment (T1, T3, T6, T15, T17, and T19), tumour relationship (T2, T7, T12, and T16), histopathology (T4, T8, T9, T11, T14, and T20), diagnosis (T10, T13, and T18) and prognosis (T5). Treatment for brain metastases was the commonest topic, and in particular radiotherapy played a key role in two leading topics (T1 and T3). Another major theme was around discovering the fundamental mechanism of brain metastases.

3.4 Topic Trends in Different Periods

There were 251 studies whose publishing years were not available. Thus, these studies were excluded from the trend analysis. Given the small number of studies ($n = 77$) published between 1947 and 1949, we included studies from 1947–1949 in the 1950's cohort (**Supplementary Table S2**).

During 1947–1959, most of the articles tended to be descriptions of the symptoms, primary cancer and survival of patients with brain metastases. In the 1960's, more studies reported data on the detection of metastatic lesions from autopsy or scan. Chemotherapy became the commonest term for the first time in four of the top 10 topics in the 1970's (T1, T5, T6, and T10). Topic trends also revealed the involvement of improving imaging techniques in diagnosing brain metastases. Computed tomography (CT) firstly occurred in the term list of the 1970's, followed by magnetic resonance imaging (MRI) and positron emission tomography (PET) in the topic terms of the 1980's and 1990's. These imaging techniques remained of interest during 2000 and 2019. The involvement of new techniques in radiation also emerged since the 1990's focusing on stereotactic radiotherapy (SRS). Starting from the 2010's, terms related to gene mutations, such as epidermal growth factor receptor (EGFR), tyrosine kinase inhibitor (TKI) and anaplastic lymphoma kinase (ALK), became increasingly common in the articles. Targeted therapy was a popular topic of brain metastases research after 2020.

4 DISCUSSION

Brain metastases occur in more than 20% of all cancer patients and carry a poor prognosis (Nayak et al., 2012; Tabouret et al., 2012). Research into brain metastases is important, and existing

TABLE 1 | The highest frequent terms for 20 topics of the brain metastases-related articles.

Topic id	Topics	Highest frequent terms of topics
1	Radiotherapy	patient, survival, brain, month, metastasis, treatment, irradiation, treat, radiation, radiotherapy, year, therapy, median, follow, chemotherapy, disease, rate, tumour, local, time
2	Lung cancer	lung, metastasis, carcinoma, patient, cell, liver, cancer, lymph, stage, node, small, pulmonary, adenocarcinoma, brain, case, bone, bronchial, squamous, non, disease
3	Stereotactic radiosurgery	dose, use, volume, field, treatment, technique, target, ray, cm, irradiation, gamma, radiation, beam, carcinoid, high, stereotactic, fraction, position, film, normal
4	Basic science	cell, tumour, mouse, human, growth, brain, antibody, line, culture, cd, metastatic, antigen, use, melanoma, lymphocyte, specific, show, virus, injection, tissue
5	Prognosis	patient, factor, analysis, value, prognostic, survival, group, significant, index, test, use, high, study, significantly, ratio, correlation, clinical, score, predict, regression
6	Treatment development	treatment, clinical, disease, therapy, brain, use, review, discuss, therapeutic, well, new, system, also, make, important, development, give, patient, possible, many
7	Primary brain tumour	tumour, case, tumour, malignant, patient, intracranial, meningioma, glioma, operation, metastasis, surgical, surgery, glioblastoma, astrocytoma, metastatic, primary, lesion, diagnosis, brain, grade
8	Brain damage research on animal	rat, day, animal, brain, injury, increase, secondary, damage, follow, injection, induce, effect, change, min, spinal, control, ischemia, cord, decrease, study
9	Protein structure	structure, secondary, protein, form, olfactory, gene, type, sequence, dendrite, terminal, region, beta, different, find, contain, analysis, bind, suggest, site, study
10	Symptoms	seizure, patient, secondary, epilepsy, eeg, syndrome, generalize, discharge, focal, focus, disorder, epileptic, paralysis, temporal, cause, onset, occur, partial, type, drug
11	Nervous system	neuron, nucleus, cortex, secondary, response, area, stimulation, activity, primary, dopamine, increase, change, effect, motor, study, cortical, nerve, system, suggest, evoke
12	Primary cancer	metastasis, cancer, brain, patient, metastatic, breast, carcinoma, primary, tumour, bone, site, lung, cns, case, disease, diagnosis, renal, survival, liver, time
13	Symptoms & lesion characteristics	case, report, year, patient, lesion, present, old, symptom, diagnosis, show, cerebral, right, examination, leave, reveal, sign, brain, nerve, clinical, disease
14	Pharmacology	brain, effect, receptor, acid, activity, increase, cell, induce, release, membrane, concentration, bind, also, enzyme, mechanism, mouse, protein, system, rat, drug
15	Chemotherapy	patient, response, dose, day, chemotherapy, treatment, week, toxicity, therapy, mg, combination, disease, treat, complete, receive, month, study, phase, drug, cycle
16	Tumour type	tumour, cell, tissue, case, show, carcinoma, type, primary, find, stain, brain, cat, large, positive, small, muscle, thyroid, kidney, body, contain
17	Palliative care	patient, secondary, care, symptom, study, brain, disorder, injury, use, result, depression, pain, problem, life, hospital, function, medical, condition, stress, general
18	Imaging	lesion, brain, image, use, contrast, method, mr, study, high, mri, imaging, value, obtain, time, weight, patient, technique, result, normal, magnetic
19	Clinical trials	group, effect, control, study, treatment, significant, trial, difference, compare, significantly, result, week, measure, patient, improvement, reduce, placebo, receive, primary, test
20	White matter and cognitive deficit	patient, matter, subject, white, secondary, control, disease, change, brain, study, dementia, normal, memory, task, atrophy, ad, word, deficit, frontal, hemisphere

approaches, such as systematic reviews, while important, are limited in their scope. In this study, we retrieved articles on brain metastases and analysed the topics with the LDA model. Furthermore, we split the articles into cohorts according to their published dates and illustrated topics of different periods. An increasing number of articles on brain metastases have been published since 1950, rising at a higher rate than overall cancer research (**Figure 1**). We identified 20 main topics for articles and grouped these into 5 themes, of which treatment was the commonest.

We used LDA in this work as it allows us to identify research topics in the text of published studies and importantly does not require predefined target topics. In that sense, LDA allows us to develop an unbiased report of the literature, in contrast to systematic reviews which impose strict criteria. It also has convenient computational properties that allow us to scale up the analysis, and thus assess very large bodies of work.

Within the 20 different topics, we identified five themes that show the main areas of interest in publications about brain metastases. As expected, treatment was the commonest

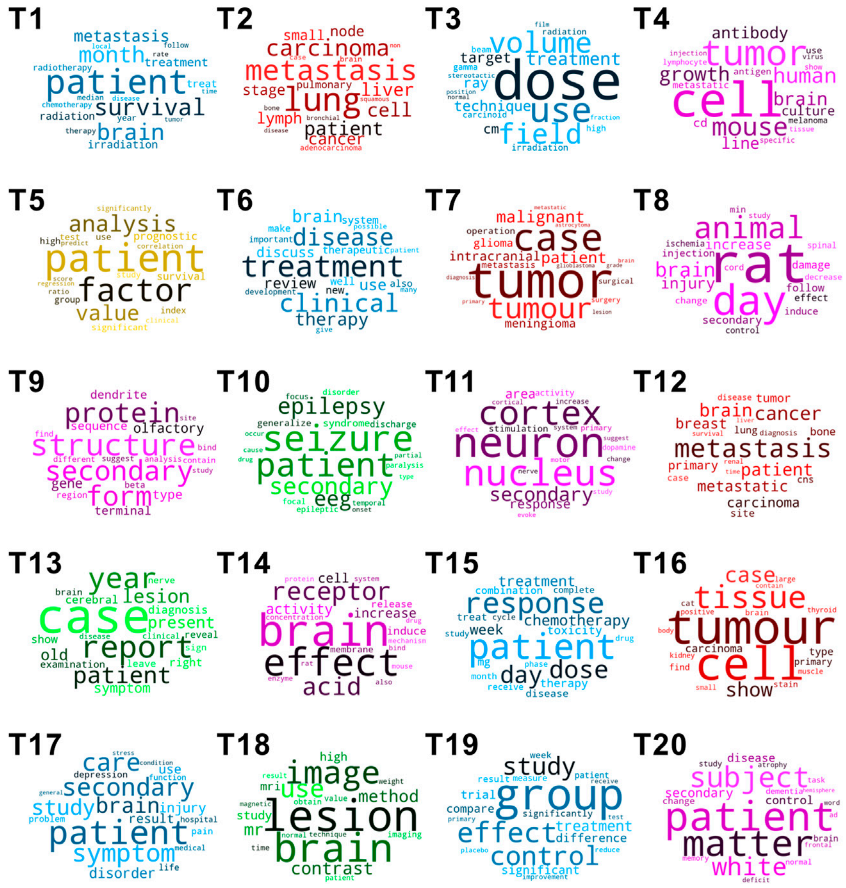


FIGURE 4 | Term frequency clouds of 20 topics on brain metastases. Topics in blue: topics related to treatment (T1, T3, T6, T15, T17, and T19). Topics in red: topics related to tumour relationship (T2, T7, T12, and T16). Topics in purple: topics related to histopathology (T4, T8, T9, T11, T14, and T20). Topics in mocha: topic related to prognosis (T5). Topics in green: topics related to diagnosis (T10, T13, and T18).

theme which accounted for six out of 20 topics, and importantly, the first and third commonest topics. Other topics were focused on the relationship between brain metastases and primary tumours, as well as histopathology, with an interest in understanding the mechanism of metastasis. Apart from these, prognosis and diagnosis were also important themes.

Radiotherapy appeared in both the first and third topics, indicating its crucial role in the treatment of brain metastases (**Table 1**; **Figure 4**). Traditionally, radiotherapy has been delivered as whole brain radiotherapy (WBRT) (Soffietti et al., 2005). However, there have been long-standing attempts to improve outcomes by varying dose and fractionation since the early 1960's (Chu and Hilaris, 1961; Peirce, 1964; Hindo et al., 1970; Hendrickson, 1977), which correlates with the importance of radiotherapy in our data since the 1960's (**Supplementary Table S2**).

The development of SRS which offers better local control and less normal tissue dosimetry (Soffietti et al., 2005; Graham et al., 2010; Abraham et al., 2018), has influenced the development of the literature. Since the 1990's, nearly all topics which included radiotherapy focused on the use of stereotactic techniques and related topics accounted appeared in at least one of the top 10 topics in those decades (T1 in the 1990's, T7 and T8 in the 2000's, T3 and T9 in the 2010's) (**Supplementary Table S2**).

Chemotherapy and related terms first occurred as the topmost terms in four of the top 10 topics in the 1970's (T1, T5, T6, and T10) (**Supplementary Table S2**) when there were a variety of studies trying to improve the outcome of brain metastases patients with chemotherapy (Gercovich et al., 1975; Black, 1979). However, this then decreased after the 1970's as people became aware of the effect of the blood-brain barrier in reducing the effect of chemotherapy in brain metastases. More recent work focuses on combining chemotherapy and radiotherapy (T9 in the 1990's, T1 in the 2000's) (**Supplementary Table S2**).

There has been a substantial increase in interest in the basic science associated with brain metastases since the 2000's (T6 in the 2000's, T2, T7, T8, and T10 in the 2010's, T2, T8, T9, and T10 since 2020). Targetable mutations, such as EGFR and ALK, were commonly reported in the studies since the 2010's (T2, T7, T8, and T10 in the 2010's), along with the relevant targeted therapy agents, including crizotinib, alectinib, and lorlatinib (T4 since 2020) (Hida et al., 2017; Martínez et al., 2017; Camidge et al., 2018; Khandekar et al., 2018) (**Supplementary Table S2**).

Imaging plays an important role in the diagnosis and management of brain metastases, and the topic trends follow this. CT first occurred in the term list of the 1970's (T3 in the 1970's), followed by MRI and PET which appeared in the top 10 topics of the 1980's (T10 in the 1980's) and 1990's (T1 in the 1990's). Imaging techniques for brain metastases remained a popular topic between 2000 and 2019 (T5 and T9 in the 2000's, T1 and T4 in the 2010's, T5 since 2020) (**Supplementary Table S2**).

The major omission is surgery. Despite the key role of surgery in the management of brain metastases, especially for large metastases (Soffietti et al., 2017; Rosenfelder and Brada, 2019) and several randomized trials showing the benefits of surgery combined with WBRT (Mintz et al., 1996; Gállego Pérez-Larraya and Hildebrand, 2014), surgically associated terms such as resection occurred only in topics related to radiosurgery (T7 in the 2000's, T3 in the 2010's) (**Supplementary Table S2**). This is in keeping with a general lack of research in surgery, and is a good example where the small number of studies examining surgery is a reflection of the weakness of the literature, rather than a measure of the relative importance of surgery.

There are some limitations to this study. First, even though we did not set limitations on languages or publication time, it is difficult to include all articles especially those written not in English or published before 1947 due to the restrictions in the databases we used. Meanwhile, the number of articles in the initial period was relatively small so that we combined articles between 1947 and 1949 with those of the 1950s when analysing topics of different periods. Second, an inspection of titles and abstracts shows many recent articles used new words and terms; however, these newer topics did not occur often enough to make the top 20 topics overall. Thirdly, the LDA model breaks sentences into a package of separate words and is more likely to consider their frequency. Therefore, the results may not convey the original context and significance of some phrases.

Overall, brain metastases remain a challenging clinical problem with high morbidity and poor prognosis. We have used LDA to provide an unbiased report of all the research into brain metastases since the last 1940's, and compared it to the baseline amount of research into cancer. It is notable that the literature on brain metastases has risen to occupy a larger proportion of the published cancer literature over time, and that the main therapeutic approach that dominates the literature is radiotherapy. While we do not suggest that simple count is sufficient to measure importance (i.e. there may be a few key, practice changing trials that involve surgery or chemotherapy), it does help us understand the scope of the literature. We think that this is important for several reasons. Firstly, it helps us understand that general scope of all literature in brain metastases. Secondly, it highlights where we might to focus our efforts for systematic reviews, where there may be more literature to review. Thirdly, it highlights both where they may be options to optimize existing treatments (e.g. optimizing radiotherapy) and also to address deficits in the literature (e.g. the relative absence of literature on surgery). This is important for clinicians, and also for research funders, who may want to reflect on the potential routes to improving the areas of research conducted in brain metastases.

5 CONCLUSION

In this paper, we presented an analysis of topics on brain metastases research by utilizing LDA modelling, which revealed the history of brain metastases studies and illustrated how treatment and diagnostic techniques developed in different periods. We found that brain metastases attracted increasing attention with a higher rate than overall cancer research, especially since 2000. Among all research on brain metastases, the most common themes were treatment and histopathology and radiotherapy occupied the first and third places in the top 20 topics, demonstrating its crucial role in brain metastases research.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Conceptualization, SS. Methodology, SS. Software, SS and JC. Validation, JC, SS, MW, and YH. Formal analysis, JC and YH. Investigation, JC and MW. Resources, JC. Data curation, JC and MW. Original draft preparation, JC. Writing review and editing, MW and YH. Project administration, JC. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the Guangdong International Young Research Talents Training Programme for Postdoctoral Researchers, the China Postdoctoral Science Foundation Grant (No. 2019M653210), the Science and Technology Project of Jiangmen (No. 2019030102480013011), the Medical Science Foundation of Jiangmen Central Hospital (No. J201901). MW is funded by Imperial College Healthcare NHS Trust and the Imperial NIHR BRC.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.858577/full#supplementary-material

REFERENCES

Abraham, C., Garsa, A., Badiyan, S. N., Drzymala, R., Yang, D., DeWees, T., et al. (2018). Internal Dose Escalation Is Associated with Increased Local Control for Non-Small Cell Lung Cancer (NSCLC) Brain Metastases Treated with Stereotactic Radiosurgery (SRS). *Adv. Radiat. Oncol.* 3 (2), 146–153. doi:10.1016/j.adro.2017.11.003

Achrol, A. S., Rennert, R. C., Anders, C., Soffietti, R., Ahluwalia, M. S., Nayak, L., et al. (2019). Brain Metastases. *Nat. Rev. Dis. Primers* 5 (1), 5. doi:10.1038/s41572-018-0055-y

Amado, A., Cortez, P., Rita, P., and Moro, S. (2018). Research Trends on Big Data in Marketing: A Text Mining and Topic Modeling Based Literature Analysis. *Eur. Res. Manage. Business Econ.* 24 (1), 1–7. doi:10.1016/j.iedeen.2017.06.002

Black, P. (1979). Brain Metastasis. *Neurosurgery* 5 (5), 617–631. doi:10.1227/00006123-197911000-0001510.1097/00006123-197911000-00015

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Machine Learn. Res.* 3, 993–1022.

Brastianos, P. K., Carter, S. L., Santagata, S., Cahill, D. P., Taylor-Weiner, A., Jones, R. T., et al. (2015). Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov.* 5 (11), 1164–1177. doi:10.1158/2159-8290.Cd-15-0369

Cheng, C-H., and Hung, W-L. (Editors) (2018). "Tea in Benefits of Health: A Literature Analysis Using Text Mining and Latent Dirichlet Allocation," 18.*ICMHI*.

Camidge, D. R., Kim, H. R., Ahn, M.-J., Yang, J. C.-H., Han, J.-Y., Lee, J.-S., et al. (2018). Brigatinib versus Crizotinib in ALK-Positive Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* 379 (21), 2027–2039. doi:10.1056/NEJMoa1810171

Cheng, X., Cao, Q., and Liao, S. S. (2020). An Overview of Literature on COVID-19, MERS and SARS: Using Text Mining and Latent Dirichlet Allocation. *J. Inf. Sci.* 0 (0), 016555152095467. doi:10.1177/0165551520954674

Chu, F. C. H., and Hilaris, B. B. (1961). Value of Radiation Therapy in the Management of Intracranial Metastases. *Cancer* 14, 577–581. doi:10.1002/1097-0142(199005/06)14:3<577::aid-cncr2820140318>3.0.co;2-f

Delen, D., and Crossland, M. D. (2008). Seeding the Survey and Analysis of Research Literature with Text Mining. *Expert Syst. Appl.* 34 (3), 1707–1720. doi:10.1016/j.eswa.2007.01.035

Galldiks, N., Kocher, M., Ceccon, G., Werner, J.-M., Brunn, A., Deckert, M., et al. (2020). Imaging Challenges of Immunotherapy and Targeted Therapy in Patients with Brain Metastases: Response, Progression, and Pseudoprogression. *Neuro Oncol.* 22 (1), 17–30. doi:10.1093/neuonc/noz147

Gállego Pérez-Larraya, J., and Hildebrand, J. (2014). Brain Metastases. *Handb. Clin. Neurol.* 121, 1143–1157. doi:10.1016/b978-0-7020-4088-7.00077-8

Gaspar, L., Scott, C., Rotman, M., Asbell, S., Phillips, T., Wasserman, T., et al. (1997). Recursive Partitioning Analysis (RPA) of Prognostic Factors in Three Radiation Therapy Oncology Group (RTOG) Brain Metastases Trials. *Int. J. Radiat. Oncology*Biology*Physics* 37 (4), 745–751. doi:10.1016/s0360-3016(96)00619-0

Gercovich, F. G., Luna, M. A., and Gottlieb, J. A. (1975). Increased Incidence of Cerebral Metastases in Sarcoma Patients with Prolonged Survival from Chemotherapy. Report of Cases of Leiomysarcoma and Chondrosarcoma. *Cancer* 36 (5), 1843–1851. doi:10.1002/1097-0142(197511)36:5<1843::aid-cncr2820360541>3.0.co;2-v

Graham, P. H., Bucci, J., and Browne, L. (2010). Randomized Comparison of Whole Brain Radiotherapy, 20 Gy in Four Daily Fractions versus 40 Gy in 20 Twice-Daily Fractions, for Brain Metastases. *Int. J. Radiat. Oncology*Biology*Physics* 77 (3), 648–654. doi:10.1016/j.ijrobp.2009.05.032

Hall, W., Djalilian, H., Nussbaum, E., and Cho, K. (2000). Long-Term Survival with Metastatic Cancer to the Brain. *Med. Oncol.* 17 (4), 279–286. doi:10.1007/bf02782192

Hendrickson, F. R. (1977). The Optimun Schedule for Palliative Radiotherapy for Metastatic Brain Cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 2 (1-2), 165–168. doi:10.1016/0360-3016(77)90024-4

Hida, T., Nokihara, H., Kondo, M., Kim, Y. H., Azuma, K., Seto, T., et al. (2017). Alectinib versus Crizotinib in Patients with *ALK* Positive Non-Small-Cell Lung Cancer (J-ALEX): An Open-Label, Randomised Phase 3 Trial. *The Lancet* 390 (10089), 29–39. doi:10.1016/s0140-6736(17)30565-2

Hindo, W. A., DeTrana, F. A., 3rd, Lee, M.-S., and Hendrickson, F. R. (1970). Large Dose Increment Irradiation in Treatment of Cerebral Metastases. *Cancer* 26 (1), 138–141. doi:10.1002/1097-0142(197007)26:1<138::aid-cncr2820260117>3.0.co;2-5

Khandekar, M. J., Piotrowska, Z., Willers, H., and Sequist, L. V. (2018). Role of Epidermal Growth Factor Receptor (EGFR) Inhibitors and Radiation in the Management of Brain Metastases from EGFR Mutant Lung Cancers. *Oncologist* 23 (9), 1054–1062. doi:10.1634/theoncologist.2017-0557

Kozlowski, D., Semeshenko, V., and Molinari, A. (2021). Latent Dirichlet Allocation Model for World Trade Analysis. *PloS One* 16 (2), e0245393. doi:10.1371/journal.pone.0245393

Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An Overview of Topic Modeling and its Current Applications in Bioinformatics. *SpringerPlus* 5 (1), 1608. doi:10.1186/s40064-016-3252-8

Martínez, P., Mak, R. H., and Oxnard, G. R. (2017). Targeted Therapy as an Alternative to Whole-Brain Radiotherapy in EGFR-Mutant or ALK-Positive Non-Small-Cell Lung Cancer with Brain Metastases. *JAMA Oncol.* 3 (9), 1274–1275. doi:10.1001/jamaoncol.2017.1047

Min, K.-B., Song, S.-H., and Min, J.-Y. (2020). Topic Modeling of Social Networking Service Data on Occupational Accidents in Korea: Latent Dirichlet Allocation Analysis. *J. Med. Internet Res.* 22 (8), e19222. doi:10.2196/19222

Mintz, A. H., Kestle, J., Rathbone, M. P., Gaspar, L., Hugenholtz, H., Fisher, B., et al. (1996). A Randomized Trial to Assess the Efficacy of Surgery in Addition to Radiotherapy in Patients with a Single Cerebral Metastasis. *Cancer* 78 (7), 1470–1476. doi:10.1002/(sici)1097-0142(19961001)78:7<1470::aid-cncr14>3.0.co;2-x

Nayak, L., Lee, E. Q., and Wen, P. Y. (2012). Epidemiology of Brain Metastases. *Curr. Oncol. Rep.* 14 (1), 48–54. doi:10.1007/s11912-011-0203-y

Niranjan, A., Lunsford, L. D., and Ahluwalia, M. S. (2019). Targeted Therapies for Brain Metastases. *Prog. Neurol. Surg.* 34, 125–137. doi:10.1159/000493057

Ostrom, Q. T., Wright, C. H., and Barnholtz-Sloan, J. S. (2018). Brain Metastases: Epidemiology. *Handb. Clin. Neurol.* 149, 27–42. doi:10.1016/b978-0-12-811161-1.00002-5

Park, E. S., Kim, S. J., Kim, S. W., Yoon, S.-L., Leem, S.-H., Kim, S.-B., et al. (2011). Cross-Species Hybridization of Microarrays for Studying Tumor Transcriptome of Brain Metastasis. *Proc. Natl. Acad. Sci. U.S.A.* 108 (42), 17456–17461. doi:10.1073/pnas.1114210108

Peirce, C. B. (1964). The Efficacy of Radiation Therapy in the Treatment of Tumors of the Brain and Brain Stem. *Clin. Neurosurg.* 10, 195–211. doi:10.1093/neurosurgery/10.cn_suppl_1.195

Percy, A. K., Elveback, L. R., Okazaki, H., and Kurland, L. T. (1972). Neoplasms of the Central Nervous System: Epidemiologic Considerations. *Neurology* 22 (4), 40. doi:10.1212/WNL.22.1.40

Rosenfelder, N., and Brada, M. (2019). Integrated Treatment of Brain Metastases. *Curr. Opin. Oncol.* 31 (6), 501–507. doi:10.1097/cco.0000000000000573

Soffietti, R., Abacioglu, U., Baumert, B., Combs, S. E., Kinhult, S., Kros, J. M., et al. (2017). Diagnosis and Treatment of Brain Metastases from Solid Tumors: Guidelines from the European Association of Neuro-Oncology (EANO). *Neuro Oncol.* 19 (2), 162–174. doi:10.1093/neuonc/now241

Soffietti, R., Costanza, A., Laguzzi, E., Nobile, M., and Ruda, R. (2005). Radiotherapy and Chemotherapy of Brain Metastases. *J. Neurooncol.* 75 (1), 31–42. doi:10.1007/s11060-004-8096-3

Soliman, H., Das, S., Larson, D. A., and Sahgal, A. (2016). Stereotactic Radiosurgery (SRS) in the Modern Management of Patients with Brain Metastases. *Oncotarget* 7 (11), 12318–12330. doi:10.18632/oncotarget.7131

Soon, Y. Y., Tham, I. W. K., Lim, K. H., Koh, W. Y., and Lu, J. J. (2014). Surgery or Radiosurgery Plus Whole Brain Radiotherapy versus Surgery or Radiosurgery Alone for Brain Metastases. *Cochrane Database Syst. Rev.* 2016 (3), Cd009454. doi:10.1002/14651858.CD009454.pub2

Shirota Y., Hashimoto T., and Sakura, T. (Editors) (2014)"Extraction of the Financial Policy Topics by Latent Dirichlet Allocation," TENCON 2014-2014 IEEE Region 10 Conference, 22-25 Oct. 2014, Bangkok, Thailand. (IEEE).

Tabouret, E., Chinot, O., Metellus, P., Tallet, A., Viens, P., and Gonçalves, A. (2012). Recent Trends in Epidemiology of Brain Metastases: An Overview. *Anticancer Res.* 32 (11), 4655–4662.

Tsukada, Y., Fouad, A., Pickren, J. W., and Lane, W. W. (1983). Central Nervous System Metastasis from Breast Carcinoma Autopsy Study. *Autopsy Study Cancer* 52 (12), 2349–2354. doi:10.1002/1097-0142(19831215)52:12<2349::aid-cncr2820521231>3.0.co;2-b

Williams, M., Chen, J., Hart, M. G., Hunter, A., Hawkins, N., Si, S., et al. (2018). First-Line Treatments for People with Single or Multiple Brain Metastases. *Cochrane Database Syst. Rev.* 2018(12). CD013223. doi:10.1002/14651858.CD013223

Zhao, N., Wilkerson, M. D., Shah, U., Yin, X., Wang, A., Hayward, M. C., et al. (2014). Alterations of LKB1 and KRAS and Risk of Brain Metastasis: Comprehensive Characterization by Mutation Analysis, Copy Number, and Gene Expression in Non-Small-Cell Lung Carcinoma. *Lung Cancer* 86 (2), 255–261. doi:10.1016/j.lungcan.2014.08.013

Zheng, M.-h., Sun, H.-t., Xu, J.-g., Yang, G., Huo, L.-m., Zhang, P., et al. (2016). Combining Whole-Brain Radiotherapy with Gefitinib/Erlotinib for Brain Metastases from Non-Small-Cell Lung Cancer: A Meta-Analysis. *Biomed. Res. Int.* 2016, 1–9. doi:10.1155/2016/5807346

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 | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 10 August 2022
DOI 10.3389/fmolb.2022.910688

Machine learning prediction of postoperative unplanned 30-day hospital readmission in older adult

Linji Li[1,2†], Linna Wang[3†], Li Lu[3*] and Tao Zhu[1*]

[1]Department of Anesthesiology, West China Hospital, Sichuan University and The Research Units of West China (2018RU012), Chinese Academy of Medical Sciences, Chengdu, China, [2]Department of Anesthesiology, The Second Clinical Medical College, North Sichuan Medical College, Nanchong Central Hospital, Nanchong, China, [3]College of Computer Science, Sichuan University, Chengdu, China

Background: Although unplanned hospital readmission is an important indicator for monitoring the perioperative quality of hospital care, few published studies of hospital readmission have focused on surgical patient populations, especially in the elderly. We aimed to investigate if machine learning approaches can be used to predict postoperative unplanned 30-day hospital readmission in old surgical patients.

Methods: We extracted demographic, comorbidity, laboratory, surgical, and medication data of elderly patients older than 65 who underwent surgeries under general anesthesia in West China Hospital, Sichuan University from July 2019 to February 2021. Different machine learning approaches were performed to evaluate whether unplanned 30-day hospital readmission can be predicted. Model performance was assessed using the following metrics: AUC, accuracy, precision, recall, and F1 score. Calibration of predictions was performed using Brier Score. A feature ablation analysis was performed, and the change in AUC with the removal of each feature was then assessed to determine feature importance.

Results: A total of 10,535 unique surgeries and 10,358 unique surgical elderly patients were included. The overall 30-day unplanned readmission rate was 3.36%. The AUCs of the six machine learning algorithms predicting postoperative 30-day unplanned readmission ranged from 0.6865 to 0.8654. The RF + XGBoost algorithm overall performed the best with an AUC of 0.8654 (95% CI, 0.8484–0.8824), accuracy of 0.9868 (95% CI, 0.9834–0.9902), precision of 0.3960 (95% CI, 0.3854–0.4066), recall of 0.3184 (95% CI, 0.259–0.3778), and F1 score of 0.4909 (95% CI, 0.3907–0.5911). The Brier scores of the six machine learning algorithms predicting postoperative 30-day unplanned readmission ranged from 0.3721 to 0.0464, with RF + XGBoost showing the best calibration capability. The most five important features of RF + XGBoost were operation duration, white blood cell count, BMI, total bilirubin concentration, and blood glucose concentration.

Conclusion: Machine learning algorithms can accurately predict postoperative unplanned 30-day readmission in elderly surgical patients.

OPEN ACCESS

EDITED BY
Charles Hsu,
Qatar University, Qatar

REVIEWED BY
Ana Cláudia Coelho,
University of Trás-os-Montes and Alto Douro, Portugal
Xianfeng Shen,
Hubei University of Medicine, China

*CORRESPONDENCE
Li Lu,
luli@scu.edu.cn
Tao Zhu,
xwtao_zhu@sina.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Molecular Diagnostics and Therapeutics,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 01 April 2022
ACCEPTED 06 July 2022
PUBLISHED 10 August 2022

CITATION
Li L, Wang L, Lu L and Zhu T (2022), Machine learning prediction of postoperative unplanned 30-day hospital readmission in older adult. *Front. Mol. Biosci.* 9:910688.
doi: 10.3389/fmolb.2022.910688

KEYWORDS

machine learning, unplanned hospital readmission, surgery, prediction, elderly

Trial registration: http://www.chictr.org.cn/showproj.aspx?proj=35795, ChiCTR, ChiCTR1900021290

Background

The unplanned hospital readmission rate is one of the most widely used indicators to assess hospital care quality (Gupta and Fonarow, 2018). Due to its substantial contribution to medical resource costs, unplanned hospital readmission is increasingly recognized as an important public health concern, especially in developed countries (Jencks et al., 2009; Axon and Williams, 2011). Geriatric surgical patients, vulnerable to chronic illnesses, are at higher risk of unplanned hospital readmission with compounded factors. Although not all of these readmissions are preventable, it is critical to propose an effective framework for their early identification. A substantial body of models exists to identify patients at risk for unplanned readmission (Miotto et al., 2016; Kansagara et al., 2011; van Walraven et al., 2012; ohnson et al., 2019). However, most of them were created based on a specific disease cluster and cannot be extrapolated to the entire postoperative population, particularly elderly surgical patients (Ali and Gibbons, 2017; Ko et al., 2020; Kong and Wilkinson, 2020; Mišić et al., 2020; Sander et al., 2020; Shebeshi et al., 2020; Wasfy et al., 2020; Amritphale et al., 2021).

Recently, machine learning (ML) algorithms were considered to be potential tools for developing clinical predictive models because of their ability to deal with multidimensional datasets and make accurate predictions (Deo, 2015; Jordan and Mitchell, 2015). Since ML algorithms can process nonlinear relationships and interactions between predictors, they may be increasingly used in medical modeling. In this study, we aimed to investigate if ML-based algorithms can accurately predict postoperative unplanned 30-day readmission in an elderly surgical patient cohort using input features, such as demographic, comorbidity, laboratory, surgical, and medication data.

Methods

Data extraction

This study has been registered in the Chinese Clinical Trial Registry (ChiCTR-1900021290), and ethical approval was obtained from the Ethical Review Board of West China Hospital, Sichuan University, China. All the relevant clinical data were prospectively collected during the course of our routine anesthesia risk assessment, intraoperative records, and postoperative follow-up using a structured data schema designed by our institution. We extracted perioperative information of elderly patients older than 65 who underwent surgeries under general anesthesia in West China Hospital, Sichuan University from July 2019 to February 2021. For patients who had multiple admission records, we only included their first admissions for analysis. Meanwhile, for patients who underwent multiple surgeries during a single hospitalization, we included all their surgeries for analysis. A flow chart describing the inclusion and exclusion process is shown in Figure 1.

Model endpoint definition

The label “postoperative 30-day unplanned readmission” was defined as follows: readmission due to the same surgical disease or postoperative complications within 30 days postoperatively in an unplanned fashion. Our professional follow-up personnel collected this information by telephone 30 days after surgery.

Data preprocessing

There were few admissions with missing data. Variables with a missing data rate greater than 30% were not included for model development. For numeric variables with a missing data rate less than 5%, the median of each variable was used for imputation. For numeric variables with a missing data rate between 5% and 30%, we performed various imputation techniques using mean absolute error (MAE) scores as estimated metrics for comparison. To estimate the score on an original full dataset, we excluded all missing value rows and randomly removed some values to create a new version of the dataset with artificially missing data. Then, we compared the performance of the random forest (RF) regressor on the complete original dataset with that on the altered dataset that used different imputation techniques. The comparison results presented in Figure 2 showed that we could find the lowest MAE to impute the missing values.

Considering the extreme imbalanced classification between the readmitted samples and non-readmitted samples (the readmission rate is only 3.36%), we both oversampled and undersampled the training set using the Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN). The SMOTE generated noisy samples by interpolating new points between marginal outliers and inliers, while ENN cleaned the space resulting from oversampling. Utilizing the SMOTE + ENN (SMOTEENN) algorithm provided by the imbalanced-learn *Python* library, we achieved a more balanced data distribution of readmitted samples and non-readmitted samples (Lemaître et al., 2017).

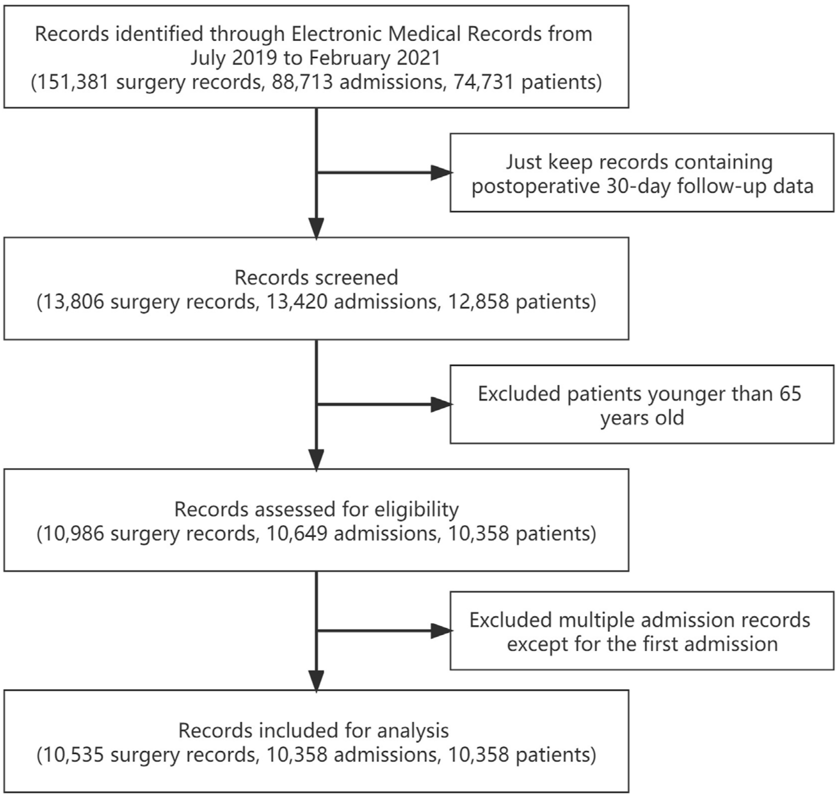

Records identified through Electronic Medical Records from July 2019 to February 2021
(151,381 surgery records, 88,713 admissions, 74,731 patients)
Just keep records containing postoperative 30-day follow-up data
Records screened
(13,806 surgery records, 13,420 admissions, 12,858 patients)
Excluded patients younger than 65 years old
Records assessed for eligibility
(10,986 surgery records, 10,649 admissions, 10,358 patients)
Excluded multiple admission records except for the first admission
Records included for analysis
(10,535 surgery records, 10,358 admissions, 10,358 patients)

FIGURE 1
Flow chart of inclusion and exclusion process for overall data set.

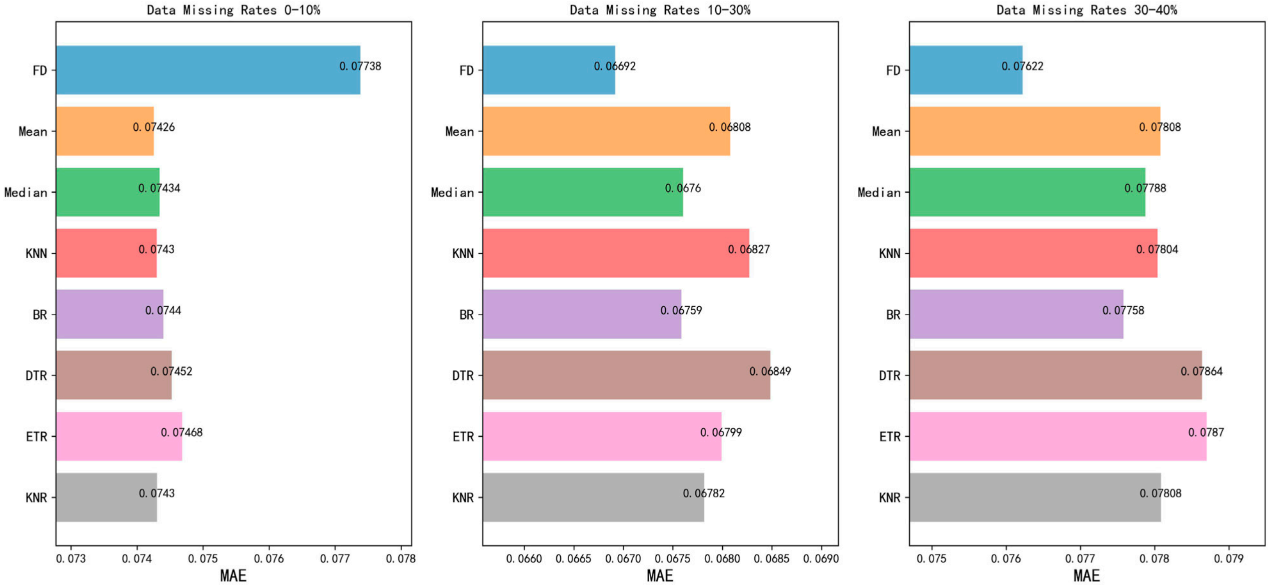


FIGURE 2
Imputation techniques in different missing data groups. FD, Full data; KNN, k nearest neighbor; BR, BayesianRidge; DTR, DecisionTreeRegressor; ETR, ExtraTreesRegressor; KNR, KNeighborsRegressor; MAE, Mean Absolute Error. BayesianRidge performed the best with the lowest MAE among all imputation techniques.

Our data were randomly divided into a training set and a test set according to a 70–30 split. We estimated models based on the training data (70%) and evaluated models based on the test data (30%). Each split was carried out to preserve the proportion of readmitted and not readmitted cases in the entire dataset. This random split was repeated ten times.

Feature selection

We focused on features that are easily accessible and not only available after discharge. For the preoperative laboratory data, we kept the last value prior to surgery. Before feature selection, we obtained 145 initial available variables. In model development, variable selection reduces the number of attributes and allows the selection of a subset of relevant features. Generally, there are three classes of optimal feature selection algorithms as follows: filter, wrapper, and embedded methods. In this study, we used the wrapper method because it can measure the usefulness of features based on the classifier performance through the search process, where different combinations of features are evaluated and compared by scores based on predictive model accuracy (Chandrashekar and Sahin, 2014).

To eliminate irrelevant, weakly relevant, or redundant features and reduce model overfitting as well as improve model generalization ability, we used a multilayer perceptron (MLP) as an estimator to implement a genetic algorithm (GA), which is a stochastic search algorithm based on the mechanics of evolution and natural selection (Torkamanian-Afshar et al., 2021). GA uses three operators, that is, selection, crossover, and mutation to improve the quality of solutions. We used Distributed Evolutionary Algorithms in *Python* to implement GA, while the function returns the optimal setting of feature selection as a binary array with the best accuracy score (Rainville et al., 2014). The independent probability for each attribute to be flipped was 0.1 in multiple flip-bit mutations. Tournament selection was set as the selection operator with a tournament size of 3. The population size was 100, the crossover probability was 0.5, and the mutation probability was 0.2.

The full list of features includes demographic data (e.g., age, gender, and body mass index [BMI]), available obtained laboratory tests prior to surgery (e.g., glucose concentration and oxygen saturation), descriptive intraoperative vital signs (e.g., systolic blood pressure), comorbidity (e.g., hypertension), and surgery descriptions (e.g., surgery type and anesthesia).

Model creation, training, and testing

This study considered different widespread types of models, that is, logistic regression, MLP, RF, extreme gradient boosting (XGBoost), and light gradient boosting machine (LGBM). The latter three are bagging or boosting ensemble learning algorithms. XGBoost is an optimized distributed gradient boosting library designed to have strong predictive power. It does not build the full tree structure but builds it greedily. It provides a parallel tree boosting that solves scientific problems, such as regression, classification, and ranking, in a fast and accurate way. LGBM is a high-performance gradient lifting framework that is based on a decision tree. Thus, it splits the tree leaf-wise with the simplest fit, whereas other boosting algorithms split the tree depth- or level-wise instead of leaf-wise. LGBM is quick because it uses a histogram-based algorithm that quickens the training procedure. We calculated MAEs as weights to combine RF and XGBoost into a hybrid model.

One of the advantages of using the abovementioned algorithms is that we can easily calculate the scores for all the input features, which represent the importance of each feature. A specific feature with a higher score means that it will have a larger effect on the model prediction. Random Forest Classifier, Logistic Regression, and MLP Classifier used in this study are from Scikit-learn. The XGB Classifier and LGBM Classifier were implemented using the xgboost and lightgbm packages (*Python* Software Foundation, 9450 SW Gemini Dr., ECM# 90772, Beaverton, OR 97008, United States) separately.

Model hyperparameters were set before training to improve the performance of the algorithms. We used RandomizedSearchCV and GridSearchCV provided by Scikit-learn. Five-fold cross-validation was applied to the training set, meaning that we calculated the average metrics while each of the five partitions was treated only once as a test set and four times as a training set. Before parameter optimization, all model classifier parameters were set to default values. We first used a random search with 200 iterations, and then a smaller range was determined based on the parameter selected in the previous step, and Grid Search worked with a small number of hyperparameters.

We used block bootstrapping to generate confidence intervals (CIs) for the performance metrics on the test set. Rather than randomly sampling procedures, we randomly sampled patients 1,000 times, included all predictions in the bootstrap sample, and sorted the performance metrics of each bootstrap sample.

Evaluation metrics

Model performance was assessed using the following metrics: area under the ROC curve (AUC), accuracy, precision, recall, and F1 score. ROC curve, as a visualization tool, can infer model performance by illustrating the relationship between precision and recall as we vary the threshold for selecting positives. Each time a different threshold was selected, a set of false-positive and true-positive rates were obtained. The calibration of the model was evaluated by Brier score and calibration plots. The 95%

TABLE 1 Summary of demographic characteristics and perioperative data in this cohort.

Variables	Training set	Testing set
Patients, n	**6,916**	**3,442**
Surgery, n (%)	**7,058(67.0)**	**3,477(33.0)**
Age (SD)	**72.1(5.8)**	**71.9(5.7)**
Female, n (%)	**2,990(43.2)**	**1,507(43.8)**
Readmission, n (%)	**237(3.36)**	**117(3.36)**
ASA		
I, **n** (%)	**11(0.16)**	**5(0.14)**
II, **n** (%)	**3,426(48.54)**	**1,667(47.94)**
III, **n** (%)	**3,531(50.03)**	**1764(50.73)**
IV, **n** (%)	**84(1.19)**	**39(1.12)**
V, **n** (%)	**6(1.0)**	**2(0.06)**
Surgery type		
Abdominal, n (%)	**3,711(52.58)**	**1782(51.25)**
Orthopedic, n (%)	**1,246(17.65)**	**673(19.36)**
Thoracic, n (%)	**636(9.01)**	**304(8.74)**
Cardiac, n (%)	**295(4.18)**	**163(4.69)**
Neuro, n (%)	**11(0.16)**	**5(0.14)**
Other, n (%)	**1,159(16.42)**	**550(15.82)**

The values in bold mean that they have the best performance in the metrics compared with all the other ML algorithms.

CIs of the abovementioned indicators were calculated through 1,000 repeated sampling. A feature ablation analysis was performed, and the change in AUC with the removal of each feature was then assessed to determine feature importance.

Results

Characteristics of the patients

Inclusion and exclusion criteria were strictly followed during the entire screening process. A flow chart indicating the inclusion and exclusion process is shown in Figure 1. Finally, a total of 10,358 elderly patients were included. The overall 30-day unplanned readmission rate was 3.36%. The demographic data and surgery-related information of patients are shown in Table 1.

Model performance

The AUCs of the six ML algorithms predicting postoperative 30-day unplanned readmission ranged from 0.6371 to 0.7686 including all features (Table 2) and from 0.6865 to 0.8654 including selected features (Table 3). The RF + XGboost classifier including selected features overall performed the best with an AUC of 0.8654 (95% CI, 0.8484-0.8824), the accuracy of 0.9868 (95% CI, 0.9834–0.9902), the precision of 0.3960 (95% CI, 0.3854–0.4066), recall of 0.3184 (95% CI, 0.259–0.3778), and F1 score of 0.4909 (95% CI, 0.3907–0.5911) (Table 3); The ROC curves of all the six ML algorithms predicting postoperative unplanned 30-day hospital readmission are shown in Figure 3, and the Precision-Recall (P-R) curves of all the six ML algorithms are also shown in Figure 4.

The Brier score of the RF + XGboost model predicting postoperative 30-day unplanned readmission was 0.0372 (95% CI, 0.0371–0.0372), showing the best calibration capability among all the ML algorithms (Table 4).

Feature importance

After performing a feature ablation analysis, we found that the five most important features of the RF + XGboost model were operation duration, white blood cell count, BMI, total bilirubin concentration, and blood glucose concentration. Figure 5 presents the feature importance of three models (RF、XGboost, and RF + XGboost) predicting postoperative unplanned 30-day hospital readmission.

Discussion

We used five ML models separately and one hybrid model to predict the 30-day postoperative unplanned readmission of elderly patients. To analyze the performance of the proposed framework,

TABLE 2 Performance of classification models including all features.

Model	AUC (95% CI)	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1 (95% CI)
RandomForest	0.7105 (0.6860–0.7350)	**0.9620(0.9610–0.9630)**	**0.3501(0.3000–0.4001)**	0.0120 (0.0110–0.0130)	0.0240 (0.0230–0.0250)
LogisticRegression	0.7145 (0.7110–0.7180)	0.9580 (0.9570–0.9590)	0.2160 (0.1820–0.2500)	0.0250 (0.0240–0.0260)	0.0442 (0.0431–0.0452)
XGBoost	0.6795 (0.6750–0.6840)	0.9606 (0.9601–0.9611)	0.2665 (0.2000–0.3333)	0.0125 (0.0120–0.0130)	0.0237 (0.0233–0.0240)
LGBM	0.6725 (0.6690–0.6760)	0.9595 (0.9590–0.9600)	0.2085 (0.1670–0.2500)	0.0125 (0.0120–0.0130)	0.0230 (0.0220–0.0240)
MLP	0.6371 (0.5741–0.7000)	0.9475 (0.9380–0.9570)	0.1621 (0.0630–0.2611)	0.0740 (0.0250–0.1230)	0.0920 (0.0350–0.1490)
Random + XGBoost	**0.7686(0.7396–0.7977)**	0.9524 (0.9523–0.9525)	0.3471 (0.3315–0.3627)	**0.1030(0.0950–0.1110)**	**0.1120(0.1100–0.1140)**

The values in bold mean that they have the best performance in the metrics compared with all the other ML algorithms.

TABLE 3 Performance of classification models including selected features.

Model	AUC (95% CI)	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1 (95% CI)
RandomForest	0.7566 (0.7481–0.7651)	0.9862 (0.9838–0.9885)	0.3950 (0.3900–0.4000)	0.3089 (0.1600–0.4578)	0.4287 (0.3952–0.4622)
LogisticRegression	0.7384 (0.7357–0.7411)	0.9503 (0.9474–0.9532)	0.2936 (0.2252–0.3620)	0.155 (0.1223–0.1878)	0.1957 (0.1406–0.2508)
XGBoost	0.7230 (0.7136–0.7324)	0.9862 (0.9835–0.9889)	**0.3977(0.3854–0.4100)**	**0.3289(0.2622–0.3955)**	0.4371 (0.3931–0.4812)
LGBM	0.7161 (0.6778–0.7544)	0.9867 (0.9855–0.988)	0.3882 (0.3763–0.4000)	0.3261 (0.2945–0.3578)	0.4385 (0.4197–0.4573)
MLP	0.6865 (0.6504–0.6226)	0.9744 (0.9711–0.9778)	0.2683 (0.2226–0.3140)	0.2434 (0.1568–0.3300)	0.2653 (0.6026–0.3281)
Random + XGBoost	**0.8654(0.8484–0.8824)**	**0.9868(0.9834–0.9902)**	0.3960 (0.3854–0.4066)	0.3184 (0.259–0.3778)	**0.4909(0.3907–0.5911)**

The values in bold mean that they have the best performance in the metrics compared with all the other ML algorithms.

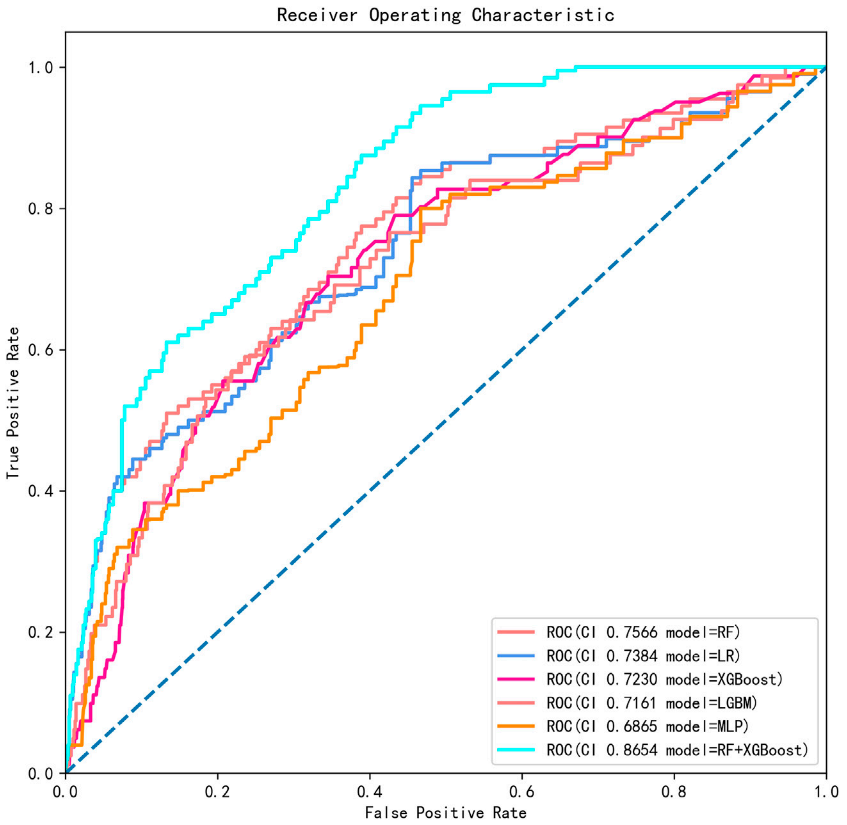


FIGURE 3
The ROC curves and AUCs of six ML algorithms predicting postoperative unplanned 30-day hospital readmission in this cohort. ROC, receiver operating characteristic; AUC, area under the curve; RF, random forest; LR, logistic regression; XGBoost, eXtreme Gradient Boosting; LGBM, Light Gradient Boosting Machine; MLP, Multilayer Perceptron.

we investigated the advantages and benefits of the proposed model over traditional ML models. Among all the algorithms, the RF + XGboost hybrid model generally performed relatively better, with an AUC of 0.8654 (95% CI, 0.8484–0.8824) and a Brier score of 0.0372 (95% CI, 0.0371–0.0372). For a single ML algorithm, RF nearly had the best performance in predicting the 30-day postoperative unplanned readmission, which has previously been reported (Peng et al., 2010; Hsieh et al., 2011; Alickovic and Subasi, 2016; Gowd et al., 2019). In addition, all ML models tended to perform similarly or better than the traditional approach (van Walraven et al., 2010; Cotter et al., 2012; Donzé et al., 2013; Low et al., 2017).

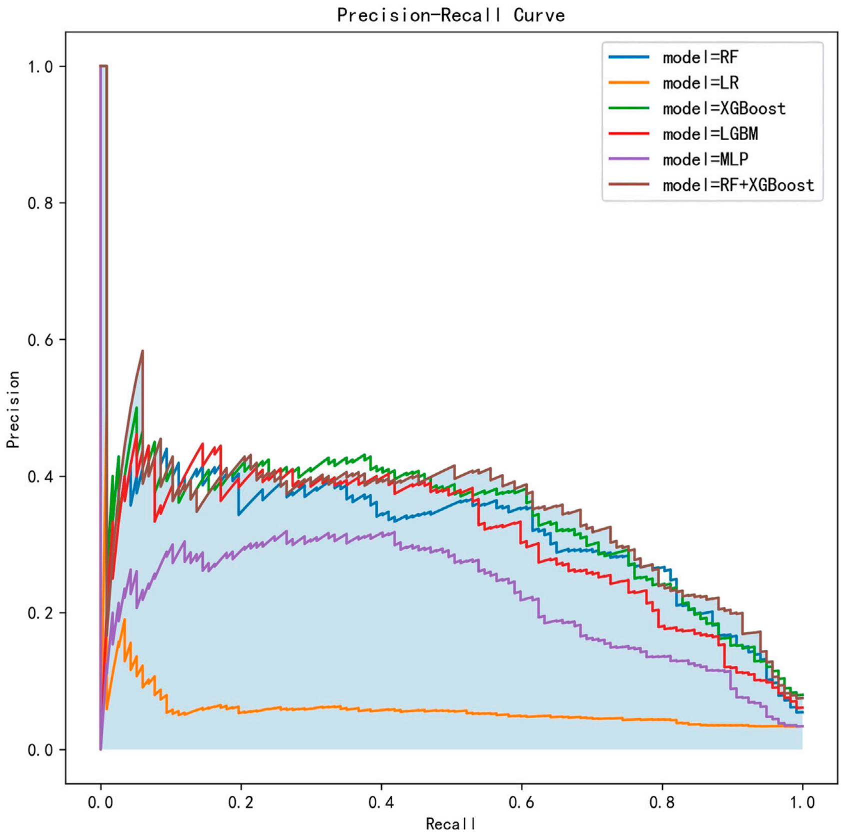


FIGURE 4
The P-R curves of six ML algorithms predicting postoperative unplanned 30-day hospital readmission in this cohort. P-R, Precision-Recall; RF, random forest; LR, logistic regression; XGBoost, eXtreme Gradient Boosting; LGBM, Light Gradient Boosting Machine; MLP, Multilayer Perceptron.

TABLE 4 Calibration of classification models including selected features.

Model	Brier Score (95% CI)
RandomForest	0.0383 (0.0377–0.0388)
LogisticRegression	0.0399 (0.0394–0.0403)
XGBoost	0.0389 (0.0386–0.0392)
LGBM	0.0377 (0.0375–0.0379)
MLP	0.0464 (0.0408–0.0519)
Random + XGBoost	**0.0372(0.0371–0.0372)**

The values in bold mean that they have the best performance in the metrics compared with all the other ML algorithms.

In the RF + XGboost model, the five most important features were operation duration, white blood cell count, BMI, total bilirubin concentration, and glucose concentration. Long duration of surgery is an important factor resulting in multiple postoperative complications, including unplanned 30-day postoperative readmission (Phan et al., 2017; Polites et al., 2017). Increased white blood cell count usually indicates an increased likelihood of infection. Postoperative infection is also an important reason for unplanned readmissions, such as lung infection requiring anti-infective treatment or wound infection requiring readmission for debridement or surgery. An increase in BMI is closely associated with higher incidence of hypertension, coronary heart disease, and diabetes, while reduced BMI, on the other hand, is also a sign of malnutrition and frailty status in the elderly (Graboyes et al., 2018; Sperling et al., 2018; Workman et al., 2020; Cutler et al., 2021). Hyperbilirubinemia reflects underlying hemolysis and hepatic dysfunction. Such patients have decreased tolerance for massive intraoperative blood loss, hypotension, and hepatic ischemia (Liao et al., 2013; Arvind et al., 2021). Elevated blood glucose level, usually including type 2 diabetes mellitus and impaired fasting glucose, is associated with postoperative infections, which are common causes of postoperative unplanned readmissions (Jones et al., 2017; Martin et al., 2019).

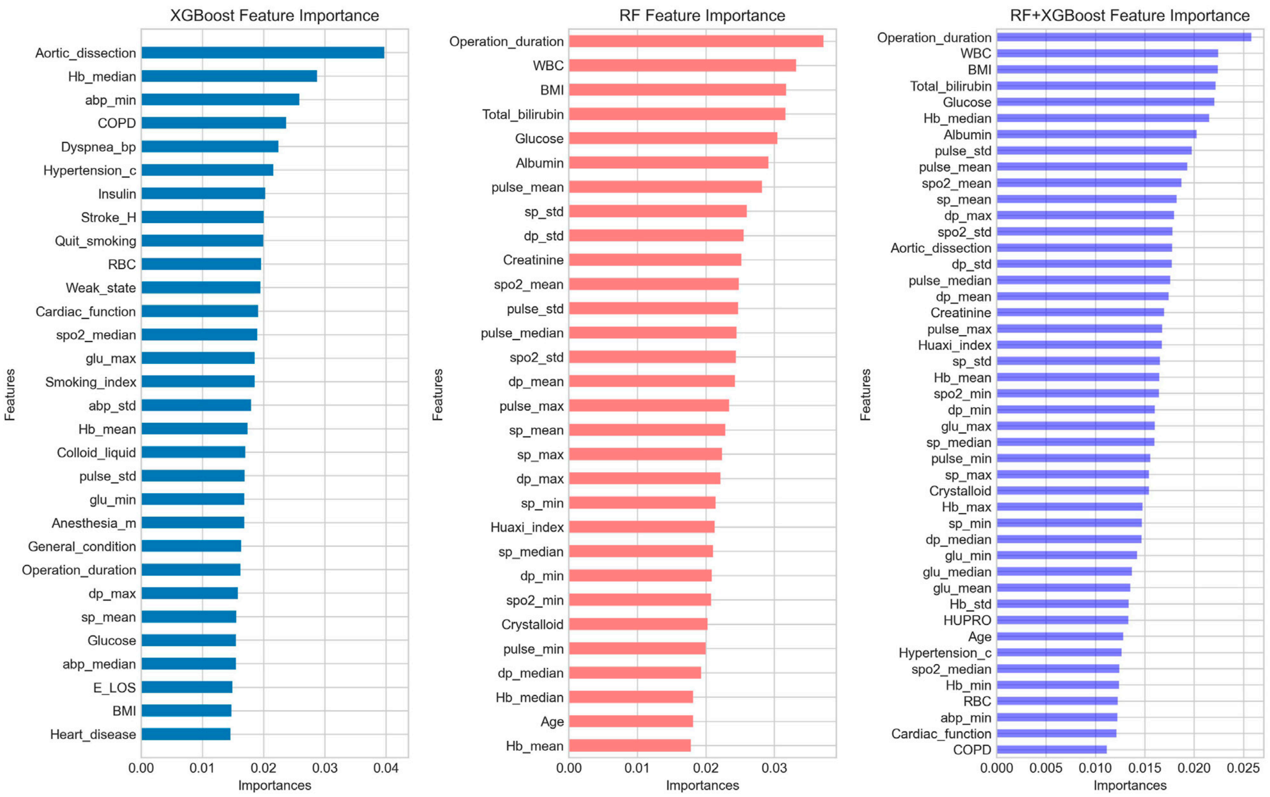


FIGURE 5
The feature importance of three ML algorithms predicting postoperative unplanned 30-day hospital readmission. RF, random forest; XGBoost, eXtreme Gradient Boosting.

To improve the performance of unplanned readmission risk prediction, we combined the RF and XGBoost classifiers by setting weights according to MAE. Our study demonstrates that the combined model could perform significantly better than individual models in predicting unplanned readmission. Meanwhile, among all the models, MLP did not achieve relatively good scores, which may be because the neural network algorithm is relatively complex for small unbalanced text datasets. Actually, the performance of ML algorithms is closely related to the imbalance rate of a label (e.g., imbalance rate of unplanned readmission). When the number of positive samples is excessively low (<10%), ML algorithms are easily overfitted. In this study, the 30-day unplanned readmission rate was lower than 5%, indicating a high probability of predicting patients as negative samples. Although we used SMOTEENN as a sampling method to reduce the imbalance rate, the classification performance has much room for improvement, as seen from the recall and F1 scores. The Brier score of the hybrid model is 0.0372 (95% CI, 0.0371–0.0372), which is also the lowest among all the algorithms.

Our analysis of postoperative patients provides us with three key insights into the prediction of unplanned readmission. First, ML is a powerful artificial intelligence approach to using data to imitate the way that humans learn and make decisions, gradually improving its accuracy. In this study, nearly all models achieved an AUC of more than 0.7, whereas studies predicting unplanned readmissions achieved AUC in the range of 0.54–0.92 (Artetxe et al., 2018). Second, hybrid models may perform better than individual models. Third, effective data processing is essential to assist decision-making. Strategies to reduce potentially avoidable 30-day readmissions may help improve the quality of care and outcomes.

Limitations

Some potential limitations should be considered. First, we did not include the information of hospital personnel for analysis. There is no doubt it is closely related to the patients' outcome; second, this is a monocenter study, and most of the patients came from western China. As a result, further external validation is needed. Third, the sample size is relatively small compared to some retrospective studies. Fourth, during data collection and follow-up, it is inevitable that some data will be missing.

Conclusion

ML algorithms can accurately predict postoperative unplanned 30-day readmission in elderly surgical patients.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the ethical review board of West China Hospital, Sichuan University, China. The ethics committee waived the requirement of written informed consent for participation.

Author contributions

Conception and design: LLi and TZ. Administrative support: LLu and TZ. Collection and assembly of data: LLi and LW. Data analysis and interpretation: LLi and LW. Manuscript writing and editing: LLi, LW, LLu, and TZ. All authors read and approved the final manuscript.

Funding

This study was supported by the National Key R&D Program of China (2018YFC2001800), Sichuan Provincial Science and Technology Key R&D Projects (2019YFG0491), and research funds from the Sichuan Provincial Health Commission (20PJ298).

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.

References

Ali, A. M., and Gibbons, C. E. (2017). Predictors of 30-day hospital readmission after hip fracture: a systematic review. *Injury* 48 (2), 243–252. doi:10.1016/j.injury.2017.01.005

Alickovic, E., and Subasi, A. (2016). Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier. *J. Med. Syst.* 40 (4), 108. doi:10.1007/s10916-016-0467-8

Amritphale, A., Fonarow, G. C., Amritphale, N., Omar, B., and Crook, E. D. (2021). All-cause unplanned readmissions in the United States. Insights from the Nationwide readmission database. *Intern. Med. J.* [Epub ahead of print]. doi:10.1111/imj.15581

Artetxe, A., Beristain, A., and Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Comput. Methods Programs Biomed.* 164, 49–64. doi:10.1016/j.cmpb.2018.06.006

Arvind, V., London, D. A., Cirino, C., Keswani, A., and Cagle, P. J. (2021). Comparison of machine learning techniques to predict unplanned readmission following total shoulder arthroplasty. *J. Shoulder Elb. Surg.* 30 (2), e50–e59. doi:10.1016/j.jse.2020.05.013

Axon, R. N., and Williams, M. V. (2011). Hospital readmission as an accountability measure. *JAMA* 305 (5), 504–505. doi:10.1001/jama.2011.72

Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. *Comput. Electr. Eng.* 40 (1), 16–28. doi:10.1016/j.compeleceng.2013.11.024

Cotter, P. E., Bhalla, V. K., Wallis, S. J., and Biram, R. W. (2012). Predicting readmissions: Poor performance of the LACE index in an older UK population. *Age Ageing* 41 (6), 784–789. doi:10.1093/ageing/afs073

Cutler, H. S., Collett, G., Farahani, F., Ahn, J., Nakonezny, P., Koehler, D., et al. (2021). Thirty-day readmissions and reoperations after total elbow arthroplasty: a national database study. *J. Shoulder Elb. Surg.* 30 (2), e41–e49. doi:10.1016/j.jse.2020.06.033

Deo, R. C. (2015). Machine learning in medicine. *Circulation* 132 (20), 1920–1930. doi:10.1161/CIRCULATIONAHA.115.001593

Donzé, J., Aujesky, D., Williams, D., and Schnipper, J. L. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Intern. Med.* 173 (8), 632–638. doi:10.1001/jamainternmed.2013.3023

Gowd, A. K., Agarwalla, A., Amin, N. H., Romeo, A. A., Nicholson, G. P., Verma, N. N., et al. (2019). Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. *J. Shoulder Elb. Surg.* 28 (12), e410–e421. doi:10.1016/j.jse.2019.05.017

Graboyes, E. M., Schrank, T. P., Worley, M. L., Momin, S. R., Day, T. A., Huang, A. T., et al. (2018). Thirty-day readmission in patients undergoing head and neck microvascular reconstruction. *Head. Neck* 40 (7), 1366–1374. doi:10.1002/hed.25107

Gupta, A., and Fonarow, G. C. (2018). The hospital readmissions reduction program-learning from failure of a healthcare policy. *Eur. J. Heart Fail.* 20 (8), 1169–1174. doi:10.1002/ejhf.1212

Hsieh, C. H., Lu, R. H., Lee, N. H., Chiu, W. T., Hsu, M. H., Li, Y. C., et al. (2011). Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. *Surgery* 149 (1), 87–93. doi:10.1016/j.surg.2010.03.023

Jencks, S. F., Williams, M. V., and Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *N. Engl. J. Med.* 360 (14), 1418–1428. doi:10.1056/NEJMsa0803563

Jones, C. E., Graham, L. A., Morris, M. S., Richman, J. S., Hollis, R. H., Wahl, T. S., et al. (2017). Association between preoperative hemoglobin A1c levels, postoperative hyperglycemia, and readmissions following gastrointestinal surgery. *JAMA Surg.* 152 (11), 1031–1038. doi:10.1001/jamasurg.2017.2350

Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science* 349 (6245), 255–260. doi:10.1126/science.aaa8415

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., et al. (2011). Risk prediction models for hospital readmission: a systematic review. *JAMA* 306 (15), 1688–1698. doi:10.1001/jama.2011.1515

Ko, D. T., Khera, R., Lau, G., Qiu, F., Wang, Y., Austin, P. C., et al. (2020). Readmission and mortality after hospitalization for myocardial infarction and heart failure. *J. Am. Coll. Cardiol.* 75 (7), 736–746. doi:10.1016/j.jacc.2019.12.026

Kong, C. W., and Wilkinson, T. M. A. (2020). Predicting and preventing hospital readmission for exacerbations of COPD. *ERJ Open Res.* 6 (2), 00325. doi:10.1183/23120541.00325-2019

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18 (1), 559–563. doi:10.48550/arXiv.1609.06570

Liao, J. C., Chen, W. J., Chen, L. H., Niu, C. C., Fu, T. S., Lai, P. L., et al. (2013). Complications associated with instrumented lumbar surgery in patients with liver cirrhosis: a matched cohort analysis. *Spine J.* 13 (8), 908–913. doi:10.1016/j.spinee.2013.02.028

Low, L. L., Liu, N., Ong, M. E. H., Ng, E. Y., Ho, A. F. W., Thumboo, J., et al. (2017). Performance of the LACE index to identify elderly patients at high risk for hospital readmission in Singapore. *Med. Baltim.* 96 (19), e6728. doi:10.1097/MD.0000000000006728

Martin, L. A., Kilpatrick, J. A., Al-Dulaimi, R., Mone, M. C., Tonna, J. E., Barton, R. G., et al. (2019). Predicting ICU readmission among surgical ICU patients: Development and validation of a clinical nomogram. *Surgery* 165 (2), 373–380. doi:10.1016/j.surg.2018.06.053

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* 6, 26094. doi:10.1038/srep26094

Mišić, V. V., Gabel, E., Hofer, I., Rajaram, K., and Mahajan, A. (2020). Machine learning prediction of postoperative emergency department hospital readmission. *Anesthesiology* 132 (5), 968–980. doi:10.1097/ALN.0000000000003140

ohnson, P. C., Xiao, Y., Wong, R. L., D'Arpino, S., Moran, S. M. C., Lage, D. E., et al. (2019). Potentially avoidable hospital readmissions in patients with advanced cancer. *J. Oncol. Pract.* 15 (5), e420–e427. doi:10.1200/JOP.18.00595

Peng, S. Y., Chuang, Y. C., Kang, T. W., and Tseng, K. H. (2010). Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. *Eur. J. Neurol.* 17 (7), 945–950. doi:10.1111/j.1468-1331.2010.02955.x

Phan, K., Kim, J. S., Capua, J. D., Lee, N. J., Kothari, P., Dowdell, J., et al. (2017). Impact of operation time on 30-day complications after adult spinal deformity surgery. *Glob. Spine J.* 7 (7), 664–671. doi:10.1177/2192568217701110

Polites, S. F., Potter, D. D., Glasgow, A. E., Klinkner, D. B., Moir, C. R., Ishitani, M. B., et al. (2017). Rates and risk factors of unplanned 30-day readmission following general and thoracic pediatric surgical procedures. *J. Pediatr. Surg.* 52 (8), 1239–1244. doi:10.1016/j.jpedsurg.2016.11.043

Rainville, F. M. D., Fortin, F. A., Gardner, M. A., Parizeau, M., and Gagné, C. (2014). Deap: enabling nimbler evolutions. *SIGEVOlution* 6 (2), 17–26. doi:10.1145/2597453.2597455

Sander, C., Oppermann, H., Nestler, U., Sander, K., von Dercks, N., Meixensberger, J., et al. (2020). Early unplanned readmission of neurosurgical patients after treatment of intracranial lesions: a comparison between surgical and non-surgical intervention group. *Acta Neurochir.* 162 (11), 2647–2658. doi:10.1007/s00701-020-04521-4

Shebeshi, D. S., Dolja-Gore, X., and Byles, J. (2020). Unplanned readmission within 28 Days of hospital discharge in a longitudinal population-based cohort of older Australian women. *Int. J. Environ. Res. Public Health* 17 (9), 3136. doi:10.3390/ijerph17093136

Sperling, C. D., Xia, L., Berger, I. B., Shin, M. H., Strother, M. C., Guzzo, T. J., et al. (2018). Obesity and 30-day outcomes following minimally invasive Nephrectomy. *Urology* 121, 104–111. doi:10.1016/j.urology.2018.08.002

Torkamanian-Afshar, M., Nematzadeh, S., Tabarzad, M., Najafi, A., Lanjanian, H., Masoudi-Nejad, A., et al. (2021). *In silico* design of novel aptamers utilizing a hybrid method of machine learning and genetic algorithm. *Mol. Divers.* 25 (3), 1395–1407. doi:10.1007/s11030-021-10192-9

van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., et al. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ* 182 (6), 551–557. doi:10.1503/cmaj.091117

van Walraven, C., Jennings, A., and Forster, A. J. (2012). A meta-analysis of hospital 30-day avoidable readmission rates. *J. Eval. Clin. Pract.* 18 (6), 1211–1218. doi:10.1111/j.1365-2753.2011.01773.x

Wasfy, J. H., Hidrue, M. K., Ngo, J., Tanguturi, V. K., Cafiero-Fonseca, E. T., Thompson, R. W., et al. (2020). Association of an acute myocardial infarction readmission-reduction program with mortality and readmission. *Circ. Cardiovasc. Qual. Outcomes* 13 (5), e006043. doi:10.1161/CIRCOUTCOMES.119.006043

Workman, K. K., Angerett, N., Lippe, R., Shin, A., and King, S. (2020). Thirty-day unplanned readmission after total knee arthroplasty at a teaching community hospital: Rates, reasons, and risk factors. *J. Knee Surg.* 33 (2), 206–212. doi:10.1055/s-0038-1677510

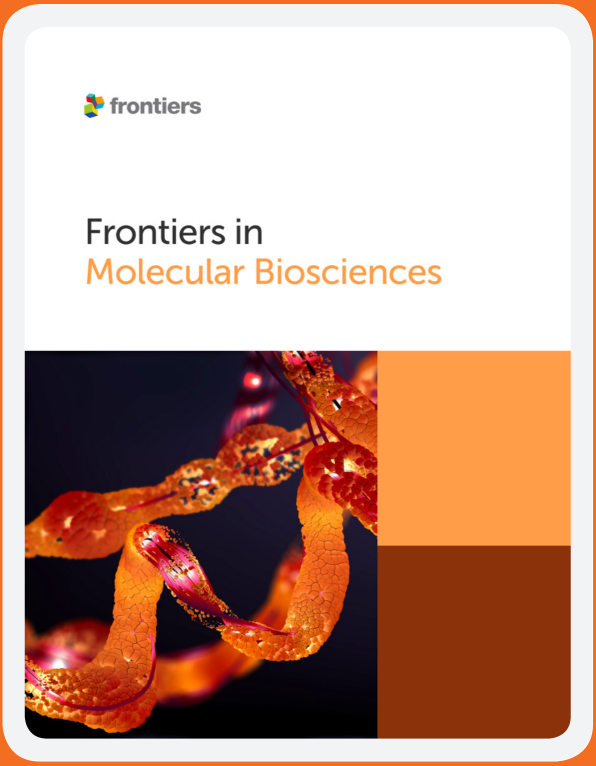
Frontiers in Molecular Biosciences

Explores biological processes in living organisms on a molecular scale

Focuses on the molecular mechanisms underpinning and regulating biological processes in organisms across all branches of life.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Molecular Biosciences

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact